

Supplementary Material

A Additional related work

Diffusion Models [6, 13], generate images by approximating the reverse of a stochastic process transforming data to noise, have become the standard approach for text-to-image generation [4]. Diffusion models modify data, e.g., images, through two processes, namely the forward and backward processes. The forward process consists of transforming data into pure Gaussian noise by gradually adding Gaussian noise across multiple diffusion steps. At the same time, the backward process aims to gradually restore the noisy data back to a clear image by training a denoiser. Models typically considered for fine-tuning fall into either the U-Net [11] or Image Transformer (ViT) [9] categories. The trained denoiser is then used to generate images from random Gaussian noise.

B Further Analysis and Considerations

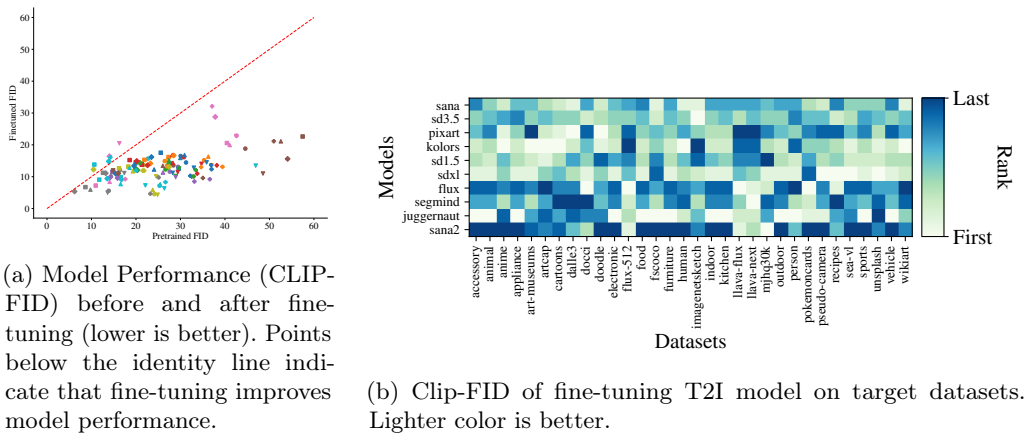


Fig. 5: The necessity and the performance variation on fine-tuning different models on a given target data set.

Before formally defining and addressing the model selection problem, we first investigate its relevance, guided by two key questions: Q1) Can fine-tuning improve quality? and Q2) Can one model outperform all others on all datasets? The former motivates the benefits and necessity of fine-tuning, the latter excludes a trivial solution to the problem.

Table 4: Algorithm performance on diverse splits of the evaluation set. We report the OSR and Kendall τ_w for selected subsets, highlighting what makes ranking prediction harder.

Criterion	Split	OSR $\uparrow$	Kendall τ_w $\uparrow$
Average neighbor distance	lowest 10	80.0 %	0.84
	highest 10	50.0 %	0.62
Performance range	lowest 10	90.0 %	0.93
	highest 10	40.0 %	0.34

B.1 Benefit of Fine-tuning

Figure 5a plots the generated image quality (based on CLIP-FID) before versus after fine-tuning for different datasets (represented by different colors) and different models (represented by different point shapes). The lower the CLIP-FID score, the higher the quality of generated images. As expected, most points lie below the diagonal line, indicating that fine-tuning improved the model’s performance. The figure also underlines that different model-dataset combinations achieve different performance gains, and performance even decreases for a tiny portion of models. However, these are never the best model for the given dataset, and overall, we can answer Q1 that fine-tuning indeed improves quality.

B.2 No Trivial Solution

Having assessed that fine-tuning benefits generation quality to different degrees, one could wonder if there is no trivial solution to the model selection problem. If there exists a model that achieves the best performance across all or most datasets, that model would be a trivial solution to our problem. Figure 5b shows that this is not the case. The heatmap plots the generation quality after fine-tuning for all combinations of datasets (on the x-axis) and models (on the y-axis). The lighter the color, the higher the model’s performance. While some models, e.g., Kolrs or SDXL, obtain on average better quality across the datasets, none are exempt from problematic darker spots. Moreover, the darker spots do not always correspond to the same datasets. For example, Kolrs and SDXL have almost opposite results for datasets like Anime, Cartoons, FS-COCO, Imagesketch, and Pokemoncards. This answers Q2, underlining that a trivial one-model-fits-all solution does not exist and motivates the importance of model selection for achieving the best results.

C Additional Ablation Studies

C.1 Sensitivity to Data Density

We now study the performance of PicTune on different contexts to see how the distribution of models and datasets impact the results. A set were data points are

highly similar enables the algorithm to ground its prediction on close to training data but closeness also makes it harder to distinguish between models.

We then select four splits of our datasets zoo: the 10 datasets closest to one another and conversely the 10 more distant, as well as the 10 datasets where the performance of models is more similar and conversely the more heterogeneous. The results reported in Table 4 show the density of datasets in the training (e.g. a high number of similar datasets) impacts positively the accuracy of the prediction, with +30 percentage points between the 10 closest and the 10 most distant. In the meantime, the degradation of the prediction on the more sparse datasets remains limited, as the τ_w correlation remains high at 0.62 even on the harder setup. Regarding the performance distribution among the models, the gap between tighter and more diverse results is significant, with very strong 90% OSR on the top datasets and poor performance on the bottom ones. This indicates the performance gap is a decisive factor in the ability to correctly predict the ranking.

Overall, these observations comforts our previous analysis that ranking is harder on datasets that trigger outlying performance of some models. This is a positive signal for larger scale operations, suggesting that an higher number of data points would increase normality of the performance distribution and therefore improve prediction quality.

C.2 Alternative construction

Table 5: Evaluation results with various metrics in PicTune. To enable valid comparisons between methods, we report O2B and O2O as a proportion of their respective value ranges. As such, the evaluation based on ImageBind shows very small values because the computed distances are of high magnitude.

Metric	OSR $\uparrow$	Kendall τ_w $\uparrow$	O2B $\downarrow$	O2O $\downarrow$
CLIP-FID	61.3 %	0.51	-1.1 %	6.8 %
ImageBind-FID	56.6 %	0.36	-0.1 %	0.1 %
FID	37.9 %	0.40	1.7 %	6.6 %
KID	30.7 %	0.30	2.4 %	4.3 %
CMMD	48.3 %	0.41	0.9 %	7.0 %

Alternate Quality Metrics Despite remaining the gold standard for T2I evaluation, FID has two major drawbacks. First, it relies on the assumption that the image embeddings follow a multivariate Gaussian distribution. Second, it requires many samples to provide a reliable estimate. Several studies have attempted to address these issues (such as KID [1], FID $_{\infty}$ [3], and CMMD [7]), but they remain less widely used.

We therefore explore the performance of PicTune when using other synthetic image quality measures. In addition to FID, we present results using KID and CMMD: these methods measure the similarity between real and generated images using Maximum Mean Discrepancy (MMD), offering an alternative to Fréchet distance by comparing feature distributions using a polynomial kernel or conditionally, respectively. Moreover, for FID, by default in the rest of the paper, we use the CLIP model for feature extraction, while here we also report the results when using ImageBind and InceptionV3 (as in the original definition) as embedding models.

The results reported in Table 5 show that we maintain consistent performance on the other distance-based metrics. We can surmise that CLIP embeddings are superior in our case since they are specifically designed for T2I generation with a model similar to those often used in the text encoders of our pipelines.

Table 6: Evaluation results when considering different modalities for dataset feature extraction.

Input	OSR $\uparrow$	Kendall $\tau_w \uparrow$
ImageBind	48.1 %	0.50
Sentence-T5	44.4 %	0.48
CLIP	61.3 %	0.51

Alternate Features Extractors We also consider different probe models we can use to construct our datasets features to try to capture properties of different data modalities. We used CLIP as a baseline since it was trained on a large number of image-model pairs and has been commonly used as an image feature extractor [12]. We now also consider two other models to account not only for the images of the datasets. Sentence-T5 [8] is an adapted Text-to-Text transfer transformer (T5) model designed for sentence embeddings. We use it to parse the captions of the different datasets, enabling us to define a dataset feature. ImageBind [5] is a multi-modality model that maps both the text and the images to the same embedding space.

The results reported in Table 6 indicate that the models based on modalities other than images fail to perform as well as CLIP. It can be argued that using an evaluation framework only based on image quality favors an image-only embedding model. We can also advance that, different from CLIP, these models are not designed specifically for the T2I context, and the specificity of image generation prompts may hinder their abilities.

Alternate Ranking Predictors Predicting the rank of a model in PicTune is represented as a classification problem. Therefore, we compare the performance of

Table 7: Prediction Result of different estimators. Notable hyper-parameters are: we used a hidden dimension of 256 for the MLP and 300 estimators for CatBoost.

Input	OSR $\uparrow$	Kendall τ_w $\uparrow$
MLP	45.2 %	0.45
Catboost	38.7 %	0.50
XGBoost	45.2 %	0.46
LimiX	51.3 %	0.51

several common classification algorithms to explore their impact on the prediction accuracy.

Multi-Layer Perceptron (MLP) is the most basic example of a feedforward neural network, composed of a simple sequence of linear layers. Extreme Gradient Boosting (XGBoost [2]) and CatBoost [10] are two more recent tree-based ensemble models that implement gradient boosting, where CatBoost natively handles categorical features.

We report in Table 7 the different results of these models, observing that tree-based predictors, XGBoost and CatBoost, all achieve decent results, as is commonly the case in such a classification setting. However, LimiX [14], which uses a Tabular transformer architecture, fares 6.1 points better. Hence, we make the choice to use LimiX.

References

1. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD gans. In: ICLR, OpenReview.net (2018)
2. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: KDD, pp. 785–794, KDD ’16, Association for Computing Machinery (2016)
3. Chong, M.J., Forsyth, D.A.: Effectively unbiased FID and inception score and where to find them. In: CVPR, pp. 6069–6078, Computer Vision Foundation / IEEE (2020)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
5. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR (2023)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS, vol. 33, pp. 6840–6851, Curran Associates, Inc. (2020)
7. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. In: CVPR, pp. 9307–9315 (2024)
8. Ni, J., Hernandez Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D., Yang, Y.: Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In: ACL, pp. 1864–1874, Association for Computational Linguistics (2022)

9. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV, pp. 4172–4182 (2023)
10. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. In: NeurIPS, vol. 31, Curran Associates, Inc. (2018)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI, pp. 234–241, Springer International Publishing (2015)
12. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K., Yao, Z., Keutzer, K.: How much can CLIP benefit vision-and-language tasks? In: ICLR, OpenReview.net (2022)
13. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML, pp. 2256–2265, ICML’15, JMLR.org (2015)
14. Zhang, X., Ren, G., Yu, H., Yuan, H., Wang, H., Li, J., Wu, J., Mo, L., et al.: Limix: Unleashing structured-data modeling capability for generalist intelligence (2025)