nature portfolio

Peer Review File

Unifying multimodal learning with next-token prediction for large multimodal models

Corresponding Author: Dr Xinlong Wang

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The manuscript introduces Emu3, a multimodal model trained using next-token prediction across text, images, and video tokens. It claims state-of-the-art performance on various benchmarks, surpassing task-specific models in image and video generation as well as vision-language understanding. The work's attempt to unify multimodal learning through next-token prediction is innovative and holds potential significance. However, prior works such as Chameleon and Emu2 have already established foundational approaches in this area, rendering the novelty of this work somewhat limited. Furthermore, the novelty of this work is weak. Applying next-token prediction to unify multimodal model is not new, while this model does not achieve great progress.

Main Concerns and Limitations
1. Choice of Baseline Models: The comparison with baseline models in image and video generation is unconvincing. The selected models, including SD-1.5 and SDXL for image generation and OpenSoraV1.2 for video generation, are outdated and not representative of the state of the art. For instance, SD3 and Flux are superior recent models for image generation, while HunyuanVideo significantly outperforms OpenSora in video generation tasks. Demonstrating superiority over non-state-of-the-art baselines undermines the validity of the claims, especially given the manuscript's goal of validating next-token prediction as a unified approach.
2. Underutilization of Pretrained Language Models: While next-token prediction is promising, as evidenced by the success of large language models (LLMs), this work opts to train text modeling from scratch rather than leveraging pretrained LLMs. Incorporating pretrained parameters from advanced LLMs could enhance text modeling performance and efficiency. Moreover, the manuscript fails to explore strategies for adapting pretrained LLM backbones for visual data tasks, such as using a mixture-of-experts (MoE) architecture where specific experts handle pretrained text and visual tasks independently before integration.
3. Inadequate Vocabulary Size for Vision Tokenizer: The vocabulary size of the Vision Tokenizer (32,768 tokens) appears insufficient for capturing the complexity of visual data. Textual LLMs typically utilize vocabularies exceeding 200,000 tokens to handle linguistic diversity. Visual data, often more information-dense than text, likely requires an even larger vocabulary to achieve comparable expressiveness. The limited codebook size may constrain the model's performance, particularly in tasks requiring detailed visual understanding. Additionally, the compression performance achieved by the proposed Vision Tokenizer is weaker than that of other state-of-the-art video VAEs.
4. Weak Video Generation Results: The paper claims competitive video generation performance, but the provided results do not support this claim. The generated video demos are not visually impressive when compared to leading video generation models. Furthermore, the model's inability to generate long-duration videos (limited to less than 10 seconds) diminishes its impact, especially given that producing longer coherent videos is a primary challenge for generative models.
5. Lack of Exploration of Key Technical Challenges: The manuscript inadequately addresses several critical challenges in developing unified multimodal models, including: (a) The inherent modality gap between image/video data and text data, which requires sophisticated integration strategies beyond simple token concatenation; (b) The design of unified visual encoders capable of supporting both generation and understanding tasks. (b) Architectural considerations for building a robust multimodal backbone. (c) The selection of an optimal codebook size for Vision Tokenizer to balance token granularity and computational efficiency. By neglecting these challenges, the manuscript fails to advance the community's understanding of the complexities involved in applying next-token prediction to multimodal models.
6. Overgeneralized Claims about the contributions: The manuscript includes speculative claims about its contributions toward artificial general intelligence (AGI), which are unsupported by evidence. The benchmarks presented are limited to controlled tasks and do not demonstrate broader generalization capabilities. This undermines the credibility of the manuscript's claims and detracts from its otherwise strong empirical focus.

Conclusion
While the manuscript presents an interesting attempt to apply next-token prediction to unified multimodal modeling, its contributions are significantly limited by the choice of baselines, underexplored technical challenges, and unsubstantiated claims. To strengthen the impact of this work, future revisions should include:
1. Comparisons with state-of-the-art models across all tasks.
2. More rigorous experiments to explore critical design choices (e.g., codebook size, integration of pretrained LLMs).
3. A balanced discussion of limitations, technical challenges, and the model's real-world applicability.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The manuscript introduces Emu3, a suite of multimodal models trained exclusively via next-token prediction. By tokenizing images, text, and video into a discrete space and training a unified Transformer, Emu3 achieves state-of-the-art performance in both generation and perception tasks. This approach eliminates the need for diffusion models or compositional architectures like CLIP + LLMs. Key highlights include: (1) Unified Multimodal Learning: Emu3 trains a single Transformer on tokenized video, image, and text data; (2) Performance: It outperforms task-specific models, such as SDXL and LLaVA, on benchmarks like text-to-image generation, vision-language understanding, and video generation; and (3) Open Source: The model and key techniques are publicly available.

Despite these strengths, I have several major concerns:

1. Lack of Conceptual Novelty: The method is not conceptually new to the community. Unifying generation and understanding by framing the task as next-token prediction has already been explored in prior works, including:
[a] Chameleon: Mixed-Modal Early-Fusion Foundation Models, Chameleon Team, ArXiv 2024.
[b] Liu et al., World Model on Million-Length Video and Language with Ring Attention, ArXiv 2024.
[c] Li et al., SynerGen-VL: Towards Synergistic Image Understanding and Generation with Vision Experts and Token Folding, ArXiv 2024.
[d] Wu et al. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation, ArXiv 2024.
[e] Wu et al. Liquid: Language Models are Scalable Multi-modal Generators, ArXiv 2024.
From the current manuscript, it is unclear what new insights this work offers compared to these existing references.

2. Lack of Novel Technical Contributions: The method does not present new technical innovations, but rather appears to be an engineering effort combining existing techniques. For example:
[a]The vision tokenizer is based on SBER-MoVQGAN.
[b]The architecture for generation and synthesis is similar to that of Llama-2.
[c] Multilingual text tokenization uses QwenTokenizer.
The manuscript would benefit from clearer justifications of the novel contributions. Without such justifications, it reads more like an integration of existing methods rather than a new innovation.

3. Unimpressive Results: The reported results are not particularly compelling. As shown in Tables 4, 5, and 6, Emu3 underperforms compared to existing methods in many of the benchmark settings, let alone when compared to state-of-the-art commercial models. More justification is needed to establish the significance of these results.

In addition, I have several minor concerns:

1. Comparison with Diffusion and Encoder-Based Models: A detailed comparison with diffusion models and encoder-based architectures across benchmarks and metrics would strengthen the manuscript.

2. Generalization and Robustness: More analysis is needed on the model’s generalization to unseen datasets, as well as its robustness to noisy data.

3. Efficiency: Providing quantitative insights on training and inference efficiency, including computational costs, would be valuable for understanding the model's practicality.

4. Applications and Limitations: A broader discussion of potential applications and limitations of the model in real-world scenarios would be helpful.

5. Architecture Design Details: The manuscript would benefit from more architectural design details, including additional figures to better illustrate the technical components.

(Remarks on code availability)
They provided instructions on how to reproduce their results. Although I did not attempt to reproduce the results due to the

required computational resources, the code appears to be reproducible.

Referee #4

(Remarks to the Author)
Summary

This works proposes to tokenize image, videos and text and then utilize next token prediction to learn a single transformer model that can learn all of these modalities across a bunch of computer vision tasks.

Novelty

Next token prediction and autoregressive modeling has been becoming more popular in computer vision. Works like "4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities" and "GiT Towards Generalist Vision Transformer through Universal Language Interface" have explored similar ideas for a subset of computer vision tasks. What sets this work apart is that it extends this to generation tasks too and achieves very competitive results.

Scientific conclusions

It is hard for me to evaluate the quality of data provided by the authors as the work does not give enough technical details for a myriad of design choices made. Everything from how the dataset was created, how the various training details like which loss to propagate at what stage, their weights etc. are not revealed. The paper currently reads more like a technical report rather than a scientific paper. Since all of the baselines use different datasets, model size and different amount of compute, it is hard to disentangle scientific novelty from good engineering. The paper is very computer vision focused which is not a bad thing but without significant technical details, rigorous ablation studies and interesting scientific conclusions, it is hard to recommend an acceptance for a venue like nature. I don't take away from the author's achievement regarding a unified model across these modalities and tasks. However, for this to be accepted at a scientific venue, we need scientific questions and conclusions.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Thanks for the authors' detailed responses and new experiments.

I still have many questions about the new version:

1. About model architecture: the MoE architecture used for comparison is not well-designed and only adopt naïve MoE design for unified model, resulting in an increase in memory requirements, communication overhead, and training complexity. There are more advanced MoE designs can be adopted, such as the generation expert and understanding expert used in Bagel (ByteDance Seed team), but this work ignores them. I believe the current comparison of MoE and LLM-based initialization is unfair, and my main concern remains unaddressed.
2. Regarding the "Unified Vision Tokenization," the adopted VQ method of the current unified tokenizer is not significantly different from the original VQ quantization strategy, except that the convolutional decoder has a slightly modified architecture; however, the VQ quantization strategy remains the same. If no additional entropy loss is added to constrain the codebook, the codebook will basically learn to be the cluster center at the end of training and fall into the local optimum. Although the adopted VQ has designed the gradient for hard-assigned operations such as argmin, gradient-based optimization has the "winner takes all" problem, which should be the root cause of the collapse. In one batch, it is impossible for all tokens in the codebook to have gradients. The optimized tokens may be optimized again in the next batch, and gradually become a winner-takes-all situation.
3. About codebook size, the adopted VQ method is still a traditional VQ solution, the utilization rate is only 68% when the codebook size reaches 65k. I don't think the current "Unified Vision Tokenization" is well-designed and well-trained. So I think the current conclusion on codebook size is correct ("The tokenizer with codebook size of 32k largely outperform the one with codebook size of 8k in both the reconstruction (rFID) and generation (gFID). However, performance drops significantly when the codebook size is further increased to 65,536, likely due to the increased difficulty of training with such a large codebook.")
4. About the performance, the current version has improved in many benchmarks but is still leg behind many advanced unified models, such as Bagel (2025-05), in both generation and understanding. As for generation and editing, the current version lags far behind Qwen-image-edit and Nona banana. I think this work always ignores doing comparisons with more advanced unified work and advanced models.

Overall, even though this work tries to solve my problems, the main issues raised in my first round of review were not well addressed.

(Remarks on code availability)
The codes are still for the last version, not updated for this version. So there's no need to review.

Referee #3

(Remarks to the Author)
The authors have largely addressed my previous concerns. The quality of the paper has been substantially improved with new experiments, scientific insights and extensions to robustness and generality. In terms of novelty, Emu3 is the first in proving this paradigm scales competitively across modalities with rigorous scaling analyses, detailed ablations, and demonstrations on new domains (robotics, interleaved generation). This would be beneficial for further development of large models.
Therefore, I feel the paper has reached the bar of nature and I recommend it for publication after addressing the following minor remaining issues:
1. The unified tokenizer claims 4× token reduction. Could the authors provide or comment on more detailed qualitative visualizations of failure cases (e.g., fine-grained textures, high-motion video regions) where compression harms fidelity?
2. vLLM is reported to give major efficiency gains. How sensitive are these results to different hardware backends (A100, TPU v5, AMD GPUs)? Could there be limitations when deploying beyond NVIDIA GPUs?
3. While authors benchmark against open-source SOTA, could authors comment on how Emu3 qualitatively compares with closed commercial systems (e.g., Gemini, GPT-4V, Claude Opus with vision)?
4. The CALVIN benchmark is simulation-based. Could the authors comment on the feasibility of validating Emu3's vision-language-action performance in real-world robotic environments with noisy sensors and delayed feedback?

(Remarks on code availability)
I have reviewed the repository. It provides very comprehensive instructions for reproducing the results, though I have not tested them myself.

Referee #4

(Remarks to the Author)
The authors present Emu3, a large-scale multimodal model trained exclusively on a next-token prediction objective. The central and highly ambitious thesis is that this single, unified framework which tokenizes images, video, text, and actions into a shared discrete space to unify several branches of multimodal AI. The work claims this simple objective can replace the need for specialized, complex, and distinct architectures, such as diffusion models for visual generation and compositional (e.g., CLIP+LLM) models for visual perception and understanding.

I previously recommended this manuscript for rejection. My core objections were threefold, as the authors have correctly identified and summarized in their rebuttal:

Lack of Scientific Detail: The original paper was a "what" paper with no "why." It presented a model and its results without any deep investigation into the underlying principles. It lacked scaling laws, architectural ablations, or principled justifications for critical design choices (e.g., loss weighting, tokenizer design).

Inadequate Comparisons: The claims of state-of-the-art (SOTA) performance were unsubstantiated. The paper lacked rigorous, head-to-head comparisons against established methods.

Unsubstantiated Conclusions: The paper's conclusion claimed broad "generality" of the next-token framework but provided little empirical evidence beyond simple image-text generation and understanding tasks. This central claim was, therefore, an unproven assertion.

*Acknowledgement of Revisions
I must begin by commending the authors for their exceptionally thorough and detailed response to my (and other reviewers') critiques. The revised manuscript and the detailed rebuttal represent a monumental expansion of the original work from an engineering report to a scientific paper.

*Overview of New Material
The authors have directly and systematically addressed my three core objections by adding a wealth of new material:

For "Lack of Scientific Detail": The authors have added a new, extensive section on multimodal scaling laws (Sec 3.1, Figs 6-9), and new in-depth ablation studies on vision tokenization (Sec 4.1, Figs 14-16), multimodal architectures (Sec 4.2, Figs 17-18), and the training recipe (Sec 4.3, Figs 19-21).

For "Inadequate Comparisons": The paper now includes significantly expanded and comprehensive benchmark tables for image generation (Sec 3.2, Table 8), video generation (Sec 3.3, Table 9), and vision-language understanding (Sec 3.5, Table 10), supplemented by new human evaluations (Sec 3.2.2, Figs 10, 11).

For "Unsubstantiated Conclusions": The authors have added an entirely new "Extensive Applications" section (Sec 5) to prove generality, demonstrating applications in vision-language-action modeling for robotics (Sec 5.1), interleaved image-text generation (Sec 5.2), and flexible, non-raster prediction paradigms (Sec 5.3).

Issues and suggestions:
Video Generation (Sec 3.3, Table 9)
The authors claim Emu3 "demonstrates highly competitive results". Emu3, an autoregressive (AR) model, achieves a VBench score of 80.96. This is a strong score, beating most open-source models (e.g., OpenSora V1.2 at 79.76). However, Table 9 also lists multiple SOTA diffusion-based models that beat Emu3: T2V-Turbo (81.01), CogVideoX-5B (81.61), Kling (81.85), Gen-3 (82.32), and Wan2.1 (83.69).

The claim "highly competitive" is an overstatement. The real story, which is still highly impressive, is that this is likely the first unified, non-diffusion, AR-only model to get this close to SOTA video generation. But it is not "highly competitive" with the best models.

Vision-Language Understanding (Sec 3.5, Table 10)
The authors claim Emu3 (an encoder-free method) is "notably surpassing its counterparts". The authors claim Emu3 (an encoder-free method) is "notably surpassing its counterparts".

A detailed check of the new, comprehensive Table 10 shows the data is mixed.

Emu3: SEEDBench-Img = 68.2, OCRBench = 687, MMMU = 31.6

LLaVA-1.6(HD) (Compositional SOTA): SEEDBench = 64.7, OCRBench = 532, MMMU = 35.1

Qwen2.5-VL (Newer SOTA): OCRBench = 864, MMMU = 58.6

The data shows Emu3 does convincingly beat the flagship LLaVA-1.6 model on SEEDBench and especially on OCRBench (a 155-point margin), which is a very strong result. However, it loses to LLaVA-1.6 on the difficult MMMU reasoning benchmark. Furthermore, the authors' own table shows that newer models like Qwen2.5-VL significantly outperform Emu3. The authors have successfully demonstrated their "from-scratch" encoder-free model can be competitive with, and on some axes superior to, the dominant LLaVA-1.6 compositional model. This is a solid finding. But it is not SOTA, and the claim "surpassing its counterparts" is only true for other encoder-free models, not the field at large.

* Conclusion
I think the authors have done an impressive job in the new draft. I think the manuscript can still benefit from proper alignment of claims.

(Remarks on code availability)

Reply to the Reviewers

Manuscript 2024-11-24134

Unifying Multimodal Learning with Next-Token Prediction for Large Multimodal Models

Nature

1 General Response

We sincerely thank the reviewers for their thorough evaluation of our manuscript and for recognizing its significance of unifying multimodal learning. We greatly appreciate the feedback and insights of the reviewers, which have been invaluable in improving our work.

We have thoroughly addressed all of the reviewers' comments and incorporated all the feedback into our updated manuscript. The revisions fall into two major directions. First, we have added extensive new experiments to deeply engage with the scientific questions of what technical aspects of the large multimodal model drive strong performance. Second, we have extended the framework to challenging new capabilities and applications to demonstrate the robustness and generalizability of the next-token prediction framework for unified multimodal learning.

In this response, we first briefly summarize the new results, including: (1) in-depth analysis of scientific questions, (2) extensive ablation experiments, (3) new capabilities and applications, and (4) comprehensive details of the training and inference. Then we detail the response to address two major comments, i.e., (1) scientific questions, including multimodal scaling laws, unified tokenizer, architectural designs, etc., and (2) generalizability of the framework, including vision-language-action models for robotic manipulation, interleaved vision-language generation, and flexible prediction paradigm. Lastly, we address all the reviewers' concerns point by point. We have revised the manuscript accordingly to incorporate all the new results and the analysis.

We would like to highlight the significance of unifying multimodal learning at scale. After *AlexNet in 2012 successfully unified feature learning at scale*, and *Transformers in 2017 successfully unified sequence learning at scale*, multimodal learning still relies on tons of hand-crafted designs. But multimodal learning is crucial for next-generation foundations such as physical AI, as we live and interact in a context of multimodality. Inspired by how prior milestones simplified learning objectives at scale, our results suggest that a single next-token objective is a viable path for unified multimodal learning across text, image, video, and action. ***Emu3 successfully unified multimodal learning at scale***, and achieved competitive results compared to highly specialized frameworks optimized by the community for years. To our knowledge, Emu3 is the first system to show that a single next-token objective can deliver competitive performance across generation and understanding and extend to robotic manipulation and interleaved generation under a unified architecture. *Importantly, we also revealed the scientific questions and technical details behind the results.* We believe our work unlocks great opportunities, including native multimodal assistants, multimodal world models, multimodal robotic models, and beyond.

High-level summary of new results

We performed extensive experiments to validate Emu3's generality and scalability. Below we concisely summarize the new results; full details are provided in the general comments and the

point-by-point responses to reviewers.

1. **In-depth analysis of scientific questions**

    - **Scaling laws for multimodal learning:** We revealed a scaling law for unified multimodal learning, showing that increasing model size and training data yields consistent, synergistic improvements across tasks.
    - **Unified tokenization:** We showed that unified tokenization produces compact token sequences that achieve comparable performance to standalone image tokenizers while using roughly $4\times$ fewer tokens.
    - **General-purpose architecture designs:** We demonstrated that a decoder-only architecture attains comparable or superior training efficiency to encoder+LLM and diffusion counterpart under fair comparisons, while offering a more unified, general-purpose design.

2. **Extensive ablation experiments**

    - **Tokenizer codebook size:** We trained and analyzed a series of tokenizers with codebook sizes from 4,096 to 65,536 to quantify effects on reconstruction and generation.
    - **Architectural comparisons:** We performed careful comparisons against vision–language architectures (encoder-based and continuous-space encoder-free) and diffusion baselines such as Flan-T5-XL + DiT.
    - **LLM initialization and MoE variant:** We ran controlled experiments with LLM-based initialization and with MoE variants to validate robustness of our findings.
    - **Multimodal dropout:** We analyzed the role of multimodal dropout for stable training across modalities.
    - **Token loss weight:** We ablated the relative weights of visual and text token losses during pretraining.

3. **New capabilities and applications**

    - **Vision-language-action models for robotic manipulation:** We demonstrate that next-token prediction can naturally extends to robotic manipulation by treating vision, language, and actions as unified token sequences.
    - **Interleaved vision-language generation:** We show that the next-token prediction framework supports interleaved generation of image and text tokens.
    - **Flexible prediction paradigm:** We validate the robustness and flexibility of our framework by adapting to different token prediction orders (e.g., diagonal, block-raster, spiral-in).

4. **Comprehensive details of the training and inference**

- **Training stages, computes, and data processing:** We detail the training cost, data construction, and training pipeline, including stage configurations, parallelism strategies, optimization settings, and training steps.
- **Inference infrastructure and sampling strategies:** We present the inference framework, the comparative analysis of inference latency, and the impact of different sampling strategies.

For better readability, we use blue to denote the original reviewer comments, brown to indicate our revisions reflected in the revised manuscript, and black for our general and point-by-point responses.

1.1 Responses to General Comment #1

General Comment #1 *Providing a deeper and more thorough engagement with the scientific questions of what technical aspects of the model are responsible for good performance.*

We conducted extensive new experiments to rigorously investigate the scientific questions underlying strong performance in large multimodal models. We summarize the key insights and scientific questions that enable *unified multimodal learning at scale*.

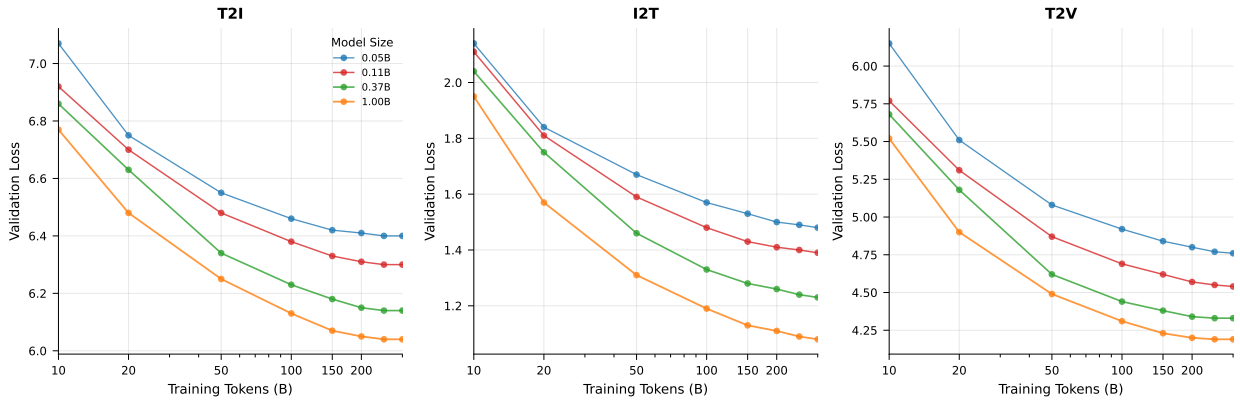


Figure 1: **Validation loss scaling with training tokens across multi-modal tasks.** Validation loss curves for models ranging from 0.05B to 1.0B parameters, excluding embedding layer parameters, across text-to-image, image-to-text, and text-to-video task, plotted as a function of the number of training tokens. Larger models consistently achieve lower validation loss, demonstrating a clear scaling trend with both model size and data volume.

1. **Scaling laws for multimodal learning.** We reveal a scaling law for unified multimodal learning. We show a clear power-law relationship: scaling up model size and training data produces consistent, synergistic improvements across text-to-image, image-to-text, and text-to-video tasks. This unified scaling law shows that a single next-token prediction model can handle these modalities without separate specialist architectures.

2. **Unified vision tokenization.** Our investigation into vision tokenizer designs reveals that mid-sized codebooks (e.g., 32k) trade-off well between reconstruction fidelity and downstream generation quality. The unified tokenizer achieves comparable performance to the

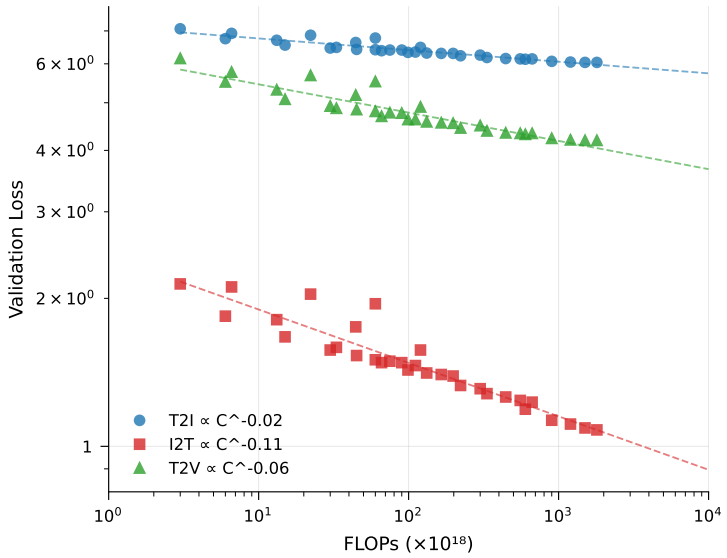


Figure 2: **Compute-optimal scaling relationships across tasks.** Log-log plot of validation loss versus training FLOPs for text-to-image (T2I), image-to-text (I2T), and text-to-video (T2V) tasks. The fitted power-law curves highlight task-specific scaling behaviors.

image-only tokenizer while using 4× fewer tokens on videos, enhancing modeling efficiency without sacrificing detail.

3. **General-purpose multimodal architecture.** We show that decoder-only token prediction architecture trained without any pretrained vision or language components can outperform or match traditional pipelines that rely on strong unimodal priors. Compared under equal compute, our unified architecture scales more efficiently and converges faster, challenging the prevailing assumption that compositional or diffusion-based models are inherently superior for multimodal learning.

4. **Training recipe matters for stable optimization.** We identify the critical training configurations, particularly dropout, token loss weighting, staged training, and learning rate schedules, that govern convergence stability in large-scale multimodal models. These findings allow us to stably train a 8B-parameter model from scratch, overcoming common issues such as mode collapse and unstable gradients during pretraining.

1.1.1 Scaling Laws in Multimodality

We reveal consistent scaling laws as a core principle underlying unified multimodal learning at scale. Inspired by the Chinchilla scaling law [14], our analysis reveals that diverse tasks including Text-to-Image (T2I), Image-to-Text (I2T), and Text-to-Video (T2V), follow a shared scaling behavior when trained jointly in a unified next-token prediction framework. We employ a power-law formulation to model the validation loss $L(N, D)$ as a function of model size N and training data size D:

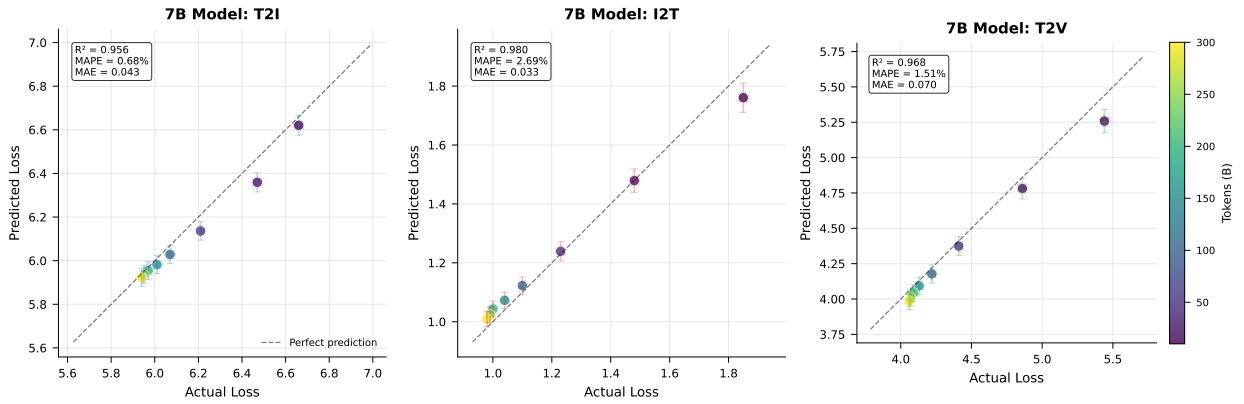


Figure 3: **Validating scaling predictions on the 7B model.** Predicted vs. actual validation losses of the 7B model (excluding embedding layer parameters) across tasks, color-coded by the number of training tokens. The strong alignment (MAPE <3% for all tasks) confirms the accuracy of our scaling law extrapolations.

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

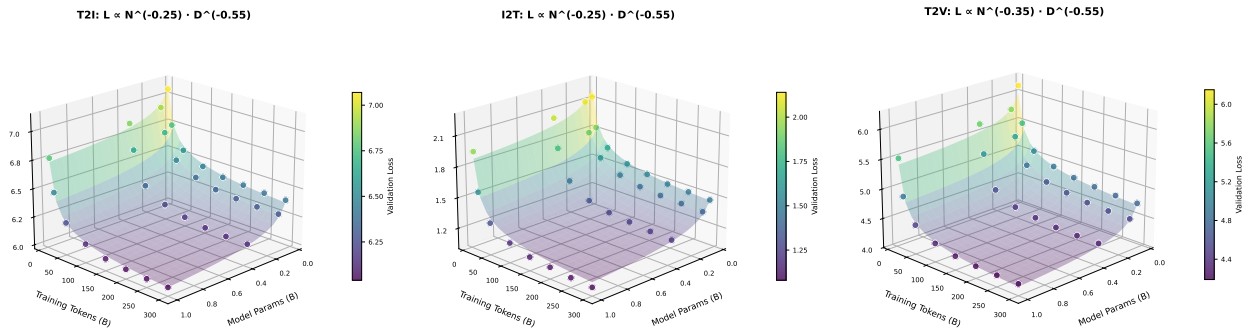
Our findings indicate that all tasks exhibit a consistent data scaling exponent $\beta = 0.55$. T2I and I2T share a model scaling exponent $\alpha = 0.25$, while T2V shows a steeper scaling with $\alpha = 0.35$. These results are supported by high-quality fits, with mean absolute percentage error (MAPE) below 3% and R^2 values exceeding 0.99.

Figures 1–4 present a coherent scaling narrative across data, compute, and predictive accuracy in unified multimodal training. Fig. 1 establishes the *data scaling trend*: for all three tasks, validation loss decreases predictably with more training tokens, with larger models benefiting more from additional data. Moving from data to compute, Fig. 2 shows *compute-optimal scaling* behaviors, where distinct power-law slopes reveal modality-specific trade-offs between scaling model parameters and dataset size. This links raw data scaling to practical compute allocation strategies. Fig. 3 then verifies the *predictive power* of these laws: extrapolations from smaller models closely match actual 7B (excluding embedding layer parameters) results ($R^2 \geq 0.95$, MAPE $< 3\%$), enabling reliable forecasts before full-scale training. Finally, Fig. 4 unifies the two scaling axes into *3D scaling surfaces* for each task, visualizing the joint dependence on model size and data volume, and yielding stable, modality-specific exponents.

These findings reinforce our central claim that a unified next-token prediction paradigm, when scaled appropriately, can serve as a simple yet powerful mechanism for multimodal learning, obviating the need for complex modality-specific fusion strategies.

1.1.2 Unified Vision Tokenization

Impact of codebook size. We train a series of image tokenizers with varying codebook sizes, ranging from 4,096 to 65,536, on ImageNet [32] 256×256, to analyze the impact of codebook size. Following SBER-MoVQGAN [31], we adopt the 67M parameters variant, keeping all configurations identical except for the codebook size. After the tokenizers are trained, we further train



(a) T2I Task (b) I2T Task (c) T2V Task

Figure 4: **3D visualization of scaling surfaces across tasks.** Validation loss surfaces plotted against model size (x-axis) and training token count (y-axis) for (a) text-to-image (T2I), (b) image-to-text (I2T), and (c) text-to-video (T2V) tasks. Each surface represents a fitted power-law scaling function: $L \propto N^{-0.25} \cdot D^{-0.55}$ for T2I and I2T, and $L \propto N^{-0.35} \cdot D^{-0.55}$ for T2V. Scatter points denote empirical measurements, while the semi-transparent surfaces illustrate the predicted scaling trends.

a 775M parameters image generation model [35] for each tokenizer. All image generation models are evaluated using classifier-free guidance of 1.80, top-k equal to the codebook size, top-p of 1.0, and a temperature of 0.98.

As shown in Fig.5 and Tab.1, increasing the codebook size consistently improves reconstruction metrics such as rFID, PSNR, and SSIM, up to a size of 32,768. The tokenizer with codebook size of 32k largely outperform the one with codebook size of 8k in both the reconstruction (rFID) and generation (gFID). However, performance drops significantly when the codebook size is further increased to 65,536, likely due to the increased difficulty of training with such a large codebook. The codebook utilization also drops to 68%, which further supports the hypothesis that training becomes more challenging as the codebook size grows excessively. Notably, the best gFID is also achieved with the codebook size of 32,768.

We also illustrate how gFID changes as both the codebook size and training steps increase in Fig. 6. The gFID consistently improves as training progresses for all experiments, with the model using a codebook size of 32,768 achieving the lowest gFID of 2.468.

Comparison between unified tokenizer and standalone image tokenizer. To evaluate the effectiveness of our unified video tokenizer, we compare its video reconstruction performance on UCF-101 [34] with the image tokenizer counterpart, which we use the SBER-MoVQ model with 270M parameters. We randomly sample 16 consecutive frames from each video in UCF-101. As shown in Tab. 2, under the same input resolution, our video tokenizer achieves comparable rFVD (27.893 vs. 26.675) and PSNR (27.546 vs. 30.499) while *using* $4\times$ *fewer tokens*. Moreover, when using the same number of latent tokens, the unified video tokenizer significantly outperforms the standalone image tokenizer, especially in terms of rFVD (27.893 vs. 139.930), demonstrating both its efficiency and effectiveness.

We also provide a qualitative comparison in Fig. 7. While the video tokenizer uses $4\times$ fewer latent tokens, it shows comparable reconstruction quality to the image tokenizer. It also preserves

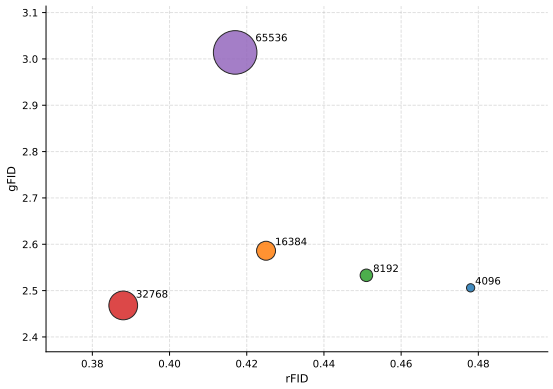


Figure 5: **Trade-off between reconstruction and generation across codebook sizes.** Scatter plot of reconstruction quality (rFID) versus generation quality (gFID) for different codebook sizes. Circle color and size indicate codebook size. Codebook size of 32k achieves the best trade-off.

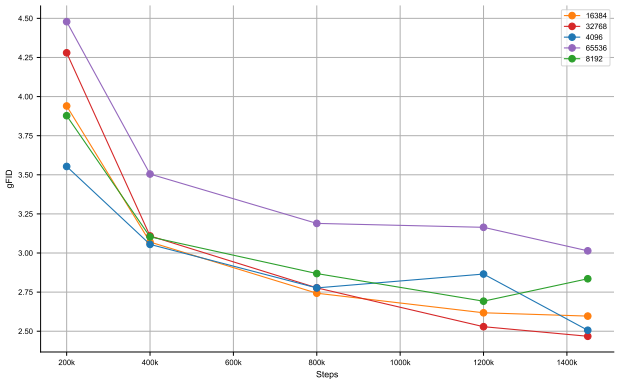


Figure 6: **Generation performance over training with different codebook sizes.** gFID curves over training steps for various codebook sizes. Each curve is color-coded by codebook size, revealing that excessively large vocabularies may hinder convergence.

finer details than the image tokenizer when using the same number of latent tokens.

1.1.3 General-Purpose Multimodal Architecture

Architectural comparisons with diffusion models. To fairly compare the next-token prediction paradigm with diffusion models for visual generation tasks, we use Flan-T5-XL [8] as the text encoder and train both a 1.5B diffusion transformer [28, 7] and a 1.5B decoder-only transformer [40] on the OpenImages [18] dataset. The diffusion model leverages the VAE from SDXL [30], while the decoder-only transformer uses the video tokenizer in **Emu3** to encode images into latent tokens. Both models are trained with identical configurations, including a linear warm-up of 2,235 steps, a constant learning rate of 1e-4, and a global batch size of 1,024. As shown in Fig. 8, the next-token prediction model consistently converges faster than the diffusion counterpart under equal training samples, underscoring the potential of next-token prediction as a more data-efficient framework for visual generation.

Architectural comparisons with encoder+LLM compositional paradigm. To fairly evaluate different vision-language architectures, we compare four model variants with comparable parameter counts and compute budgets on the image-to-text validation set (i.e., an image understanding task), as shown in Fig.9. The models compared are: (1) a decoder-only model that consumes discrete image tokens as input (Emu3 variant, 1.22B parameters); (2) an early-fusion decoder-only model using continuous image patch embeddings (EVE-style variant, 1.17B); (3) a late-fusion architecture comprising a vision encoder and decoder (LLaVA-style variant, 1.22B = 1.05B decoder + 0.17B vision encoder); and (4) a late-fusion architecture initialized with a larger CLIP-based vision encoder (LLaVA-style variant, 1.35B = 1.05B decoder + 0.30B vision encoder).

The above models are trained without any pretrained LLM initialization. FLOPs are estimated as $6ND$ for early-fusion models and $6(N_v D_v + ND)$ for late-fusion models, where N denotes the

codebook size	utilization	rFID(↓)	PSNR(↑)	SSIM(↑)	gFID(↓)
4,096	100%	0.478	24.38	0.7128	2.506
8,192	100%	0.451	24.62	0.7205	2.533
16,384	100%	0.425	24.75	0.7265	2.586
32,768	100%	**0.388**	**25.12**	**0.7404**	**2.468**
65,536	68%	0.417	24.91	0.7293	3.014

Table 1: **Comparison of vision tokenizers with varying codebook sizes.** Increasing the codebook size consistently improves both reconstruction quality (rFID, PSNR, SSIM) and generation performance (gFID), with gains observed up to a size of 32,768. However, further increasing the size to 65,536 leads to a significant drop in codebook utilization (68%), resulting in degraded performance.

	Resolution	Latent Size	Temporal Ratio	Spatial Ratio	rFVD(↓)	PSNR(↑)
Video MoVQ	320×240	$4 \times 40 \times 30$	4	8×8	27.893	27.546
Image MoVQ	320×240	$16 \times 40 \times 30$	-	8×8	**26.675**	**30.499**
	160×120	$16 \times 20 \times 15$			139.930	26.718

Table 2: **Unified video tokenizer vs. standalone image tokenizer.** Zero-shot reconstruction performance on 16-frame clips from UCF-101. The video tokenizer achieves competitive performance using 4× fewer latent tokens compared to the image tokenizer at the same resolution. When the image tokenizer is applied at a lower resolution to match the total latent count, its reconstruction quality drops significantly.

number of transformer decoder parameters, D the number of multimodal tokens, N_v the number of vision encoder tokens, and D_v the number of vision encoder parameters.

Notably, when trained from scratch, the presumed advantage of the encoder-based LLaVA-style compositional architecture largely diminishes. The decoder-only next-token prediction model not only achieves comparable performance but also converges faster, challenging the prevailing belief that encoder+LLM architectures are inherently superior for multimodal understanding.

While the EVE-style early-fusion model also adopts a decoder-only architecture, it lacks generative capability. In summary, the next-token prediction architecture supports both understanding and generation, and can match or even surpass compositional encoder+LLM paradigms in learning efficiency when compared under equal conditions.

1.1.4 Training Recipe for Stable Optimization

Large-scale unified multimodal learning is highly sensitive due to the diverse distributions of multimodal data. An improper training recipe can easily leads to training collapse, underscoring the fundamental difficulty of stable optimization at scale. We conduct ablation experiments of several key factors below. More training details such as training stages, parallelism strategies, loss weights, optimization hyperparameters, and training steps are shown in Tab. 3

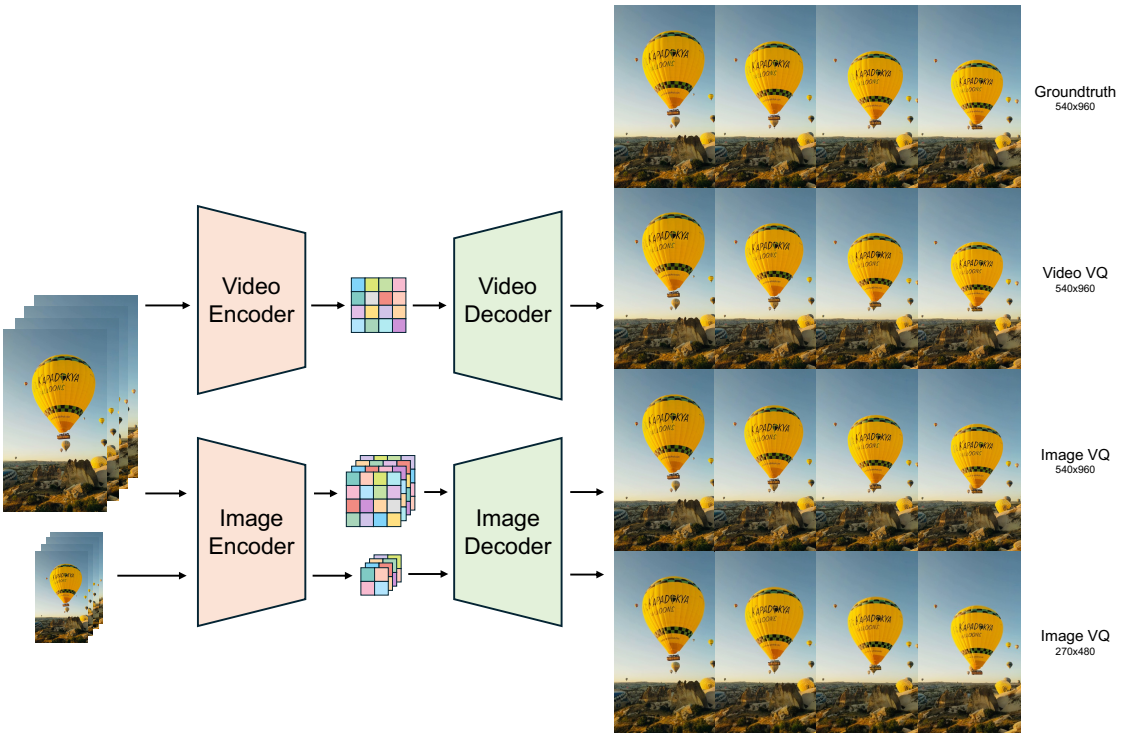


Figure 7: **Comparison of unified video tokenizer and standalone image tokenizer.** The video tokenizer achieves comparable reconstruction with 4× fewer latent tokens at the same resolution. When the image tokenizer is downsampled to match the total token count, its reconstruction quality degrades noticeably. Zoom in for details.

Multimodal dropout for training stability. During the pretraining, we initially trained the model with a dropout rate of 0. However, as shown in Fig. 12(a), the training became unstable and eventually collapsed in a later period. We conducted extensive ablations to address this issue, exploring z-loss, adding masked tokens, reducing the learning rate for fine-tuning, skipping optimizer state loading, label smoothing, and packing-based training. None of these strategies effectively resolved the instability. Only after introducing a small dropout rate of 0.1 and resuming from a stable checkpoint did training recover and proceed smoothly. This result highlights the critical role of dropout in maintaining stable convergence during large-scale multimodal learning.

LLM-based initialization and MoE architecture. While the effectiveness of next-token prediction in large language models (LLMs) is well-established, our work focuses on exploring whether this paradigm can serve as a unified and scalable solution for multimodal learning, beyond language alone. To this end, we deliberately avoid incorporating strong priors from pretrained LLMs in our primary experiments, so as to clearly evaluate the capability of next-token prediction from scratch in a multimodal setting.

That said, we also conducted controlled experiments with LLM-based initialization. As shown in Fig. 10, we show that using a pretrained LLM to initialize the language component of our model does provide a better starting point in early training stages (e.g., faster initial loss reduction). However, as training progresses, the gap between models with and without LLM initialization diminishes. Both converge to comparable validation loss, and the latter even converges better on

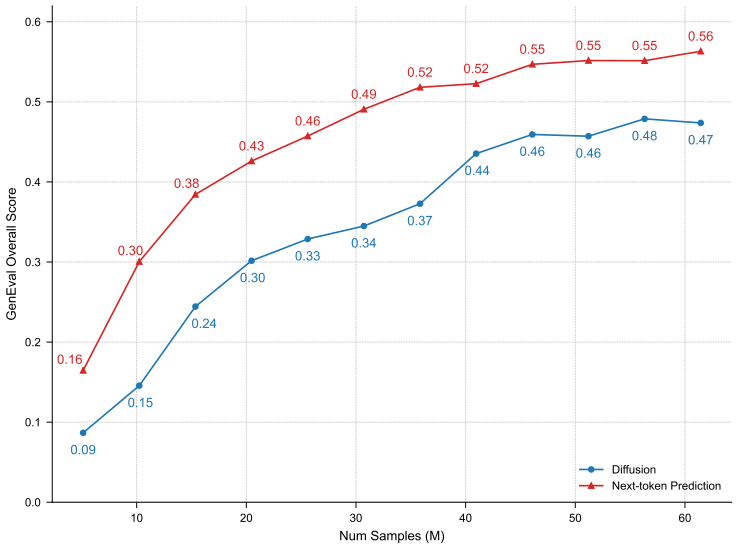


Figure 8: **Comparison of latent diffusion and next-token prediction paradigms.** GenEval overall scores as a function of training samples. Under equal training samples, the next-token prediction model consistently achieves higher scores and converges faster than its diffusion-based counterpart.

vision token loss. This indicates that our multimodal learning is robust and effective even without relying on pretrained language representations. While incorporating LLMs may offer efficiency benefits in the short term, Emu3's design can scale effectively from scratch, supporting its potential as a general-purpose, unified multimodal learner.

We also explored a mixture-of-experts (MoE) architecture variant in our framework. We conducted controlled experiments with MoE architectures to further validate our findings. Fig. 11 presents validation loss trajectories for Text-to-Image (T2I) and Image-to-Text (I2T) tasks in MoE models, providing additional empirical evidence for our architectural decisions. In fact, loading pretrained models can be detrimental to text-to-image tasks, as it introduces biases that hinder the model's ability to learn effective visual representations from scratch. Furthermore, our investigation into MoE architectures reveals significant efficiency limitations. Our implementation, which uses 4 experts with 1 activated per token, requires a 4x larger total parameter count in its expert layers compared to a dense equivalent to match the number of activated parameters. This design increases memory requirements, communication overhead, and training complexity.

These comparisons, combined with the observed asymmetry between understanding and generation tasks, support our design choice of a unified dense architecture with random initialization. Our comprehensive empirical analysis demonstrates that the unified next-token prediction paradigm, when properly scaled and trained from scratch, provides a robust and efficient foundation for general multimodal learning, eliminating pretrained initialization or complex architectural modifications.

Token loss weight. We ablate the weights of visual and text token losses during pretraining. As shown in Fig. 12, visual-only or text-only supervision optimizes task-specific performance at the expense of cross-modal generalization. Joint supervision (visual: 1.0, text: 1.0) achieves the best generation performance, demonstrating the benefit of unified learning. However, it underperforms on understanding compared to the text-only setting. To balance both tasks, we adopt a moderate

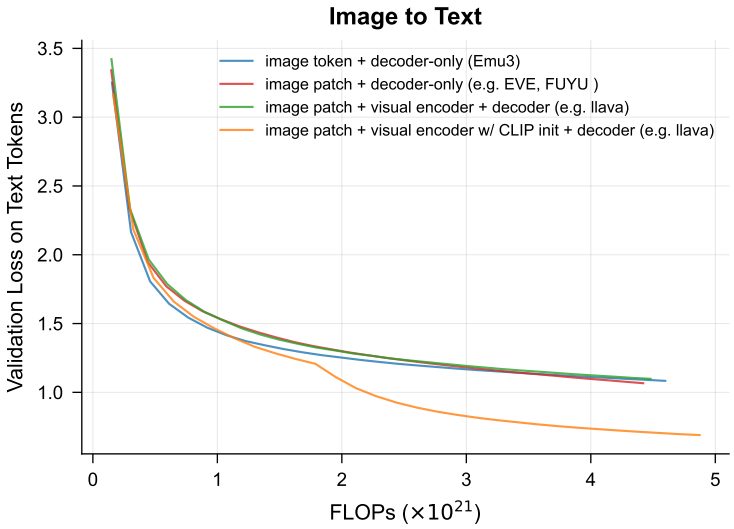


Figure 9: **Comparisons with encoder+LLM compositional paradigm.** Validation loss on text tokens vs. estimated training FLOPs ($\times 10^{12}$) for four architectures on an image-to-text validation dataset. Compared models include: (1) a decoder-only model with discrete image tokens as input (Emu3), (2) a decoder-only model with continuous image patches as input (e.g., EVE, Fuyu), (3) a late-fusion model with a vision encoder and decoder (e.g., LLaVA wo/ pretrain), and (4) the same as (3) but with a pretrained CLIP-large vision encoder (e.g., LLaVA). FLOPs are estimated as $6ND$ for early-fusion models and $6(N_v D_v + ND)$ for late-fusion models.

configuration (visual: 0.5, text: 1.0), which achieves competitive results on both tasks. This ablation underscores the necessity of careful loss balancing in large-scale multimodal learning, where naive parameter choices can lead to degraded performance or task bias.

Learning rate. We compare three different learning rates: 5×10^{-5}, 5×10^{-4}, and 5×10^{-3}. We observe that an overly large learning rate (e.g., 5×10^{-3}) quickly leads to unstable training and collapse, while a smaller one (e.g., 5×10^{-5}) greatly slows down convergence. In contrast, 5×10^{-4} achieves a good balance between training stability and efficiency, and is adopted as the default setting in our experiments.

1.1.5 Training and Inference Details

Training stages, compute, and data processing. Tab. 3 summarizes the training pipeline, including stage configurations, parallelism strategies, loss weights, optimization settings, and training steps. We estimate the training cost of Emu3 pretraining in terms of equivalent GPU hours on 1280 NVIDIA H100s. The model contains 8.49 billion parameters, with 6.98B in transformer layers and 1.51B in embedding layers, and is trained over three stages.

Stage1 and Stage2 share the same setup and respectively consume 96k GPU-hours, running at 3.0 seconds per iteration and 370 TFLOP/s per GPU. Stage3, which employs longer sequences (length is 65,536), is substantially more compute-intensive, requiring 25.1 seconds per iteration at 391 TFLOP/s per GPU, totaling 803k GPU-hours. In total, Emu3 pretraining consumes approxi-

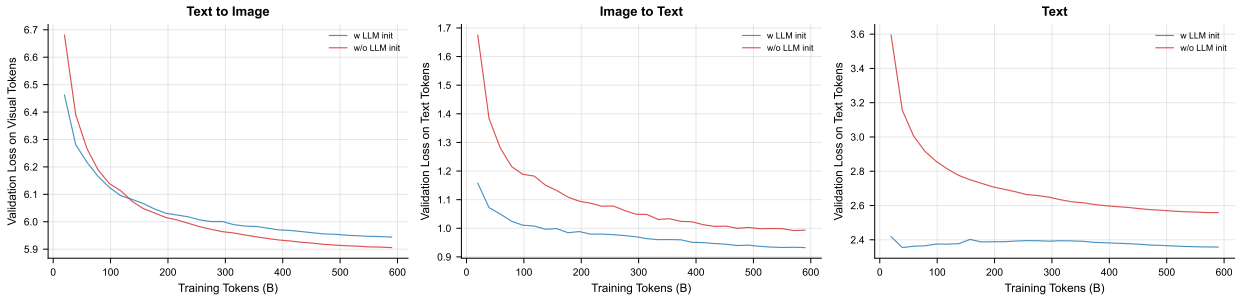


Figure 10: **Ablation experiments with LLM-based initialization on a 1.5B model.** We compare validation loss curves across three tasks: (1) vision token prediction loss on the text-to-image validation set, (2) text token prediction loss on the image-to-text validation set, and (3) text token prediction loss on the text-only validation set. Models initialized with a pretrained LLM (i.e., Qwen2.5-1.5b [40]) show faster convergence in early training stages, particularly for text-related losses. However, the performance gap narrows over time. The model without LLM initialization eventually achieves comparable or even better convergence on vision token loss, demonstrating the robustness of our unified multimodal learning.

mately 995k H100-equivalent GPU-hours.

Details of data construction including sources, filtering, and preprocessing are provided in Tab. 4.

Inference infrastructure. Our inference infrastructure is built on FlagScale [9], a multimodal serving system developed atop vLLM [19]. FlagScale inherits vLLM's key advantages, including the PagedAttention mechanism for fine-grained KV-cache management and continuous dynamic batching, which enables adaptive, per-token scheduling to maximize GPU utilization and minimize end-to-end latency. It is further accelerated by optimized CUDA kernels such as FlashAttention [11]. On this foundation, FlagScale extends the inference backend to support Classifier-Free Guidance (CFG) [13] for autoregressive multimodal generation. Specifically, we integrate CFG directly into the dynamic batching pipeline by co-feeding conditional and negative prompts within the same batch iteration. This design allows a unified forward pass at each decoding step, enabling correct logit fusion and guidance scaling without disrupting the batching semantics. Crucially, our CFG-aware extension incurs negligible overhead and preserves the low-latency, high-throughput characteristics of vLLM. As a result, FlagScale provides a scalable and efficient inference framework that supports fine-grained generation control across a wide range of multimodal tasks.

Tab. 6 presents a comparative analysis of inference latency and throughput between our vLLM-based backend and a standard Hugging Face (HF) Transformers implementation across three representative tasks: Text-to-Image (Emu3-Gen with CFG, averaging 8,216 tokens per image), Image-to-Text (Emu3-Chat without CFG), and Text-to-Video (Emu3 with CFG, averaging 65,536 tokens per video). All experiments were conducted on a single GPU node equipped with an NVIDIA H100 or its computational equivalent.

For Text-to-Image, vLLM attains a 3.8× speed-up at batch size 1 (68.1s vs. 259.7s) which increases to 6.7× at batch size 16; HF runs out of GPU memory beyond batch size 32. Correspondingly, token throughput under vLLM rises from 120 tok/s at batch size 1 to over 1260 tok/s

	Stage1	Stage2	Stage3
Hyperparameters			
Learning rate	5×10^{-4}	5.0×10^{-5}	5.0×10^{-5}
LR scheduler	Cosine	Constant	Constant
Weight decay	0.1		
Gradient norm clip	1.0	5.0	5.0
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1.0 \times 10^{-6}$)		
Loss weight (visual : text)	0.5 : 1	0.5 : 1	0.5 : 1
Warm-up steps	3600	1800	1800
Training steps	90K	90K	90K
Sequence length	5120	5120	65536
Training seen tokens	2.4T	2.4T	7.5T
Resolution	512	512	512
Parallelism Strategies			
Tensor parallelism	4	4	4
Pipeline parallelism	1	1	1
Context parallelism	1	1	4
Data sampling ratio			
Text	0.2	0.2	0.1
Image-Text pair (T2I)	0.64	0.64	0.08
Image-Text pair (I2T)	0.36	0.36	0.02
Video-Text pair (T2V)	0.0	0.0	0.64
Video-Text pair (V2T)	0.0	0.0	0.36

Table 3: Training recipe for different stages of Emu3 pretraining.

Modality	Data Source	Processing Steps & Filtering Criteria
Text	- Aquila corpus [43] - FineWeb-Edu [23]	- Direct use
Image Data	- Open-source Web data [33]	- Extract features using EVA-CLIP [36] - Apply K-means clustering to obtain representative image clusters - Caption generation using our image captioner
Image Data	- Open-source Web data [20, 6, 12] - Open-source Image data [18, 15] - AI-generated data [27, 38] - In-house Emu3 data	- Resolution Filter (remove resolution$<$512×512) - Aesthetic filtering by LAION aesthetic score [20] (remove aesthetic score$<$5.2) - Keep aesthetic score $\geq$5.5 and remove text-rich & monochrome image - Caption generation using our image captioner
Video-Text	- Open-source YouTube videos [26, 42] - In-house Emu3 videos	- Scene-based splitting via PySceneDetect [3] (remove clip length$<$5s) - Removed text-rich clips using PaddleOCR [10] (remove text area$>$1%) - Motion filtering by RAFT [37] (remove optical flow score$<$8) - Aesthetic filtering by LAION aesthetic score [20] (removed aesthetic score$<$5) - Clips captioned using our video captioner

Table 4: Dataset construction including data sources and preprocessing steps.

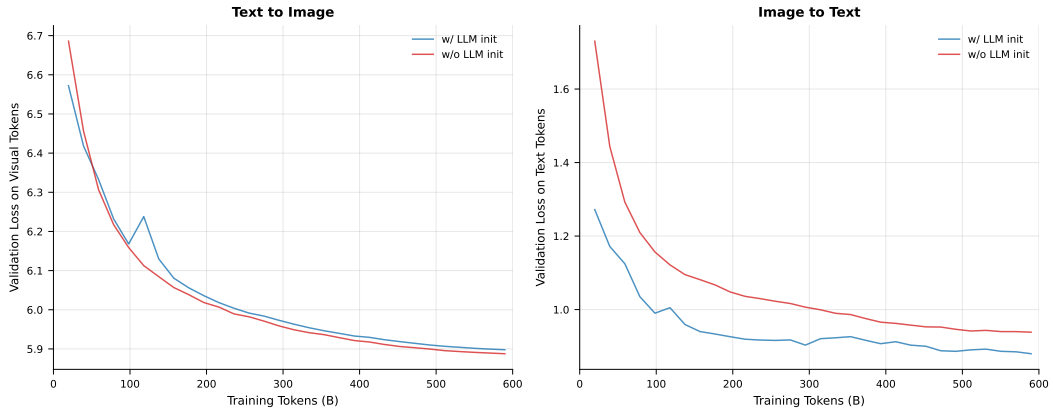


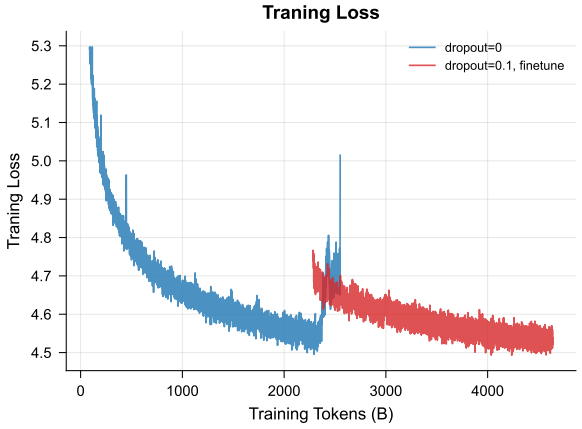
Figure 11: **Ablation experiments with LLM-based initialization on a 1.5B (active) MoE model.** Validation loss for MoE models with/without pretrained LLM initialization. Left: text-to-image generation; Right: image-to-text understanding. Pretrained initialization is detrimental to text-to-image. The MoE model is upcycled from Qwen2.5-1.5B.

	Stage1	**Stage2**	**Stage3**
Training Performance			
Elapsed time per iteration (s)	3.0	3.0	25.1
TFLOP/s per GPU	370	370	391
Equivalent H100 GPU-hours	96k	96k	803k
Model Parameters			
Model size (B)	8.49 (6.98 Transformer + 1.51 Embedding)		
Total Training Cost			
H100-equivalent GPU-hours	995k		

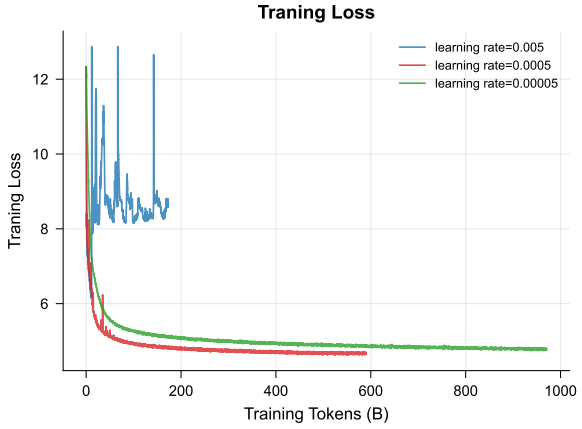
Table 5: **Training computations for different stages of Emu3 pretraining.** Training cost is measured in H100-equivalent GPU-hours, assuming an effective scale of 1280 H100 GPUs.

at batch size 64, while HF peaks at approximately 145 tok/s. In Image-to-Text inference, vLLM consistently delivers 1.6–1.9× higher throughput across batch sizes 1–8 and remains operational up to batch size 128 (20.9s), where HF again becomes out of memory (OOM). For the Text-to-Video task, which involves extremely long sequences (65k tokens), the HF backend fails due to memory limitations, even at batch size 1. In contrast, our vLLM-based system remains stable across all batch sizes, achieving up to 132.52 tokens/s. These results highlight vLLM's superior memory efficiency, thanks to the PagedAttention KV-cache management and its continuous dynamic batching, which together enable larger batches, longer contexts and substantially reduced tail latencies compared to a conventional HF Transformers setup.

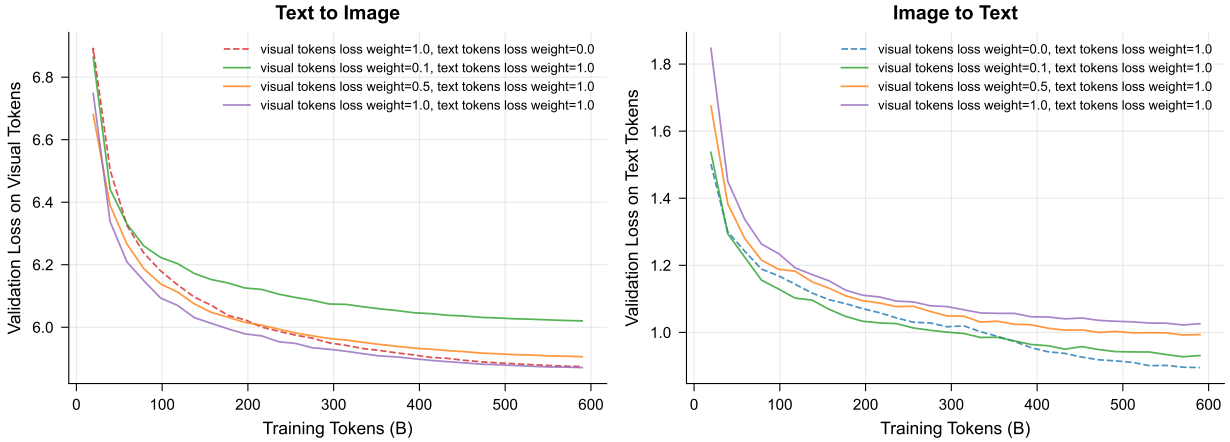
Notably, Fig. 13 illustrates a token-centric multimodal infrastructure that is efficient and extensible, underscoring the practicality and scalability of our framework for large-scale real-world deployment. In this design, data tokenization is performed directly on edge devices, and only the resulting discrete token IDs are transmitted to large-scale servers for unified multimodal training and inference. This approach greatly improves efficiency, as token IDs are substantially more compact than raw multimodal data such as images or videos.



(a) **Effect of dropout on training stability during Emu3 pretraining.** Training loss curves across Stage1 and Stage2 are shown. When `dropout`=0 in Stage1, training becomes unstable and eventually collapses. In contrast, applying a small dropout rate (`dropout`=0.1) and resuming from a checkpoint ensures stable convergence.



(b) **Effect of learning rate on training stability and convergence.** We compare three learning rates: 5×10^{-3}, 5×10^{-4}, and 5×10^{-5}. A large learning rate (e.g., 5×10^{-3}) causes divergence, while a small rate (e.g., 5×10^{-5}) significantly slows down convergence. The setting 5×10^{-4} achieves the best trade-off.



(c) **Ablation on token loss weighting.** We examine the effect of balancing visual and text token losses during pretraining. The curves show validation loss on generation and understanding tasks under different weight configurations. Dashed lines represent settings with one loss weight set to zero, resulting to single-task training.

Figure 12: **Impact of dropout, learning rate, and modality loss weight on Emu3 pretraining.** Proper choices of dropout rate, learning rate, and loss balancing are critical to ensuring training stability and achieving competitive performance in large-scale multimodal learning.

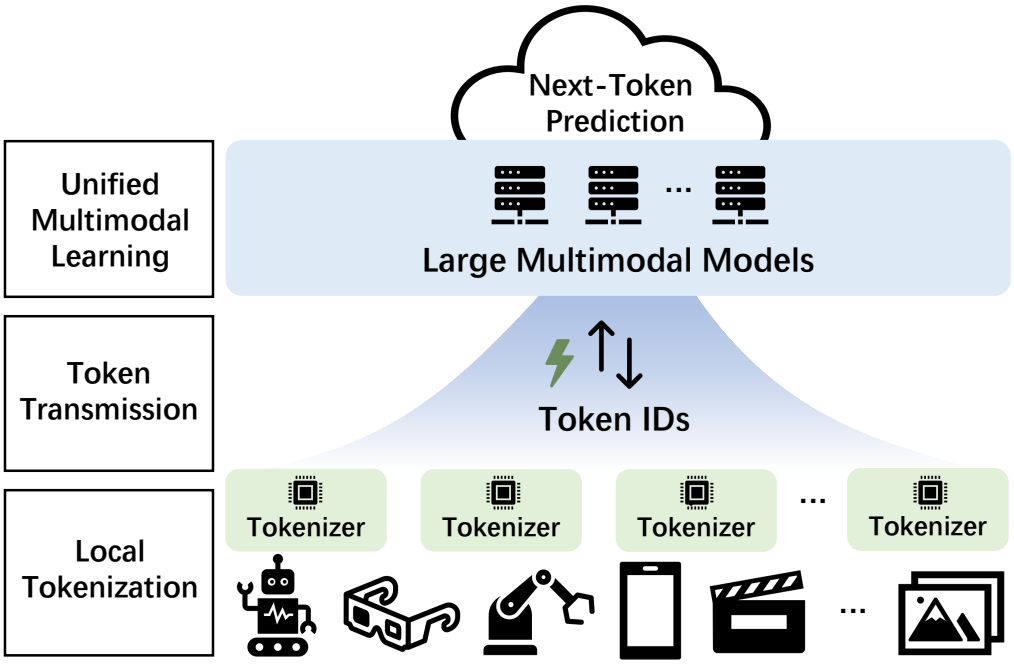


Figure 13: **Token-centric multimodal infrastructure.** Multimodal data tokenization can be performed directly on edge devices, and only the resulting discrete token IDs are compactly and efficiently transmitted to large-scale servers for unified multimodal training and inference.

Token sampling. We investigate the impact of different sampling strategies on generation quality for the text-to-image task. Unless otherwise specified, we adopt classifier-free guidance (CFG) with a scale of 5.0, maintaining default sampling parameters: top-k = 32768, top-p = 1.0, and temperature = 1.0. As shown in Fig. 14a, the performance improves significantly when increasing top-k from 1 to 2048, reaching a maximum overall GenEval score of 0.597. However, as top-k continues to increase, the score gradually declines. This suggests that a moderate top-k is crucial for high-quality generation. In addition, we observe a similar trend when changing the parameter of top-p and temperature as in Fig. 14b and Fig. 14c.

1.2 Responses to General Comment #2

General Comment #2 *How robust and generalizable the framework of next-token prediction is.*

We extend the next-token prediction framework to a range of challenging new capabilities to demonstrate its robustness and generalizability in general-purpose multimodal learning. This includes generalization to vision-language-action model for robotic manipulation, interleaved image-text generation, and flexible prediction paradigms beyond the raster order. We summarize the key capabilities and findings that emerge from scaling unified multimodal learning, highlighting the broad applicability and versatility of the next-token prediction approach.

1. **Vision-language-action model.** We demonstrate that next-token prediction seamlessly generalizes to robotic manipulation by treating vision, language, and actions as unified token sequences. This enables a single model to perform visual forecasting, action generation, and long-horizon planning, achieving state-of-the-art performance on the CALVIN benchmark.

Task	Batch Size	vLLM		HF		Speed Up
		Time (s)	Throughput (token/s)	Time (s)	Throughput (token/s)	
Text-to-Image	1	68.11	120.62	259.69	31.64	3.81
	2	74.16	221.58	303.70	54.11	4.10
	4	82.92	396.34	379.74	86.54	4.58
	8	100.48	654.13	548.44	119.85	5.46
	16	135.22	972.16	905.58	145.16	6.70
	32	220.81	1190.70	OOM	OOM	-
	64	415.70	1264.92	OOM	OOM	-
	128	844.29	1245.59	OOM	OOM	-
Image-to-Text	1	1.17	3687.90	1.92	2252.09	1.64
	2	1.32	6554.10	2.20	3929.00	1.67
	4	1.50	11497.13	2.75	6266.80	1.83
	8	2.09	16541.42	4.00	8624.96	1.92
	16	3.26	21206.78	OOM	OOM	-
	32	5.78	23897.13	OOM	OOM	-
	64	10.15	27209.77	OOM	OOM	-
	128	20.90	26426.17	OOM	OOM	-
Text-to-Video	1	768.29	81.60	OOM	OOM	-
	2	1186.79	105.65	OOM	OOM	-
	4	1986.64	126.22	OOM	OOM	-
	8	3802.48	131.89	OOM	OOM	-
	16	7568.93	132.52	OOM	OOM	-

Table 6: **Comparison of inference latency and throughput between vLLM and Hugging Face Transformers**.

2. **Interleaved image-text generation.** We show that the same next-token prediction backbone supports fine-grained interleaving of image and text tokens. By fine-tuning on recipe-style data, Emu3 can autoregressively generate coherent multimodal sequences conditioned on user instructions, showcasing compositionality across modalities.

3. **Flexible prediction paradigm.** We further validate the robustness of our model by adapting it to various token prediction orders (e.g., diagonal, block-raster, spiral-in). Emu3 transfers effectively across these settings and maintains strong performance, even enabling zero-shot image inpainting capabilities. This underscores the flexibility of next-token prediction beyond fixed raster order.

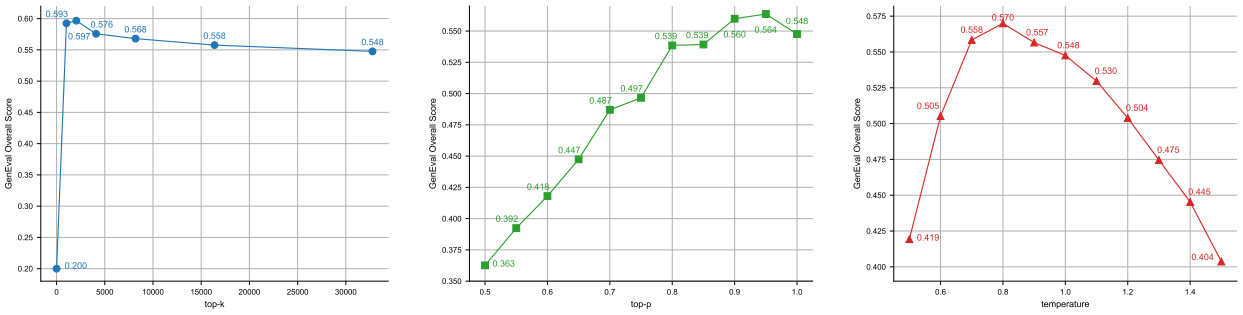
1.2.1 Vision-Language-Action Models for Robotic Manipulation

We apply our framework to robotic manipulation by transferring **Emu3** to a vision-language-action model. Our approach achieves state-of-the-art results on CALVIN benchmark [25] compared to competitive specialized approaches, including RT-1 [4] and RoboVLMs [21].

We integrate language, vision, and action modalities by converting them into discrete tokens, forming a unified multimodal sequence L:

$$L = (L_t^1, L_v^1, L_a^1, L_v^2, L_a^2, \ldots, L_v^t, L_a^t) \tag{1}$$

Here, L_t, L_v, and L_a are discrete token sequences for language, vision, and actions, respectively, sharing the same vocabulary space. Tokens from different modalities are interleaved across time steps t, with special tokens inserted between modalities to indicate modality boundaries.



(a) top-k sampling (b) top-p sampling (c) temperature sampling

Figure 14: **Impact of token sampling parameters.** We show the GenEval overall score of **Emu3** over different sampling configurations. The default settings are cfg = 5.0, top-k = 32768, top-p = 1.0 and temperature = 1.0.

For the robotic manipulation task, a single text instruction is provided at the beginning, followed by an interleaved sequence of visual observations and actions. Under this interleaved autoregressive sequence modeling framework, tasks such as text-to-video generation, visual prediction, and action prediction are naturally formulated as sub-tasks of the same training objective. The actions are tokenized using the FAST tokenizer [29]. By applying frequency-domain encoding, FAST compresses an action chunk of shape $T \times D$ into a variable-length sequence of discrete tokens, depending on the complexity of the action. T denotes the temporal length, and D means the dimension of the action tokens. Since all modality signals are fully transformed into discrete tokens, training is simplified to a next-token prediction task using the standard cross-entropy loss.

Method	**Task**	**Tasks Completed in a Row**					**Avg. Len ↑**
		1	2	3	4	5	
MCIL [24]	ABCD→D	0.373	0.027	0.002	0.000	0.000	0.40
RT-1 [4]	ABCD→D	0.844	0.617	0.438	0.323	0.227	2.45
Robo-Flamingo [22]	ABCD→D	0.964	0.896	0.824	0.740	0.660	4.09
DeeR [41]	ABCD→D	0.982	0.902	0.821	0.759	0.670	4.13
GR-1 [39]	ABCD→D	0.949	0.896	0.844	0.789	0.731	4.21
UP-VLA [44]	ABCD→D	0.962	0.921	0.879	0.842	0.812	4.42
RoboVLMs [21]	ABCD→D	0.967	0.930	0.899	0.865	0.826	4.49
Ours	ABCD→D	**0.985**	**0.959**	**0.928**	**0.899**	**0.870**	**4.64**

Table 7: Long-horizon robotic manipulation evaluation on the CALVIN benchmark.

Dataset. CALVIN [25] is a simulated benchmark tailored for long-horizon, language-conditioned robotic manipulation. It consists of four simulated environments (A, B, C, and D), each containing human-collected demonstration trajectories. The benchmark evaluates performance across 34 distinct tasks with 1,000 unique instruction sequences, assessing the model's ability to execute sequential tasks. Model performance is measured by the average number of completed tasks.

Implementation details. During training, the model is initialized with pre-trained weights from the Emu3, while the action encoder utilizes the FAST tokenizer [29] with a vocabulary size of 1024,

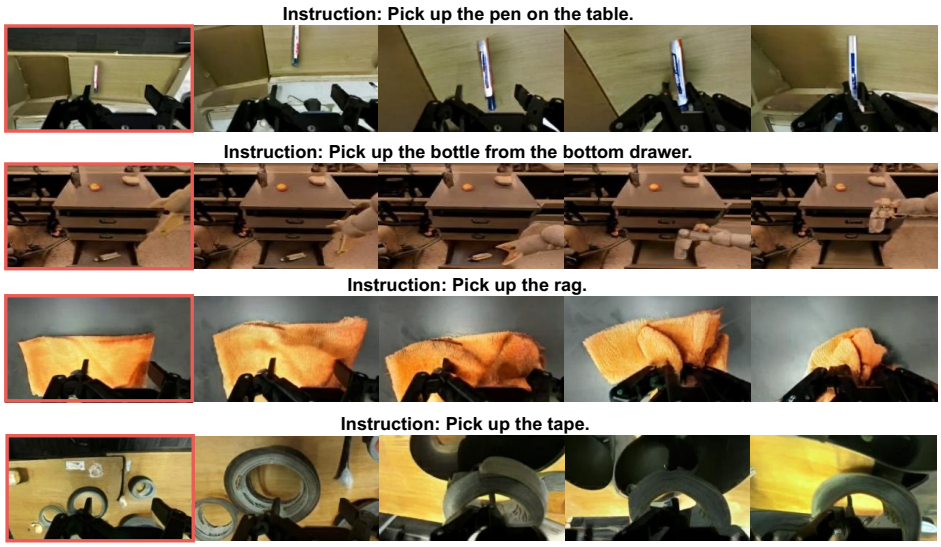


Figure 15: Visualization of visual prediction on the Droid dataset.

replacing the last 1024 token IDs of the language tokenizer. For the CALVIN dataset finetuning, RGB images from both third-person and wrist perspectives serve as observational data. The third-person RGB images have a resolution of 200×200, while the wrist RGB images are 80×80. These images are discretized into tokens using the Emu3 image tokenizer, with a spatial compression factor of 8. During model training, we use a time window of length 20 and set the action chunk size to 10. The input consists of two consecutive VAVA (Vision–Action–Vision–Action) frames. The loss function assigns a weight of 0.5 to visual tokens and 1.0 to action tokens to balance the multimodal learning objectives. Training is performed on NVIDIA A100 GPU with a batch size of 192 over 8k training steps. The learning rate follows a cosine annealing schedule, starting at 8×10^{-5}. During inference, the model operates in an online manner, maintaining a sliding two-frame observation window and predicting the action for the next time step.

Main results. Tab. 7 presents the experimental results in simulation environments. Our method achieves the state-of-the-art performance on ABCD→D tasks, demonstrating superior multitask learning and long-horizon planning capabilities. Fig. 15 illustrates the model's visual prediction capabilities by showcasing predicted future observations conditioned on the language instruction and the first frame.

1.2.2 Interleaved Image-Text Generation

We extend **Emu3** to interleaved image-text generation, where structured textual steps are accompanied by corresponding illustrative images in a single output sequence. Thanks to the flexibility and generability of the framework, we directly fine-tune the model to autoregressively generate such multimodal sequences in an end-to-end manner.

Data Format. In terms of data format for interleaved image-text generation, we construct the instruction tuning dataset as follows, where all contents that should be supervised are included within

two speical tokens (<|extra_100|> and <|extra_101|>). Among them, <|VIS_PLH|> serves as an image placeholder to indicate the actual position of the image tokens.

The Data Format Example of Interleaved Image-text Generation

You are a helpful assistant.
USER:
How to cook Dry Pot Cauliflower?
ASSISTANT:
<|extra_100|>
1. Break the cauliflower into small pieces and soak them in salted water<|VIS_PLH|>
2. Heat a little oil in a pot and stir-fry the cauliflower until it releases its moisture<|VIS_PLH|>
3. In the same pot, stir-fry ginger, garlic, dried chili peppers, and Sichuan pepper sauce until red oil forms<|VIS_PLH|>
4. Add the cauliflower and continue stir-frying. Season with soy sauce, oyster sauce, vegetable essence, and a little salt<|VIS_PLH|>
5. Serve and enjoy!<|VIS_PLH|>
<|extra_101|>

Training Setting. The model was trained for 10k steps with a global batch size of 128 and a maximum sequence length of 33,792 tokens. Each image–text interleaved sequence contained up to 8 images, where each image was resized to a maximum resolution of 512^2 pixels while preserving its aspect ratio. We employed the Adam optimizer with a base learning rate of 7×10^{-6} and a cosine decay schedule, with the minimal learning rate set to 5×10^{-7}. A dropout rate of 0.1 was applied during training, and the loss contributions from image and text modalities were weighted equally.

Visualization. As demonstrated in Fig. 16, even with basic fine-tuning using limited interleaved image-text data, the model exhibits promising capabilities in generating interleaved image-text sequences. This result suggests that our framework possesses greater versatility and scalability compared to conventional single-modality models limited to text or image generation.

1.2.3 Flexible Prediction Paradigm

To further demonstrate the robustness and flexibility of our backbone, we extended Emu3 to different token prediction orders (e.g., diagonal, block-raster, spiral-in) and found that it transfers very fast and generalizes well across these settings. (1) **Raster order**: standard left-to-right, top-to-bottom scan of image tokens. (2) **Diagonal order**: tokens are predicted along diagonals from the top-left to bottom-right, increasing spatial dispersion of context. (3) **Block-raster order**: raster scan within fixed-size local blocks (8·8 tokens per block), followed by sequential traversal of the blocks, which encourages local consistency before global composition. (4) **Spiral-in order**: decoding begins at the image boundaries and proceeds inward in a spiral pattern, prioritizing outer context before filling the center.

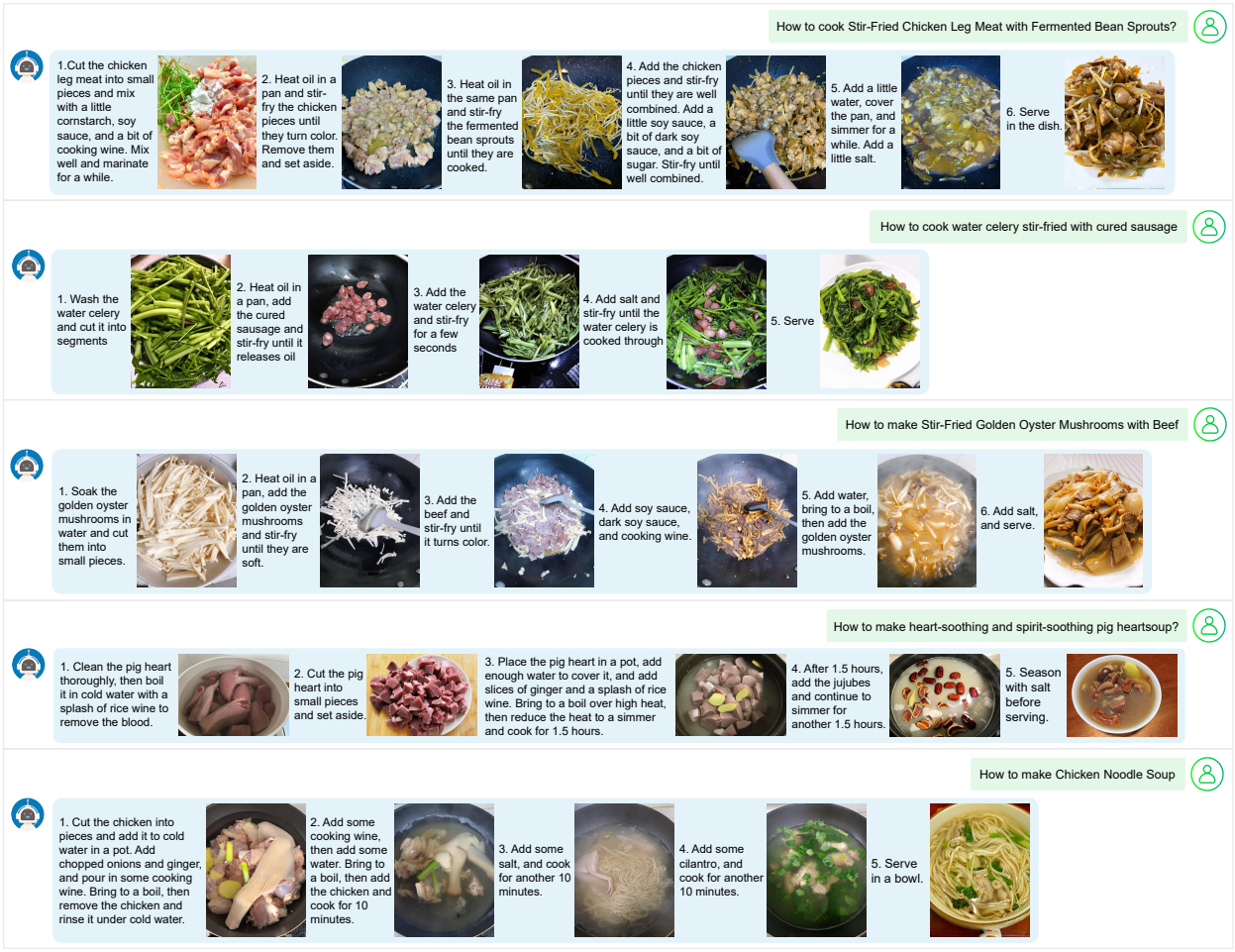


Figure 16: Visualization of interleaved image-text generation results.

Unlike the conventional raster-scan decoding, these alternative token orders modify the autoregressive dependency structure and alter the spatial locality of context tokens, thereby posing a more challenging learning and generalization problem. Our approach leverages the same pre-trained Emu3 model, originally optimized with raster next-token prediction, and fine-tunes it on the new token orders. All transfer experiments follow the training recipe of *Emu3-Gen* for fairness: The sequence length is 5120 tokens, global batch size 512, learning rate 1×10^{-5} with cosine decay and warmup fraction 0.002, and Adam optimizer with weight decay 0.1. Each model is fine-tuned for 50B tokens on data uniformly sampled from the *Emu3-Gen* SFT dataset, with full-token supervision applied throughout.

We observe that Emu3's inductive biases, learned from large-scale raster training, transfer effectively to novel token orders, resulting in significantly higher performance than training from scratch (Tab. 8). We further validate functional utility through zero-shot image inpainting (Fig. 17). The spiral-in order, in particular, aligns with the spatial requirements of region completion, enabling Emu3 to generate coherent fills without task-specific tuning, underscoring the general-purpose adaptability of the backbone.

Order	Initialization	Score (↑)
Raster	Pretrained	0.66
Diagonal	Pretrained	0.61
	From Scratch	0.17
Block-Raster	Pretrained	0.66
	From Scratch	0.18
Spiral-In	Pretrained	0.60
	From Scratch	0.16

Table 8: **Ablation study demonstrating the flexibility of Emu3 across different token prediction orders.** All transfer experiments are fine-tuned for 50B tokens. "Pretrained" denotes initialization from the Emu3 checkpoint trained with raster order. Results show that the Emu3 trained with raster next-token prediction can rapidly adapt and maintain high performance across various token orders, highlighting the generality and robustness of the framework.

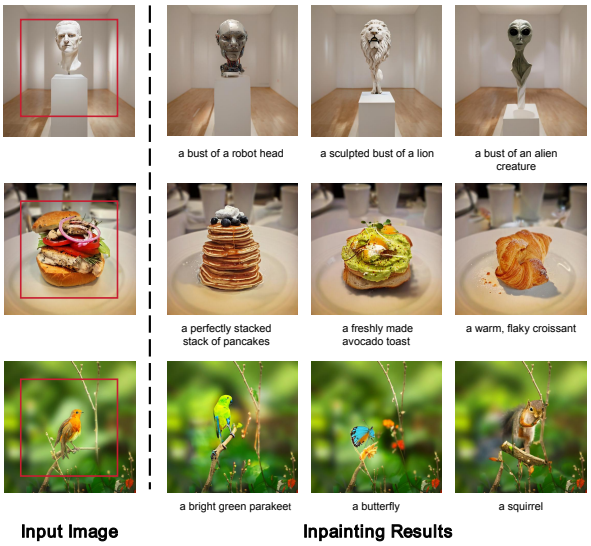


Figure 17: **Zero-shot image inpainting with Emu3 using spiral-in token order.** Given the input images (left of each row) and corresponding prompts, Emu3 accurately fills the masked regions within the bounding boxes, generating semantically aligned content without task-specific fine-tuning.

2 Detailed Responses to Reviewers

2.1 Responses to Reviewer 1

Comment #1.1

The manuscript introduces Emu3, a multimodal model trained using next-token prediction across text, images, and video tokens. It claims state-of-the-art performance on various benchmarks, surpassing task-specific models in image and video generation as well as vision-language understanding. The work's attempt to unify multimodal learning through next-token prediction is innovative and holds potential significance.

Response to #1.1

We sincerely thank the reviewer for acknowledging the innovation and potential impact of our work in unifying multimodal learning through next-token prediction. We have made substantial experiments according to the suggestions and response below point by point. For example, to further demonstrate the generality, we have extended Emu3 beyond traditional benchmarks such as embodied robotic manipulation. These tasks require not only perception and generation across modalities (e.g., vision and language), but also grounded action understanding and decision-making in dynamic environments. Emu3's ability to handle such complex tasks with a unified, autoregressive framework highlights its potential to serve as a general-purpose foundation for real-world multimodal tasks. This extension substantiates our central claim that next-token prediction alone can drive scalable and effective multimodal learning across domains.

Comment #1.2

While the manuscript presents an interesting attempt to apply next-token prediction to unified multimodal modeling, its contributions are significantly limited by the choice of baselines, underexplored technical challenges, and unsubstantiated claims. To strengthen the impact of this work, future revisions should include:

1. Comparisons with state-of-the-art models across all tasks. Choice of Baseline Models: The comparison with baseline models in image and video generation is unconvincing... Weak Video Generation Results: The paper claims competitive video generation performance, but the provided results do not support this claim...

Response to #1.2

We have added the most recent baseline models for each task as a comparison, including the Hunyuan-Video, Wan2.1 for video generation, Flux for image generation, Qwen-2.5-VL for vision-language understanding, and RoboVLM for robotic manipulation. Note that *outperforming each task-specific state-of-the-art model is not the main goal of our work*. These specialized paradigms have been extensively optimized by the community over years, relying on sophisticated engineering pipelines and tailored data, making them difficult to generalize or unify across modalities. In contrast, our work demonstrates that a dramatically simpler and more unified approach based solely on next-token prediction, without task-specific architectures or hand-crafted losses, can effectively handle diverse modalities, including vision, language, and action. More importantly, our model outperforms the state-of-the-art task-specific baseline on the robotic manipulation benchmark, a domain where existing image and video generation models fall short. We believe this result highlights the potential of unified multimodal modeling not just for perceptual generation but also for enabling real-world multimodal agents capable of perception, reasoning, and interaction.

Comment #1.3

2. More rigorous experiments to explore critical design choices (e.g., codebook size, integration of pretrained LLMs).

Response to #1.3

Thank you for the suggestion. We have conducted comprehensive ablation studies to rigorously examine key design choices in our framework. These include:

- **Effect of codebook size:** See Sec. 1.1.2 ("Impact of codebook size"), Fig. 5, Tab. 1, and Fig. 6;
- **Integration of pretrained LLMs:** See Sec. 1.1.4 ("LLM-based initialization and MoE architecture"), Fig. 10 and Fig. 11;
- **Unified vision tokenizer:** See Sec. 1.1.2 ("Comparison between unified tokenizer and standalone image tokenizer"), Tab. 2, and Fig. 7;
- **Variations in model architectures:** See Sec. 1.1.3 ("Architectural comparisons with diffusion models"), Fig. 8; Sec. 1.1.3 ("Architectural comparisons with encoder+LLM compositional paradigm"), and Fig. 9;
- **Training recipe for stable and scalable training:** See Fig. 12 and Sec. 1.1.4;
- **Inference strategies:** See Fig. 14 and Sec. 1.1.5 ("Training and Inference Details").

These additional experiments provide deeper insights into the design choices and help validate the effectiveness and generality of our approach.

Comment #1.3.1

Underutilization of Pretrained Language Models: While next-token prediction is promising, as evidenced by the success of large language models (LLMs), this work opts to train text modeling from scratch rather than leveraging pretrained LLMs. Incorporating pretrained parameters from advanced LLMs could enhance text modeling performance and efficiency. Moreover, the manuscript fails to explore strategies for adapting pretrained LLM backbones for visual data tasks, such as using a mixture-of-experts (MoE) architecture where specific experts handle pretrained text and visual tasks independently before integration.

Response to #1.3.1

We appreciate the reviewer's suggestion. While the effectiveness of next-token prediction in large language models (LLMs) is well-established, our work focuses on exploring whether this paradigm can serve as a unified and scalable solution for multimodal learning, beyond language alone. To this end, we deliberately avoid incorporating strong priors from pretrained LLMs in our primary experiments, so as to clearly evaluate the capability of next-token prediction from scratch in a multimodal setting.

That said, we also conducted controlled experiments with LLM-based initialization. As shown in Fig. 10, we show that using a pretrained LLM to initialize the language component of our model does provide a better starting point in early training stages (e.g., faster initial loss reduction). However, as training progresses, the gap between models with and without LLM initialization diminishes. Both converge to comparable validation loss, and the later even converges better on vision token loss. This indicates that our multimodal learning is robust and effective even without

relying on pretrained language representations. While incorporating LLMs may offer efficiency benefits in the short term, Emu3's design can scale effectively from scratch, supporting its potential as a general-purpose, unified multimodal learner.

We also explored a mixture-of-experts (MoE) architecture variant in our framework We conducted controlled experiments with MoE architectures to further validate our findings. Fig. 11 presents validation loss trajectories for Text-to-Image (T2I) and Image-to-Text (I2T) tasks in MoE models, providing additional empirical evidence for our architectural decisions. In fact, loading pretrained models can be detrimental to text-to-image tasks, as it introduces biases that hinder the model's ability to learn effective visual representations from scratch. Furthermore, our investigation into MoE architectures reveals significant efficiency limitations. Our implementation, which uses 4 experts with 1 activated per token, requires a 4x larger total parameter count in its expert layers compared to a dense equivalent to match the number of activated parameters. This design increases memory requirements, communication overhead, and training complexity.

These comparisons, combined with the observed asymmetry between understanding and generation tasks, support our design choice of a unified dense architecture with random initialization. Our comprehensive empirical analysis demonstrates that the unified next-token prediction paradigm, when properly scaled and trained from scratch, provides a robust and efficient foundation for general multimodal learning, eliminating pretrained initialization or complex architectural modifications.

Comment #1.3.2

Inadequate Vocabulary Size for Vision Tokenizer: The vocabulary size of the Vision Tokenizer (32,768 tokens) appears insufficient for capturing the complexity of visual data. Additionally, the compression performance achieved by the proposed Vision Tokenizer is weaker than that of other state-of-the-art video VAEs.

Response to #1.3.2

We appreciate the reviewer's observation regarding the vocabulary size and compression performance of our vision tokenizer. We conducted thorough experiments with varying codebook sizes on ImageNet [32] 256×256 in terms of both reconstruction and generation tasks. As shown in Fig. 5 and Tab. 1, increasing the codebook size consistently improves reconstruction metrics such as rFID, PSNR, and SSIM, up to a size of 32,768. However, further enlarging the codebook size (e.g., to 65,536) results in degraded rFID performance (e.g., 0.388 to 0.417) and lower codebook utilization (e.g., 68%). This suggests that scaling up the vocabulary is non-trivial and can introduce instability or inefficiency during training. Notably, the best gFID is also achieved with the codebook size of 32,768 on class-to-image generation task. More training details can be found in Sec. 1.1.2.

Besides, we also compare our video tokenizer with the image counterpart. As shown in Tab. 2, we evaluate the reconstruction performance on UCF-101 [34]. With the same input resolution, our video tokenizer achieves comparable rFVD (27.893 vs. 26.675) and PSNR (27.546 vs. 30.499) while using $4\times$ fewer tokens. Moreover, when using the same number of latent tokens, our video tokenizer significantly outperforms the image tokenizer, especially in terms of rFVD (27.893 vs. 139.930), demonstrating both its efficiency and effectiveness.

Designing an ideal tokenizer that is both highly compact and information-preserving remains an open and underexplored challenge, especially in discrete token space. While continuous-space VAEs have benefited from years of optimization and are now highly mature, discrete vision tokenizers have received comparatively less attention due to the dominance of diffusion-based ap-

proaches in generative modeling. Our work takes a step in this direction by unifying multimodal learning with next-token prediction based on a discrete vision tokenizer, which already demonstrates strong performance across perception, generation, and interaction tasks. We hope this serves as a baseline and motivates further research into building more effective vision tokenizers for multimodal learning.

Comment #1.3.3

Lack of Exploration of Key Technical Challenges: The manuscript inadequately addresses several critical challenges in developing unified multimodal models, including: (a) The inherent modality gap between image/video data and text data, which requires sophisticated integration strategies beyond simple token concatenation;

Response to #1.3.3

We thank the reviewer for raising this point. Bridging the modality gap between modalities and enable stable optimization is indeed one of the core challenges in unified multimodal modeling. Large-scale unified multimodal learning is highly sensitive due to the diverse distributions of multimodal data. Improper recipe easily leads to training collapse, underscoring the fundamental difficulty of stable optimization at scale. We conducted a series of experiments to explore training recipes that enable stable and scalable learning across modalities. These include dropout configurations, balancing token loss weight from different modalities, and learning rate schedules. More training details such as training stages, parallelism strategies, loss weights, optimization hyperparameters, and training steps can be found in Tab. 3

Multimodal dropout for training stability. During the pretraining, we initially trained the model with a dropout rate of 0. However, as shown in Fig. 12(a), the training became unstable and eventually collapsed in a later period. We conducted extensive ablations to address this issue, exploring z-loss, adding masked tokens, reducing the learning rate for fine-tuning, skipping optimizer state loading, label smoothing, and packing-based training. None of these strategies effectively resolved the instability. Only after introducing a small dropout rate of 0.1 and resuming from a stable checkpoint did training recover and proceed smoothly. This result highlights the critical role of dropout in maintaining stable convergence during large-scale multimodal learning.

Token loss weight. We conduct ablations on the weights of visual and text token losses during pretraining. As shown in Fig. 12, visual-only or text-only supervision optimizes task-specific performance at the expense of cross-modal generalization. Joint supervision (visual: 1.0, text: 1.0) achieves the best generation performance, demonstrating the benefit of unified learning. However, it underperforms on understanding compared to the text-only setting. To balance both tasks, we adopt a moderate configuration (visual: 0.5, text: 1.0), which achieves competitive results on both tasks. This ablation underscores the necessity of careful loss balancing in large-scale multimodal learning, where naive parameter choices can lead to degraded performance or task bias.

Learning rate. We compare three different learning rates: 5×10^{-5}, 5×10^{-4}, and 5×10^{-3}. We observe that an overly large learning rate (e.g., 5×10^{-3}) quickly leads to unstable training and collapse, while a smaller one (e.g., 5×10^{-5}) greatly slows down convergence. In contrast, 5×10^{-4} achieves a good balance between training stability and efficiency, and is adopted as the default setting in our experiments.

Furthermore, as shown in General Comment #1 (Sec. 1.1.1) and Fig. 1, 2, 3, 4, we observe a clear scaling law in unified multimodal training, suggesting that once the model is stably trained at scale, it can effectively learn modality alignment and achieve strong performance, even on complex tasks that previously required specialized architectures such as vision-language understanding, vi-

sual generation and robotic manipulation. These findings support our central claim: next-token prediction, when properly scaled and trained, can serve as a simple yet powerful mechanism without requiring complicated fusion strategies.

Comment #1.3.4

(b) Architectural considerations for building a robust multimodal backbone.

Response to #1.3.4

We thank the reviewer for raising this important point. We conducted a series of experiments to explore architectural choices for building a robust and effective multimodal backbone. Specifically, we compare our Emu3 architecture with several widely adopted designs and demonstrate the generality and robustness, including: (1) architectural comparisons with diffusion models.; (2) architectural comparisons with encoder+LLM compositional paradigm; and (3) fexible prediction paradigm..

Architectural comparisons with diffusion models. To fairly compare the next-token prediction paradigm with diffusion models for visual generation tasks, we use Flan-T5-XL [8] as the text encoder and train both a 1.5B diffusion transformer [28, 7] and a 1.5B decoder-only transformer [40] on the OpenImages [18] dataset. The diffusion model leverages the VAE from SDXL [30], while the decoder-only transformer uses the video tokenizer in **Emu3** to encode images into latent tokens. Both models are trained with identical configurations, including a linear warm-up of 2,235 steps, a constant learning rate of 1e-4, and a global batch size of 1,024. As shown in Fig. 8, the next-token prediction model consistently converges faster than the diffusion counterpart under equal training samples, underscoring the potential of next-token prediction as a more data-efficient framework for visual generation.

Architectural comparisons with encoder+LLM compositional paradigm. To fairly evaluate different vision-language architectures, we compare four model variants with comparable parameter counts and compute budgets on the image-to-text validation set (i.e., an image understanding task), as shown in Fig.9. The models compared are: (1) a decoder-only model that consumes discrete image tokens as input (Emu3 variant, 1.22B parameters); (2) an early-fusion decoder-only model using continuous image patch embeddings (EVE-style variant, 1.17B); (3) a late-fusion architecture comprising a vision encoder and decoder (LLaVA-style variant, 1.22B = 1.05B decoder + 0.17B vision encoder); and (4) a late-fusion architecture initialized with a larger CLIP-based vision encoder (LLaVA-style variant, 1.35B = 1.05B + 0.30B).

The above models are trained without any pretrained LLM initialization. FLOPs are estimated as $6ND$ for early-fusion models and $6(N_v D_v + ND)$ for late-fusion models, where N denotes the number of transformer decoder parameters, D the number of multimodal tokens, N_v the number of vision encoder tokens, and D_v the number of vision encoder parameters.

Notably, when trained from scratch, the presumed advantage of the encoder-based LLaVA-style compositional architecture largely diminishes. The decoder-only next-token prediction model not only achieves comparable performance but also converges faster, challenging the prevailing belief that encoder+LLM architectures are inherently superior for multimodal understanding.

While the EVE-style early-fusion model also adopts a decoder-only architecture, it lacks generative capability. In summary, the next-token prediction architecture supports both understanding and generation, and can match or even surpass compositional encoder+LLM paradigms in learning efficiency when compared under equal conditions.

Fexible prediction paradigm. To further demonstrate the robustness and flexibility of our backbone, we extended Emu3 to different token prediction orders (e.g., diagonal, block-raster,

spiral-in). As in 17, it demonstrates that the model transfers very fast and generalizes well across these settings.

Comment #1.3.5

(c) The selection of an optimal codebook size for Vision Tokenizer to balance token granularity and computational efficiency. ... The design of unified visual encoders capable of supporting both generation and understanding tasks.

Response to #1.3.5

We appreciate the reviewer's comment. We conducted experiments with different codebook sizes for the vision tokenizer. As show in Fig. 5 and Tab. 1, increasing the codebook size consistently improves reconstruction metrics such as rFID, PSNR, and SSIM, up to a size of 32,768. However, further enlarging the codebook size (e.g., to 65,536) introduces challenges in both tokenizer training and downstream optimization (e.g., lower code utilization (68%) and worse gFID (2.468 to 3.014)). The chosen size (32,768) offers a good balance, achieving strong performance while maintaining stable training.

Besides, we also evaluate the efficiency and effectiveness of our video tokenizer. As shown in Tab. 2, we compare the reconstruction performance on UCF-101 [34] with the SBER-MoVQGAN image tokenizer. With the same input resolution, our video tokenizer achieves comparable rFVD (27.893 vs. 26.675) and PSNR (27.546 vs. 30.499) while using $4\times$ fewer tokens. Moreover, when using the same number of latent tokens, our video tokenizer significantly outperforms the image tokenizer, especially in terms of rFVD (27.893 vs. 139.930), demonstrating both its efficiency and effectiveness.

Comment #1.4

3. A balanced discussion of limitations, technical challenges, and the model's real-world applicability. Overgeneralized Claims about the contributions: The manuscript includes speculative claims about its contributions toward artificial general intelligence (AGI), which are unsupported by evidence. The benchmarks presented are limited to controlled tasks and do not demonstrate broader generalization capabilities. This undermines the credibility of the manuscript's claims and detracts from its otherwise strong empirical focus.

Response to #1.4

We appreciate the reviewer's critical feedback. In response, we have added a dedicated discussion section outlining the key limitations and technical challenges of our work. We have added Sec.7 in our revised paper to discuss the limitations. These include: (1) the efficiency of training and inference, which remains a concern due to the scale of next-token prediction across modalities; (2) the current limitations of our tokenizer in terms of both compression ratio and reconstruction quality; (3) the need for better integration between tokenizer learning and downstream task optimization; and (4) the limited diversity and quality of available multimodal datasets, especially for long-horizon and video-centric tasks.

We also highlight several underexplored technical challenges, such as: developing efficient architectures for ultra-long multimodal contexts, improving tokenizer expressiveness, and establishing stronger real-world benchmarks that go beyond controlled evaluations.

To strengthen the applicability of Emu3, we extended our model to more complex and practical multimodal tasks, including robotic manipulation (as vision-language-action model) and interleaved image-text generation (e.g., generating visual cooking recipes). These additions help ground the model's capabilities in real-world scenarios and demonstrate the generality and effec-

tiveness of the proposed framework for general-purpose multimodal learning.

Finally, we have revised the manuscript to avoid speculative claims related to AGI, and instead focus on our concrete contributions and the demonstrated empirical results. We believe these revisions provide a more rigorous and credible positioning of the work, and we view this line of research as a promising step toward more general-purpose multimodal systems.

2.2 Responses to Reviewer 2

Comment #2.1

The manuscript introduces Emu3, a suite of multimodal models trained exclusively via next-token prediction. By tokenizing images, text, and video into a discrete space and training a unified Transformer, Emu3 achieves state-of-the-art performance in both generation and perception tasks. This approach eliminates the need for diffusion models or compositional architectures like CLIP + LLMs. Key highlights include: (1) Unified Multimodal Learning: Emu3 trains a single Transformer on tokenized video, image, and text data; (2) Performance: It outperforms task-specific models, such as SDXL and LLaVA, on benchmarks like text-to-image generation, vision-language understanding, and video generation; and (3) Open Source: The model and key techniques are publicly available.

Response to #2.1

We sincerely thank the reviewer for recognizing the unified modeling approach, strong performance across tasks, and our open-sourcing efforts. To further demonstrate the generality and practical value of Emu3, we have extended the model beyond standard tasks to challenging new tasks such as robotic manipulation (as vision-language-action model), interleaved image-text generation (e.g., generating visual cooking recipes). These tasks require not only visual and language understanding, but also grounded reasoning and interaction in dynamic environments. Emu3's ability to handle such complex scenarios within a unified, autoregressive framework highlights its potential as a general-purpose foundation model for real-world multimodal tasks. We hope this work not only validates the power of next-token prediction for scalable multimodal learning, but also inspires future research toward simple, unified, and effective multimodal systems.

Comment #2.2

Despite these strengths, I have several major concerns: 1. Lack of Conceptual Novelty: The method is not conceptually new to the community. Unifying generation and understanding by framing the task as next-token prediction has already been explored in prior works, including: [a] Chameleon: Mixed-Modal Early-Fusion Foundation Models, Chameleon Team, ArXiv 2024. [b] Liu et al., World Model on Million-Length Video and Language with Ring Attention, ArXiv 2024. [c] Li et al., SynerGen-VL: Towards Synergistic Image Understanding and Generation with Vision Experts and Token Folding, ArXiv 2024. [d] Wu et al. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation, ArXiv 2024. [e] Wu et al. Liquid: Language Models are Scalable Multi-modal Generators, ArXiv 2024. From the current manuscript, it is unclear what new insights this work offers compared to these existing references.

Response to #2.2

We thank the reviewer for pointing out these recent related works. We have discussed them in more detail in the Sec. 6 of the revised manuscript. While several recent efforts have explored next-token

prediction in image context, including Chameleon, VILA-U, Liquid, and others, our work differs in both motivation and scope. Many of these approaches extend the language model to enable basic image capabilities. In contrast, Emu3 is designed from the ground up as a multimodal-centric framework, aiming to unify generation, understanding, and interaction across text, image, video, and beyond through a single autoregressive Transformer.

Furthermore, our work is the first to demonstrate that next-token prediction alone, without diffusion models or separate perception/generation modules, can achieve state-of-the-art performance on both visual generation and understanding tasks, while also scaling to challenging tasks such as robotic manipulation and interleaved generation. For example, compared to Chameleon, which is limited to mixed text-image modeling and performs poorly relative to task-specific models like SDXL and LLaVA, largely due to its LLM-centric motivation rather than building a unified multimodal learner from first principles.

We do not claim to introduce an entirely new modeling paradigm from scratch. Instead, for the first time, we validated and scaled a simple yet underappreciated approach, namely next-token prediction, toward general-purpose multimodal learning. This work offers not only empirical insights but also practical value, demonstrating that a single unified model can rival specialized systems while unlocking new real-world capabilities. We believe it sets a meaningful milestone in the pursuit of scalable and unified multimodal learning.

Comment #2.3

2. Lack of Novel Technical Contributions: The method does not present new technical innovations, but rather appears to be an engineering effort combining existing techniques. For example: [a]The vision tokenizer is based on SBER-MoVQGAN. [b]The architecture for generation and synthesis is similar to that of Llama-2. [c] Multilingual text tokenization uses QwenTokenizer. The manuscript would benefit from clearer justifications of the novel contributions. Without such justifications, it reads more like an integration of existing methods rather than a new innovation.

Response to #2.3

We appreciate the reviewer's concern and would like to clarify the nature of our contributions. While it is true that some components of our system, such as the backbone architecture and tokenizer initialization build upon existing methods, the key novelty of our work lies not in proposing entirely new modules in isolation, but in demonstrating that a simple, unified next-token prediction framework can scale effectively to handle diverse modalities (text, image, video, action) and outperform or rival highly specialized systems across a wide range of tasks.

Many transformative advances in AI, such as AlexNet [17] and GPT [5], did not hinge on introducing entirely new building blocks but rather on reimagining existing techniques under a new, unifying perspective, backed by strong empirical validation. AlexNet unified feature learning at scale, while GPT unified sequence learning at scale. Inspired by how prior milestones simplified learning objectives at scale, our results suggest that a single next-token objective is a viable path for unified multimodal learning across text, image, video, and action. Our contributions include: (1) Unified tokenization that enables joint modeling across image, video, and action spaces, with discrete tokens that support generation, understanding and manipulation. See Sec. 1.1.2 for more results. (2) A general-purpose next-token prediction architecture that scales to diverse tasks, including robotic manipulation, without task-specific heads or fusion modules. See Sec. 1.1.3 for more results. (3) The observation of a scaling law in unified multimodal learning, indicating that performance consistently improves with model and data scale. See Sec. 1.1.1 for more results. (4)

A training recipe that enables stable and efficient optimization across modalities with very different statistics. See Sec. 1.1.4 and Sec. 1.1.5 for more results. (5) Strong empirical results across generation, perception, manipulation benchmarks, and generalizing to flexible prediction paradigms. See Sec. 1.2.1, Sec. 1.2.2, and Sec.1.2.3 for more results.

We believe these contributions go beyond simple engineering integration. They offer new insights into how a unified autoregressive modeling strategy, traditionally reserved for language, can serve as a general-purpose solution for multimodal learning, with implications for both research and deployment. We hope this foundation encourages future research to build more specialized techniques, similar to how advances have emerged atop LLMs following the GPT series.

Comment #2.4

3. Unimpressive Results: The reported results are not particularly compelling. As shown in Tables 4, 5, and 6, Emu3 underperforms compared to existing methods in many of the benchmark settings, let alone when compared to state-of-the-art commercial models. More justification is needed to establish the significance of these results.

Response to #2.4

We thank the reviewer for the thoughtful feedback. We would like to clarify that outperforming every task-specific model on individual benchmarks is not the main objective of this work. Many of the compared models such as DALL-E3 [2], Qwen-2.5-VL [1], and HunyuanVideo [16] are the result of years of focused community effort, often employing highly specialized architectures, task-specific objectives, and carefully curated training pipelines. These models excel at narrow tasks but are difficult to generalize across modalities or transfer to new domains. In contrast, Emu3 adopts a unified and dramatically simpler approach based purely on next-token prediction.

Importantly, we also extended Emu3 to the challenging domain of robotic manipulation and achieved state-of-the-art results on CALVIN benchmark [25] compared to specialized vision-language-action models, as in Sec. 1.2.1. This is a domain where state-of-the-art image or video generation models are not applicable. In addition, Emu3 enables new applications such as interleaved generation (e.g., cooking recipe synthesis) and transferring to flexible prediction orders, as shown in Sec. 1.2.2 and Sec. 1.2.3. These results demonstrate the value of Emu3 as a viable foundation for real-world multimodal systems capable of unified perception, generation, and interaction across diverse modalities.

Comment #2.5

In addition, I have several minor concerns: 1. Comparison with Diffusion and Encoder-Based Models: A detailed comparison with diffusion models and encoder-based architectures across benchmarks and metrics would strengthen the manuscript.

Response to #2.5

We thank the reviewer for the helpful suggestion. In the revised manuscript, we have added detailed comparisons with both diffusion-based and encoder-based architectures. To ensure a fair evaluation, we conducted experiments under comparable settings. The results show that our token-based architecture achieves competitive performance with diffusion models in visual generation and comparable results with encoder-based models in vision-language understanding tasks. A key insight from our analysis is that the often-perceived superiority of encoder-based models largely stems from pretraining on massive-scale image-text datasets (e.g., billions of samples). When trained from scratch under similar conditions, token-based and encoder-based models can achieve

comparable performance; however, encoder-based models lack generation capabilities.

Architectural comparisons with diffusion models. To fairly compare the next-token prediction paradigm with diffusion models for visual generation tasks, we use Flan-T5-XL [8] as the text encoder and train both a 1.5B diffusion transformer [28, 7] and a 1.5B decoder-only transformer [40] on the OpenImages [18] dataset. The diffusion model leverages the VAE from SDXL [30], while the decoder-only transformer uses the video tokenizer in **Emu3** to encode images into latent tokens. Both models are trained with identical configurations, including a linear warm-up of 2,235 steps, a constant learning rate of 1e-4, and a global batch size of 1,024. As shown in Fig. 8, the next-token prediction model consistently converges faster than the diffusion counterpart under equal training samples, underscoring the potential of next-token prediction as a more data-efficient framework for visual generation.

Architectural comparisons with encoder+LLM compositional paradigm. To fairly evaluate different vision-language architectures, we compare four model variants with comparable parameter counts and compute budgets on the image-to-text validation set (i.e., an image understanding task), as shown in Fig.9. The models compared are: (1) a decoder-only model that consumes discrete image tokens as input (Emu3 variant, 1.22B parameters); (2) an early-fusion decoder-only model using continuous image patch embeddings (EVE-style variant, 1.17B); (3) a late-fusion architecture comprising a vision encoder and decoder (LLaVA-style variant, 1.22B = 1.05B decoder + 0.17B vision encoder); and (4) a late-fusion architecture initialized with a larger CLIP-based vision encoder (LLaVA-style variant, 1.35B = 1.05B + 0.30B).

The above models are trained without any pretrained LLM initialization. FLOPs are estimated as $6ND$ for early-fusion models and $6(N_v D_v + ND)$ for late-fusion models, where N denotes the number of transformer decoder parameters, D the number of multimodal tokens, N_v the number of vision encoder parameters, and D_v the number of vision encoder tokens.

Notably, when trained from scratch, the presumed advantage of the encoder-based LLaVA-style compositional architecture largely diminishes. The decoder-only next-token prediction model not only achieves comparable performance but also converges faster, challenging the prevailing belief that encoder+LLM architectures are inherently superior for multimodal understanding.

While the EVE-style early-fusion model also adopts a decoder-only architecture, it lacks generative capability. In summary, the next-token prediction architecture supports both understanding and generation, and can match or even surpass compositional encoder+LLM paradigms in learning efficiency when compared under equal conditions.

Comment #2.6

2. Generalization and Robustness: More analysis is needed on the model's generalization to unseen datasets, as well as its robustness to noisy data.

Response to #2.6

Our evaluation already covers extensive tests on unseen datasets, providing a direct measure of the model's generalization capability. Specifically, the four image generation benchmarks, eight vision–language understanding benchmarks, and the video generation benchmark are all unseen during training, meaning our training process is agnostic to any specific downstream dataset, which is commonly referred to as zero-shot evaluation.

Regarding robustness, the model is trained on large-scale, noisy web data, which inherently tests its ability to learn from imperfect inputs. The strong zero-shot performance across diverse benchmarks demonstrates that it can effectively generalize despite training noise.

Furthermore, we have added new experiments and analyses of transferring to different prediction orders, interleaved multimodal generation, and vision–language–action tasks. These results further confirm that our next-token prediction framework maintains both generalization and robustness in varied and challenging conditions.

Comment #2.7

3. Efficiency: Providing quantitative insights on training and inference efficiency, including computational costs, would be valuable for understanding the model's practicality.

Response to #2.7

Training stages, compute, and data processing. Tab. 3 summarizes the training pipeline, including stage configurations, parallelism strategies, loss weights, optimization settings, and training steps. We estimate the training cost of Emu3 pretraining in terms of equivalent GPU hours on 1280 NVIDIA H100s. The model contains 8.49 billion parameters, with 6.98B in transformer layers and 1.51B in embedding layers, and is trained over three stages.

Stage1 and Stage2 share the same setup and respectively consume 96k GPU-hours, running at 3.0 seconds per iteration and 370 TFLOP/s per GPU. Stage3, which employs longer sequences (length 65,536), is substantially more compute-intensive, requiring 25.1 seconds per iteration at 391 TFLOP/s per GPU, totaling 803k GPU-hours. In total, Emu3 pretraining consumes approximately 995k H100-equivalent GPU hours.

Details of data construction including sources, filtering and preprocessing are provided in Tab. 4.

Inference infrastructure. Our inference infrastructure is built on FlagScale [9], a multimodal serving system developed atop vLLM [19]. FlagScale inherits vLLM's key advantages, including the PagedAttention mechanism for fine-grained KV-cache management and continuous dynamic batching, which enables adaptive, per-token scheduling to maximize GPU utilization and minimize end-to-end latency. It is further accelerated by optimized CUDA kernels such as FlashAttention [11]. On this foundation, FlagScale extends the inference backend to support Classifier-Free Guidance (CFG) [13] for autoregressive multimodal generation. Specifically, we integrate CFG directly into the dynamic batching pipeline by co-feeding conditional and negative prompts within the same batch iteration. This design allows a unified forward pass at each decoding step, enabling correct logit fusion and guidance scaling without disrupting the batching semantics. Crucially, our CFG-aware extension incurs negligible overhead and preserves the low-latency, high-throughput characteristics of vLLM. As a result, FlagScale provides a scalable and efficient inference framework that supports fine-grained generation control across a wide range of multimodal tasks.

Inference latency. Tab. 6 presents a comparative analysis of inference latency and throughput between our vLLM-based backend and a standard Hugging Face (HF) Transformer implementation across three representative tasks: Text-to-Image (Emu3-Gen with CFG, averaging 8,216 tokens per image), Image-to-Text (Emu3-Chat without CFG), and Text-to-Video (Emu3 with CFG, averaging 65,536 tokens per video). All experiments were conducted on a single GPU node equipped with an NVIDIA H100 or its computational equivalent.

For Text-to-Image, vLLM attains a 3.8× speed-up at batch size 1 (68.1s vs. 259.7s) which increases to 6.7× at batch size 16; HF runs out of GPU memory beyond batch size 32. Correspondingly, token throughput under vLLM rises from 120 tok/s at batch size 1 to over 1260 tok/s at batch size 64, while HF peaks at approximately 145 tok/s. In Image-to-Text inference, vLLM

consistently delivers 1.6–1.9× higher throughput across batch sizes 1–8 and remains operational up to batch size 128 (20.9s), where HF again becomes OOM. For the Text-to-Video task, which involves extremely long sequences (65k tokens), the Hugging Face backend fails due to memory limitations, even at batch size 1. In contrast, our vLLM-based system remains stable across all batch sizes, achieving up to 132.52 tokens/s. These results highlight vLLM's superior memory efficiency, thanks to the PagedAttention KV-cache management and its continuous dynamic batching, which together enable larger batches, longer contexts, and substantially reduced tail latencies compared to a conventional HF Transformer setup.

Notably, Fig. 13 also illustrates a token-centric multimodal infrastructure that is both efficient and extensible, underscoring the practicality and scalability of our framework for large-scale real-world deployment. In this design, data tokenization is performed directly on edge devices, and only the resulting discrete token IDs are transmitted to large-scale servers for unified multimodal training and inference. This approach greatly improves efficiency, as token IDs are substantially more compact than raw data such as images or videos.

Comment #2.8

4. Applications and Limitations: A broader discussion of potential applications and limitations of the model in real-world scenarios would be helpful.

Response to #2.8

Targeting real-world applications, we extended our model to more complex and practical multimodal tasks, including robotic manipulation (as vision-language-action model) and interleaved image-text generation (e.g., generating visual cooking recipes). These additions demonstrate the generality and effectiveness of the proposed framework for general-purpose multimodal learning.

We added a dedicated discussion section outlining the key limitations and technical challenges of our work. We have added Sec.7 in our revised paper to discuss the limitations. These include: (1) the efficiency of training and inference, which remains a concern due to the scale of next-token prediction across modalities; (2) the current limitations of our tokenizer in terms of both compression ratio and reconstruction quality; (3) the need for better integration between tokenizer learning and downstream task optimization; and (4) the limited diversity and quality of available multimodal datasets, especially for long-horizon and video-centric tasks.

We also highlight several underexplored technical challenges, such as: developing efficient architectures for ultra-long multimodal contexts, improving tokenizer expressiveness, and establishing stronger real-world benchmarks that go beyond controlled evaluations.

Comment #2.9

5. Architecture Design Details: The manuscript would benefit from more architectural design details, including additional figures to better illustrate the technical components.

Response to #2.9

We appreciate the reviewer's suggestion. In the revised manuscript, we have added additional architectural figures, tables and clarifying descriptions to better illustrate key components of our system, including the vision tokenizer, the unified architecture, and the training pipeline. These additions will make the technical design easier to understand and reproduce.

2.3 Responses to Reviewer 3

Comment #3.1

Summary

This works proposes to tokenize image, videos and text and then utilize next token prediction to learn a single transformer model that can learn all of these modalities across a bunch of computer vision tasks.

Novelty

Next token prediction and autoregressive modeling has been becoming more popular in computer vision. Works like "4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities" and "GiT Towards Generalist Vision Transformer through Universal Language Interface" have explored similar ideas for a subset of computer vision tasks. What sets this work apart is that it extends this to generation tasks too and achieves very competitive results.

Response to #3.1

We thank the reviewer for recognizing our contributions. While prior works have explored autoregressive modeling for subsets of vision tasks, our work differs significantly in motivation, methodology, and empirical results. We have added more analysis and comparison with existing works in the revised manuscript. Moreover, Emu3 extends beyond traditional scenarios to more multimodal tasks such as robotic manipulation and interleaved multimodal generation (e.g. cooking recipes in Fig. 16), demonstrating not only strong performance but also broad applicability of our framework to diverse modalities and task formats.

To our knowledge, Emu3 is the first system to show that a single next-token objective can deliver competitive performance across generation and understanding and extend to manipulation and interleaved generation under a unified architecture.

We would like to highlight the significance of unifying multimodal learning at scale. After *AlexNet in 2012 successfully unified feature learning at scale*, and *Transformers in 2017 successfully unified sequence learning at scale*, multimodal learning still relies on tons of hand-crafted designs. But multimodal learning is crucial for next-generation foundations such as physical AI, as we live and interact in a context of multimodality. Inspired by how prior milestones simplified learning objectives at scale, our results suggest that a single next-token objective is a viable path for unified multimodal learning across text/image/video/action under one model. ***Emu3 successfully unified multimodal learning at scale***, and achieved competitive results compared to highly specialized frameworks optimized by the community for years. Importantly, we also revealed the scientific questions and technical details behind the results. We believe our work unlocks great opportunities, including native multimodal assistants, multimodal world models, multimodal robotic models, and beyond.

Comment #3.2

It is hard for me to evaluate the quality of data provided by the authors as the work does not give enough technical details for a myriad of design choices made. Everything from how the dataset was created, how the various training details like which loss to propagate at what stage, their weights etc. are not revealed.

Response to #3.2

We thank the reviewer for pointing this out. In the revised version, we have added significantly more technical details to improve clarity and transparency. Specifically, we now include:

- **Dataset construction:** A detailed description of the dataset construction process, including data sources, filtering criteria, and preprocessing steps, is shown in Tab. 4;
- **More training details:** Training stages, parallelism strategies, loss weights, optimization hyperparameters, and training steps are shown in Tab. 3 and Tab. 5;
- **More inference details:** Decoding strategies, sampling temperatures, and inference infrastructure, are detailed in Sec. 1.1.5 ("Inference infrastructure" and "Token sampling").

Additionally, we have conducted comprehensive ablation studies to examine key design choices in our framework rigorously. These include:

- **Effect of codebook size:** See Sec. 1.1.2 ("Impact of codebook size"), Fig. 5, Tab. 1, and Fig. 6;
- **Integration of pretrained LLMs:** See Sec. 1.1.4 ("LLM-based initialization and MoE architecture"), Fig. 10 and Fig. 11;
- **Variations in model architectures:** See Sec. 1.1.3 ("Architectural comparisons with diffusion models"), Fig. 8; Sec. 1.1.3 ("Architectural comparisons with encoder+LLM compositional paradigm"), and Fig. 9;
- **Training recipe for stable and scalable training:** See Fig. 12 and Sec. 1.1.4;

These additions provide deeper insights into the design choices and make our work easier to evaluate and reproduce. All relevant information is now included in the revised submission.

Comment #3.3

I don't take away from the author's achievement regarding a unified model across these modalities and tasks. However, for this to be accepted at a scientific venue, we need scientific questions and conclusions.

Response to #3.3

We thank the reviewer for emphasizing the importance of grounding our contributions in rigorous scientific inquiry. At the core of our work lies the central scientific question: *Can next-token prediction serve as a unified and scalable learning objective for multimodal learning?*

To investigate this question, we designed a unified architecture trained via next-token prediction and performed extensive empirical studies across multiple tasks and modalities at scale. Our results lead to several scientific conclusions, which we detail below. We believe these findings advance both the understanding and practice of multimodal learning, and we hope this addresses the reviewer's concerns regarding scientific rigor and insight.

1. **Scaling laws for multimodal learning.** (Sec. 1.1.1). We establish a multimodal scaling law showing that next-token prediction leads to consistent and predictable improvements across diverse tasks, including text-to-image, image-to-text, and text-to-video generation, when increasing model size and data scale. This power-law behavior mirrors the scaling trends previously observed in unimodal language models, and offers evidence that a single learning objective can unify multiple modalities under the same regime.

2. **Unified vision tokenization.** (Sec. 1.1.2) Our systematic exploration of codebook design reveals that a unified vision tokenizer with a moderate codebook size (e.g., 32k) balances reconstruction quality and modeling efficiency. Notably, it achieves comparable performance to specialized image tokenizers, while using $4\times$ fewer tokens for videos. This finding supports the feasibility of a single tokenizer for both spatial and spatiotemporal visual data.

3. **General-purpose multimodal architecture.** (Sec. 1.1.3) We demonstrate that a decoder-only architecture trained from scratch, without relying on pretrained vision or language backbones, can match or outperform traditional modular pipelines and diffusion models, when compared under equal computes. This challenges the conventional belief that strong unimodal priors, late-fusion compositional designs, and diffusion-based models are essential for high-quality multimodal generation.

4. **Training stability at scale.** (Sec. 1.1.4) Through extensive ablations, we identify key training recipe components, such as multimodal dropout and token loss balancing, that are important to stable convergence in large-scale multimodal models. These insights enabled us to stably train a 7B-parameter model from scratch, avoiding common pitfalls such as gradient instability or mode collapse.

Taken together, these contributions provide a cohesive set of scientific insights into the mechanisms and principles that govern unified multimodal learning at scale. We believe this work offers both conceptual and empirical advances of interest to the broader scientific community.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,

Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[9] FlagScale Contributors. Flagscale: A unified meta-framework enabling adaptive heterogeneous computing for the llm ecosystem. `https://github.com/FlagOpen/FlagScale`, 2024. Accessed: 2025-06-26.

[10] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025.

[11] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[14] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang,

Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[20] LAION. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

[21] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024.

[22] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *ICLR*, 2024.

[23] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024.

[24] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[25] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[26] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[27] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Journeydb: A benchmark for generative image understanding, 2023.

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[29] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[31] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations*, 2023.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[35] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[36] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[38] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[39] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2024.

[40] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[41] Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

[42] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16375–16387, 2022.

[43] Bo-Wen Zhang, Liangdong Wang, Jijie Li, Shuhao Gu, Xinya Wu, Zhengduo Zhang, Boyan Gao, Yulong Ao, and Guang Liu. Aquila2 technical report. *arXiv preprint arXiv:2408.07410*, 2024.

[44] Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint arXiv:2501.18867*, 2025.

Reply to the Reviewers

Manuscript 2024-11-24134

Unifying Multimodal Learning with Next-Token Prediction for Large Multimodal Models

Nature

Response to Referee #1

Comment #1.1

About model architecture: the MoE architecture used for comparison is not well-designed and only adopt naïve MoE design for unified model, resulting in an increase in memory requirements, communication overhead, and training complexity. There are more advanced MoE designs can be adopted, such as the generation expert and understanding expert used in Bagel (ByteDance Seed team), but this work ignores them. I believe the current comparison of MoE and LLM-based initialization is unfair, and my main concern remains unaddressed.

Response to #1.1

We appreciate the reviewer's thoughtful observation. We agree that more advanced MoE variants, such as designs that explicitly separate "generation" and "understanding" experts, can be effective in certain domains. However, these designs rely on strong task assumptions, and their benefit does not necessarily transfer to unified multimodal modeling. In vision–language settings, the functional boundary between generation and understanding is often fluid: tasks such as captioning, grounded reasoning, or temporally coherent prediction require both forms of capability simultaneously. As the task space grows, manually defining expert roles can become brittle and may eventually constrain the model's ability to develop emergent specialization.

In this study, our goal is to conduct a controlled comparison between dense and MoE architectures under matched initialization and compute constraints. Therefore, we adopt a canonical and minimally engineered MoE design to avoid introducing additional confounding variables related to routing strategies or handcrafted expert assignments. This allows the comparison to more cleanly reflect architectural and initialization effects.

We have clarified this rationale in Supplementary Information Section 3.2.3.

Comment #1.2

Regarding the "Unified Vision Tokenization," the adopted VQ method of the current unified tokenizer is not significantly different from the original VQ quantization strategy, except that the convolutional decoder has a slightly modified architecture; however, the VQ quantization strategy remains the same. If no additional entropy loss is added to constrain the codebook, the codebook will basically learn to be the cluster center at the end of training and fall into the local optimum. Although the adopted VQ has designed the gradient for hard-assigned operations such as argmin, gradient-based optimization has the "winner takes all" problem, which should be the root cause of the collapse. In one batch, it is impossible for all tokens in the codebook to have gradients. The optimized tokens may be optimized again in the next batch, and gradually become a winner-takes-all situation.

Response to #1.2

We thank the reviewer for this comment. We agree that classical vector quantization (VQ) methods,

including ours, may still face challenges such as local optima and partial codebook utilization during gradient-based optimization. These are well-known limitations inherent to VQ training.

However, the focus of this work is not on proposing a brand new quantization technique, but on demonstrating that a well-trained VQ tokenizer, when integrated into a unified next-token prediction framework, can effectively support large-scale multimodal learning across images, videos, text, and beyond. Our contribution lies in showing that such a unified autoregressive paradigm remains stable, scalable, and competitive across modalities, even with a conventional VQ backbone.

We believe that future improvements in VQ methods (e.g., balanced codebook usage or adaptive quantization) will further improve visual fidelity and efficiency, thereby strengthening the overall potential and scalability of our proposed framework. We also further discuss related challenges in scaling the codebook size in our response to Comment #1.3.

Comment #1.3

About codebook size, the adopted VQ method is still a traditional VQ solution, the utilization rate is only 68% when the codebook size reaches 65k. I don't think the current "Unified Vision Tokenization" is well-designed and well-trained. So I think the current conclusion on codebook size is correct ("The tokenizer with codebook size of 32k largely outperform the one with codebook size of 8k in both the reconstruction (rFID) and generation (gFID). However, performance drops significantly when the codebook size is further increased to 65,536, likely due to the increased difficulty of training with such a large codebook.")

Response to #1.3

We agree that training VQ models with very large codebooks remains an open and challenging research problem. As the reviewer correctly notes, when the codebook size grows to 65k, maintaining high utilization becomes increasingly difficult, and convergence stability is often compromised, a well-known limitation of existing VQ formulations.

In this work, our goal was not to exhaustively solve this large codebook challenge, but rather to study and quantify its effect within the proposed unified multimodal framework. Our findings, that performance improves up to 32k but degrades beyond, highlight precisely this limitation, and we have made this discussion explicit in the revised manuscript.

We believe these observations provide useful empirical evidence for the community and motivate future research on improving VQ efficiency and utilization, which would in turn further strengthen the proposed next-token prediction paradigm. We have added discussions in the revised manuscript section 'Conclusion, limitation and future work' and Supplementary Information section 1.

Comment #1.4

About the performance, the current version has improved in many benchmarks but is still leg behind many advanced unified models, such as Bagel (2025-05), in both generation and understanding. As for generation and editing, the current version lags far behind Qwen-image-edit and Nona banana. I think this work always ignores doing comparisons with more advanced unified work and advanced models.

Response to #1.4

We thank the reviewer for the constructive comments. We agree that new models continue to emerge; however, outperforming every recently released or task-specific model on individual

benchmarks is not the primary goal of this work. Our focus is to demonstrate that a single next-token prediction framework can achieve strong multimodal performance across understanding, generation, and video domains, supported by systematic scaling studies, ablation analyses, and demonstrations of new capabilities such as robotics and interleaved vision–language generation.

We acknowledge that recent models such as Bagel, Qwen-Image-Edit, and Nano Banana report strong results, yet their scopes and methodologies differ substantially. Bagel is a hybrid architecture with diffusion model expert and does not handle videos; Qwen-Image-Edit is a specialized diffusion model only for image editing with curated data; and Nano Banana remains proprietary without public implementation or details.

In contrast, Emu3 presents a unified formulation that scales competitively across diverse modalities. We believe this establishes a solid foundation for future work that can further advance performance by scaling data and computation within this unified next-token prediction paradigm. We have added more discussion with most recent works in the revised manuscript section 'Related work'.

Response to Referee #3

Comment #3.1

The authors have largely addressed my previous concerns. The quality of the paper has been substantially improved with new experiments, scientific insights and extensions to robustness and generality. In terms of novelty, Emu3 is the first in proving this paradigm scales competitively across modalities with rigorous scaling analyses, detailed ablations, and demonstrations on new domains (robotics, interleaved generation). This would be beneficial for further development of large models. Therefore, I feel the paper has reached the bar of nature and I recommend it for publication after addressing the following minor remaining issues

Response to #3.1

We sincerely thank the reviewer for the positive and encouraging feedback, as well as for the constructive comments provided in earlier rounds, which have been invaluable in improving the quality and clarity of our work. We are pleased that the additional experiments, analyses, and extensions have effectively addressed the previous concerns and strengthened the contribution of this study. We greatly appreciate the reviewer's recognition of Emu3's novelty and potential impact, and we have carefully addressed all remaining minor issues in the point-by-point response and in the revised manuscript.

Comment #3.2

The unified tokenizer claims 4× token reduction. Could the authors provide qualitative visualizations of failure cases (e.g., fine-grained textures, high-motion regions) where compression harms fidelity?

Response to #3.2

We thank the reviewer for this helpful suggestion. We have added qualitative visualizations of representative failure cases in Figure 3 of the Supplementary Information, which mainly arise in extreme scenarios with exceptionally dense high-frequency details or pronounced motion patterns.

Comment #3.3

vLLM is reported to give major efficiency gains. How sensitive are these results to different hardware backends (A100, TPU v5, AMD GPUs)? Could there be limitations when deploying beyond NVIDIA GPUs?

Response to #3.3

Thanks to the generality of our framework, our multimodal inference inherits most of the key advantages of existing LLM infrastructures such as vLLM. vLLM's architectural optimizations (e.g., continuous batching, PagedAttention) are largely hardware-agnostic and have shown strong portability in community tests. On AMD GPUs, ROCm-based deployments can achieve competitive throughput when configured properly, while on TPUs, vLLM runs efficiently through the PyTorch XLA backend with minor graph compilation overhead due to static-shape constraints.

Overall, while NVIDIA backends currently provide the most mature ecosystem, vLLM's efficiency advantages are not inherently CUDA-specific. We anticipate continued improvement in ROCm and TPU support as community adoption broadens, enabling broader cross-accelerator deployment in the future.

Comment #3.4

While authors benchmark against open-source SOTA, could authors comment on how Emu3 qualitatively compares with closed commercial systems (e.g., Gemini, GPT-4V, Claude Opus with vision)?

Response to #3.4

We thank the reviewer for the insightful question. Emu3 adopts a unified Transformer trained end-to-end on interleaved text, image, and video tokens under a single next-token prediction objective. In contrast, closed commercial systems such as Gemini, GPT-4V, and Claude Opus rely on proprietary architectures whose multimodal pathways, pretrained encoders, and integration mechanisms are not publicly disclosed. We acknowledge that large commercial systems often exhibit strong performance in certain specialized areas, such as complex visual–mathematical reasoning, fine-grained document understanding, and discipline-informed recognition tasks, benefiting from their curated data pipelines and proprietary optimizations. However, we note that Emu3's fully unified design offers distinct advantages: it uses one shared token space across modalities, supports reproducible and interpretable scaling behavior, and enables native multimodal capabilities, such as video generation, interleaved reasoning-and-generation, and vision-language-action modeling, within a single causal model.

Comment #3.5

The CALVIN benchmark is simulation-based. Could the authors comment on the feasibility of validating Emu3's vision-language-action performance in real-world robotic environments with noisy sensors and delayed feedback?

Response to #3.5

We thank the reviewer for raising this question. While the CALVIN benchmark is simulation-based, Emu3's vision–language–action formulation was designed with real-world deployment challenges in mind. The next-token prediction paradigm naturally conditions on arbitrary-length histories, allowing the model to integrate feedback over time and recover from partial or imperfect sensor inputs, thereby accommodating noisy sensors or delayed feedback. In practice, real-world robotic validation requires substantial data collection (e.g., time-consuming tele-operation or on-hardware rollouts) and system-level engineering efforts to ensure safety, latency guarantees, and

reliable actuation, which makes large-scale evaluation on physical robots difficult within the scope of this work.

Although large-scale physical-robot validation is part of our future work, the simulation results show that Emu3 can model complex, interleaved perception–action sequences without task-specific components, indicating strong potential for transfer to real robotic systems. We have added further discussion of these considerations in the revised Methods section 'Vision–Language–Action Models.'

Response to Referee #4

Comment #4.1

**Acknowledgement of Revisions I must begin by commending the authors for their exceptionally thorough and detailed response to my (and other reviewers') critiques. The revised manuscript and the detailed rebuttal represent a monumental expansion of the original work from an engineering report to a scientific paper.*

Response to #4.1

We sincerely thank the reviewer for the generous and encouraging feedback. We greatly appreciate the recognition of our efforts in strengthening the scientific depth and clarity of the manuscript. The reviewers' constructive comments were invaluable in guiding the substantial improvements made in this revision.

Comment #4.2

Video Generation (Sec 3.3, Table 9) The authors claim Emu3 "demonstrates highly competitive results". Emu3, an autoregressive (AR) model, achieves a VBench score of 80.96. This is a strong score, beating most open-source models (e.g., OpenSora V1.2 at 79.76). However, Table 9 also lists multiple SOTA diffusion-based models that beat Emu3: T2V-Turbo (81.01), CogVideoX-5B (81.61), Kling (81.85), Gen-3 (82.32), and Wan2.1 (83.69).

The claim "highly competitive" is an overstatement. The real story, which is still highly impressive, is that this is likely the first unified, non-diffusion, AR-only model to get this close to SOTA video generation. But it is not "highly competitive" with the best models.

Response to #4.2

We thank the reviewer for the constructive feedback and fully agree with the clarification. Our intent in using "highly competitive" was to emphasize that Emu3, despite being a unified next-token prediction model without diffusion or task-specific modules, achieves results within a narrow margin of state-of-the-art diffusion-based systems. We have revised the wording in the manuscript to more accurately reflect this point, highlighting that Emu3 is the first to demonstrate that the next-token prediction paradigm can scale competitively and reach the performance of well-established task-specific models.

Comment #4.3

Vision-Language Understanding (Sec 3.5, Table 10) The authors claim Emu3 (an encoder-free method) is "notably surpassing its counterparts". A detailed check of the new, comprehensive Table 10 shows the data is mixed.

Emu3: SEEDBench-Img = 68.2, OCRBench = 687, MMMU = 31.6

LLaVA-1.6(HD) (Compositional SOTA): SEEDBench = 64.7, OCRBench = 532, MMMU = 35.1
Qwen2.5-VL (Newer SOTA): OCRBench = 864, MMMU = 58.6
The data shows Emu3 does convincingly beat the flagship LLaVA-1.6 model on SEEDBench and especially on OCRBench (a 155-point margin), which is a very strong result. However, it loses to LLaVA-1.6 on the difficult MMMU reasoning benchmark. Furthermore, the authors' own table shows that newer models like Qwen2.5-VL significantly outperform Emu3. The authors have successfully demonstrated their "from-scratch" encoder-free model can be competitive with, and on some axes superior to, the dominant LLaVA-1.6 compositional model. This is a solid finding. But it is not SOTA, and the claim "surpassing its counterparts" is only true for other encoder-free models, not the field at large.

Response to #4.3

We agree that it is indeed a solid finding and a strong result that Emu3 substantially outperforms LLaVA-1.6(HD) on SEEDBench and OCRBench, while remaining competitive on MMMU, despite being a from-scratch, encoder-free model that does not rely on pretrained vision encoders or compositional alignment with language models.

We have revised the text to clarify this point: the claim has been refined to state that Emu3 "reaches the performance of its counterparts", rather than claiming to surpass all vision–language models across every benchmark, particularly those that integrate strong pretrained encoders and task-specific alignment. This clarification aligns with our ablation analysis on architectural comparisons and underscores the competitiveness of unified next-token prediction framework in vision–language understanding.