Boosting Synthetic Data for VLMs via Diffusion Noise Optimization

Ren Ohkubo[1,2] Rintaro Yanagi[1] Hirokatsu Kataoka[1,3] Yutaka Satoh[1,2]
[1]National Institute of Advanced Industrial Science and Technology (AIST) [2]University of Tsukuba
[3]Visual Geometry Group, University of Oxford

Abstract

Recent advances in diffusion models have enabled the generation of synthetic images nearly indistinguishable from real ones, making them attractive for dataset construction. However, synthetic images often contain features that differ from those of real images, which can hinder the training of Vision-Language Models (VLMs). In this paper, we propose a method to construct synthetic image datasets that enable more effective VLMs training. The proposed method reduces the gap between real and synthetic images by optimizing the initial noise in diffusion models. Our approach enhances the alignment between text conditions and generated images within the embedding spaces of multiple models, in a plug-and-play manner. This approach also reduces characteristic discrepancies from real images, leading to higher-quality synthetic image data and ultimately improving VLM training. Using the CC3M dataset as a baseline, we generate synthetic datasets conditioned on the same captions. Experiments show that CLIP models trained on our datasets achieve 23.69% Ave. R@1 in zero-shot retrieval and 17.97% in zero-shot classification accuracy on ImageNet-1K, outperforming models trained on naïvely generated data. Furthermore, our method demonstrates strong scalability and sample efficiency—achieving even better performance with up to 40% fewer synthetic images.

1. Introduction

The growing reliance on large-scale image-text datasets has been a driving force in recent advances in VLMs. However, constructing such datasets with real images is costly and often limited in coverage. Synthetic image offers an attractive alternative due to its scalability, controllability, and cost-effectiveness. In particular, diffusion models can produce high-quality images by inputing texts, making them a promising approach for augmenting and constructing datasets to learn image-text correspondences.

However, synthetic images are less effective than real images for training VLMs. One reason is that not all synthetic images exhibit features equivalent to those of real images. For example, Xu et al. demonstrated that variations in initial noise can lead to images with characteristics that deviate from real ones [30]. These findings suggest that some synthetic images may involve both visual and semantic discrepancies compared to real images, potentially impeding model training.

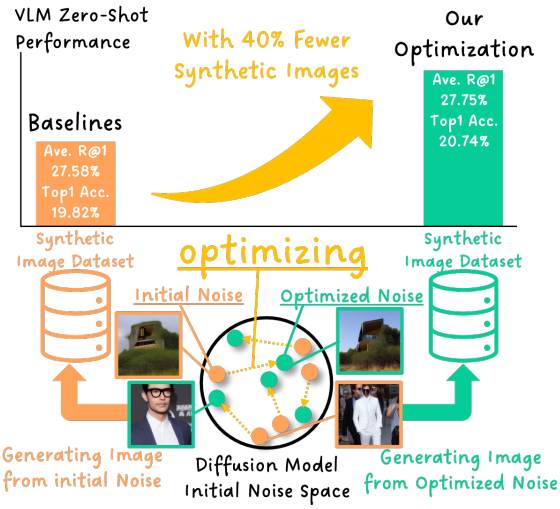

Figure 1. Synthetic image datasets created by our plug-and-play method facilitate more effective training than those created by the original model-based datasets.

To address this issue, we propose a method to construct synthetic image datasets that enable more effective VLMs training. Our approach leverages multiple pretrained embedding models to find an optimal initial noise that maximizes the text-image similarity without modifying the weights of the diffusion model. This optimization is plug-and-play and does not require additional training of the generative or embedding models. The proposed method not only enhances the consistency between the text and the synthetic image, but also reduces the gap between the synthetic and real images to a degree that is recognizable by the embedding models.

We evaluated the effectiveness of our method by comparing CLIP models [23] trained on our synthetic datasets with those trained on baseline Latent Diffusion Model (LDM) [24]-based datasets. Our method achieved the highest zero-

shot retrieval and classification performance, and showed strong scalability—as illustrated in Figure 1—achieving better results with up to 40% fewer synthetic images as illustrated in Figure1. Additionally, FID and cosine similarity evaluations indicate that our images are closer to real images and more semantically aligned with text. These results confirm that our method narrows the gap between real and synthetic images, providing more effective training data.

In summary, the main contributions of this study are as follows:

Key findings. Optimizing the initial noise in diffusion models significantly improves VLM training with synthetic images, reducing dataset construction costs while enhancing alignment with real-world characteristics.

Technical contributions. We introduce a plug-and-play method to optimize initial noise across various embedding models without tuning diffusion model, enabling flexible and broadly applicable synthetic image generation.

Experimental contribution. Our method achieves the highest zero-shot performance across multiple benchmarks and improves scalability. It requires up to 40% fewer samples than conventional approaches.

2. Related Work

2.1. Synthetic Image for Model Training

Diffusion models, originally proposed by Ho et al. [13], generate images by iteratively denoising Gaussian noise in pixel space. Among them, LDM [24] is widely utilized. LDM operates in the latent space of VQ-VAE [28], where Gaussian noise is iteratively denoised before being decoded into an image by the VQ-VAE decoder. Compared to conventional generative models [8, 17], LDM enables high-quality text-conditioned image generation, making it a promising tool for applications, ranging from data augmentation [16, 27, 32] to the generation of training images for downstream tasks such as image classification [11, 12, 26].

Recently, increasing attention has been given to training VLMs using synthetic images. For example, StableRep [26] introduces a framework that extends the InfoNCE loss [21]. This method accommodates multiple positive pairs by treating images generated by Stable Diffusion [24] as pseudo-positive pairs, thereby enhancing zero-shot performance of the CLIP model. Similarly, SynthCLIP [10] employs large language models (LLMs) to generate textual descriptions from a predefined concept bank, then it is used as input for Stable Diffusion. This approach enables the CLIP model training without relying on real images and real text.

However, prior studies assume that synthetic images are both text-aligned and realistic. In practice, naïve diffusion models frequently generate visually or semantically inconsistent images [1, 30], which limits their effectiveness for training. In this work, we challenge the implicit assumption underlying the use of synthetic images for training, and demonstrate that effective training of VLMs can be achieved via "Diffusion Noise Optimization".

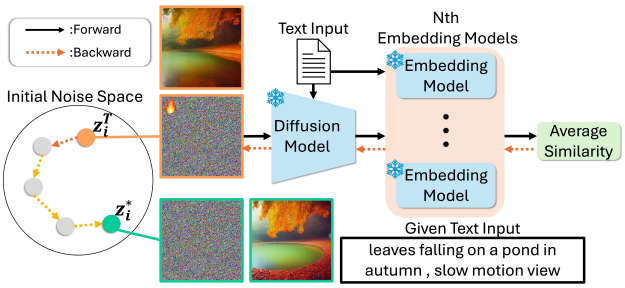

Figure 2. Overview of the proposed method. The initial noise is optimized using multiple embedding models.

2.2. Latent Optimization of Diffusion Model

Diffusion models sometimes produce misaligned images. Recent studies have identified initial noise as a critical factor influencing the quality of synthetic images [22, 30], leading to various approaches for optimizing initial noise.

Eyring et al. [4] proposed a method that leverages multiple reward models reflecting human preferences to evaluate and optimizes synthetic images. Guo et al. [9] introduced an approach that assesses whether a given initial noise can generate high-quality images via self-attention and cross-attention maps extracted from the diffusion process. Qi et al. [22] also proposed a diffusion model inversion-based approach that maps an image back to its initial noise. They found that a higher similarity between the initial noise and the noise obtained after inversion leads to better results, leading to an optimization strategy that maximizes this similarity. However, prior work mainly focuses on human-perceived quality rather than training effectiveness. In contrast, we optimize the initial noise to maximize text-image similarity across multiple embeddings, reducing the gap to real images and improving VLM training. This focus makes our approach fundamentally different from prior methods.

3. Method

3.1. Overview of the Proposed Method

As shown in Figure 2, we propose a method to construct a dataset using synthetic images aligned with the input text while reducing the gap to real images. This is achieved by optimizing the initial noise in a pre-trained LDM with fixed weights. Specifically, the initial noise is treated as a learnable parameter and optimized through a denoising process to maximize the cosine similarity between the input text and the generated image. The proposed method not only adjusts the alignment between text and images but also reduces the gap between real and generated images. This is because images tuned through noise optimization acquire distinguishable features that can be effectively evaluated by embedding

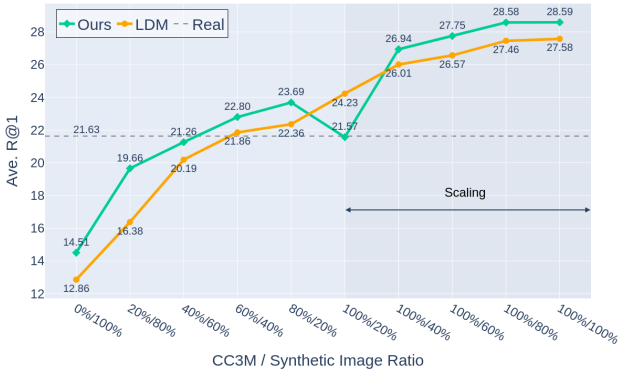

Figure 3. The results of zero-shot retrieval, where we report the average R@1 across several evaluation datasets. The x-axis indicates the proportion of real (CC3M) to synthetic images in the training data.

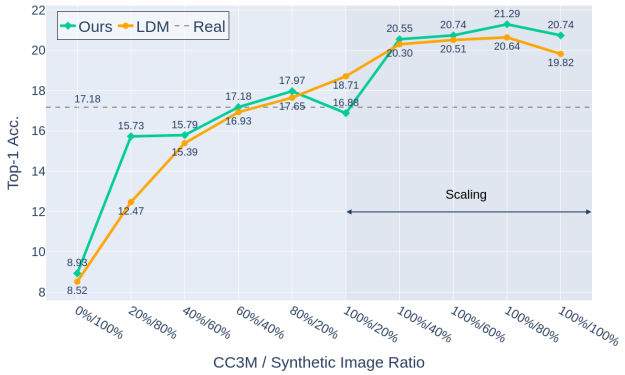

Figure 4. The results of zero-shot classification in ImageNet-1K. The x-axis indicates the proportion of real (CC3M) to synthetic images in the training data.

models trained on real images.

However, embedding models such as CLIP sometimes focus on specific words within the text while failing to capture the overall semantic meaning [29]. To mitigate this issue, we perform optimization across multiple embedding models trained on different datasets. Our method operates in a plug-and-play manner, requiring no retraining of the LDM or embedding models. Furthermore, since our method is based on optimization, it eliminates the need to predefine similarity thresholds, which was required in conventional dataset construction approaches [6].

3.2. Formulation as Noise Optimization

Given a pair consisting of an initial noise $\boldsymbol{z_i^T} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and a corresponding text prompt $\boldsymbol{s_i}$, the denoising process using an LDM with any sampling method R can be expressed as follows:

$$\begin{aligned} \boldsymbol{z_i^0} &= R(\boldsymbol{z_i^T}, \boldsymbol{s_i}) \\ &= R_1\left(R_2\left(\cdots R_T\left(\boldsymbol{z_i^T}, \boldsymbol{s_i}\right)\cdots, \boldsymbol{s_i}\right), \boldsymbol{s_i}\right) \end{aligned} \tag{1}$$

where R_t denotes the denoising step at time t under the sampling method R. Then, the synthetic image $\boldsymbol{x_i}$ can be generated as $\boldsymbol{x_i} = D(\boldsymbol{z_i^0})$, where D denotes the VQ-VAE decoder. Consequently, treating $\boldsymbol{z_i^T}$ as a learnable parameter, the optimal initial noise can be obtained by solving the following optimization problem:

$$\boldsymbol{z_i^*} = \arg\min_{\boldsymbol{z_i^T}} \left(1 - \frac{1}{N}\sum_{j=1}^{N} \mathrm{sim}^j(D(R(\boldsymbol{z_i^T}, \boldsymbol{s_i})), \boldsymbol{s_i})\right) \tag{2}$$

where $\mathrm{sim}^j(\cdot,\cdot)$ represents the cosine similarity measured by the j-th embedding model out of the total N embedding models. By re-inputting the optimized $\boldsymbol{z_i^*}$ and the corresponding text $\boldsymbol{s_i}$ into the LDM, we obtain the optimized image $\boldsymbol{x_i^*}$.

4. Experiments

4.1. Dataset Setup

To evaluate the effectiveness of our method, we use CC3M[25][1] as the baseline real-image dataset. Following [5], which suggests mixing real and synthetic images is effective, we construct datasets with varying real/synthetic ratios using our method and LDM as the baseline synthetic-image dataset. We evaluate models trained on these datasets through zero-shot retrieval (Flickr8K[14], Flickr30K[31], MSCOCO[2]) and zero-shot classification (ImageNet-1K[3]).

4.2. Image Generation and Pre-Training Details

We use a fine-tuned Stable Diffusion-v1.4 trained on LAION-Aesthetics [19] to generate 256×256 images with 10-step DPM++ sampling [20] and a guidance scale of 7.5. Our method further optimizes the initial noise for 50 steps using Adam (lr=0.01). For similarity computation, we use two high-performing CLIP models trained on DFN2B [6] and Datacomp1B [7]. We then pre-train a CLIP model with a ViT-B/16 backbone on the generated datasets for 40 epochs. Full training details are provided in Appendix A.

4.3. Results of Zero-Shot Tasks

We evaluate the effectiveness of our method on zero-shot retrieval and classification tasks using CLIP models trained on datasets containing varying proportions of real and synthetic images. For the retrieval evaluation, we use the average of Image Recall@1 (IR@1) and Text Recall@1 (TR@1), further averaged across multiple bench-

[1] As of December 2024, only 2.2M out of the original 3M samples were available.

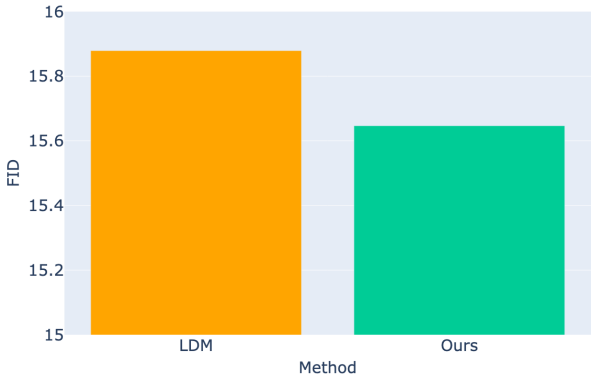

Figure 5. FID of LDM and Our method. **Lower is better.**

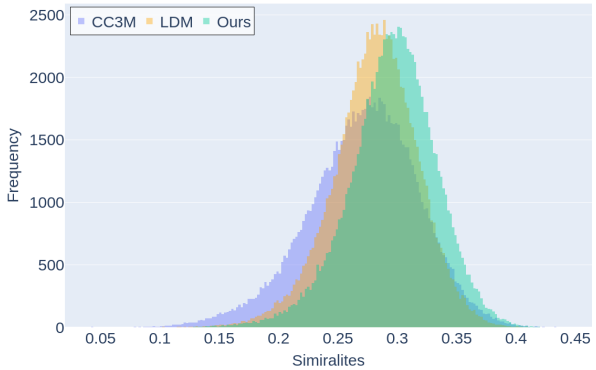

Figure 6. Histogram of cosine similarity between synthetic images and input text.

mark datasets. Detailed results for each dataset can be found in Appendix B. As shown in Figures 3 and 4, without dataset scaling, our method consistently outperforms the LDM-based approach in both retrieval and classification. For both methods, the highest performance was achieved when the CC3M/synthetic image ratio was 80%/20%, surpassing the performance obtained using only real images. Specifically, our method achieved an average R@1 of 23.69% and a Top-1 accuracy of 17.97%, while the LDM-based method reached 22.36% and 17.65%, respectively. These results indicate that our method generates synthetic images that are more effective for VLM training. The ablation study on multiple embedding models is provided in Appendix C.

Additionally, to evaluate the scalability of our method, we conducted experiments in which synthetic images were incrementally added to the CC3M dataset. As shown in Figures 3 and 4, our method achieves the same or higher peak performance as the LDM-based approach with fewer synthetic samples. For example, in zero-shot retrieval, our method achieves 27.75% R@1 at a 100%/60% ratio, surpassing the LDM's peak of 27.58% with 40% fewer synthetic samples. In zero-shot classification, our method achieves 20.74% accuracy at 100%/60%, exceeding LDM's best result of 20.64% with 20% fewer synthetic samples. These findings suggest that our method remains effective even when used to scale existing datasets, providing synthetic images that contribute meaningfully to VLM training.

4.4. Quantitative Evaluation of Synthetic Images

To assess how effectively our method reduces the gap between real and synthetic images, we computed the Fréchet Inception Distance (FID) between CC3M real images and images generated by our method and LDM, using 100K randomly sampled text prompts. As shown in Figure 5, our method achieved a lower FID (15.64) compared to LDM (15.87), suggesting that the images generated by the proposed method are closer in distribution to real images, thereby narrowing the gap between synthetic and real images to a degree that is recognizable by the embedding models. We also evaluated image-text alignment by computing CLIP-based cosine similarity scores across four CLIP models (OpenAI CLIP, LAION2B, DFN2B, DataComp1B). The averaged similarity scores were visualized in Figure 6. Our method consistently yielded higher frequencies of strong text-image similarity compared to both real and LDM-generated images, demonstrating improved semantic alignment and suggesting better suitability for VLM training. Qualitative evaluation and representative examples of the generated images can be found in Appendix D.

5. Conclusion

We proposed a method for constructing synthetic datasets by optimizing the initial noise of a pre-trained diffusion model. Leveraging multiple embedding models in a plug-and-play manner, our approach improves both semantic alignment and realism without additional model training. Experiments show that our method enhances the effectiveness of synthetic data for VLM training. A dataset with 20% optimized synthetic images outperforms training on CC3M and LDM-based datasets, achieving 23.69% retrieval and 17.97% classification accuracy. Moreover, better performance is achieved with up to 40% fewer synthetic images, demonstrating strong scalability. Although our method relies on embedding models, aggregating multiple embeddings improves robustness. These results underscore the potential of optimized synthetic data as a scalable and effective resource for VLM training.

6. Acknowldgements

This work was supported by the policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain" of AIST, Japan. Computation was performed using ABCI 3.0, provided by AIST and AIST Solutions, with support from the "ABCI 3.0 Development Acceleration Use" program.

References

[1] Qingqing Cao, Mahyar Najibi, and Sachin Mehta. Ctrlsynth: Controllable image text synthesis for data-efficient multimodal learning. *arXiv preprint arXiv:2410.11963*, 2024. 2

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *arXiv preprint arXiv:2406.04312*, 2024. 2

[5] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7382–7392, 2024. 3

[6] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 3

[7] 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, 2024. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[9] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. 2

[10] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 2

[11] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20555–20565, 2023. 2

[12] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 3

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1

[16] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024. 2

[17] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[18] LAION-AI. Clip_benchmark. https://github.com/LAION-AI/CLIP_benchmark, 2022. 1

[19] lambdalabs. minisd-diffusers. https://huggingface.co/lambdalabs/miniSD-diffusers. 3

[20] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 3

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[22] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3

[26] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[27] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2

[28] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[29] Hanyao Wang, Yibing Zhan, Liu Liu, Liang Ding, and Jun Yu. Balanced similarity with auxiliary prompts: Towards alleviating text-to-image retrieval bias for clip in zero-shot learning. *arXiv preprint arXiv:2402.18400*, 2024. 3

[30] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. *arXiv preprint arXiv:2405.14828*, 2024. 1, 2

[31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[32] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *Advances in neural information processing systems*, 36:76558–76618, 2023. 2

Boosting Synthetic Data for VLMs via Diffusion Noise Optimization

Supplementary Material

A. CLIP training details

In the experiments described in Section 4.3, we trained the models according to Table 1. All CLIP models used in this study are based on the OpenCLIP implementation [15], and all evaluations are conducted using CLIP Benchmark [18].

Table 1. CLIP pre-training settings.

config	CLIP
epochs	40
batch size	512
optimizer	AdamW
learning rate	5×10^{-4}
weight decay	0.5
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.98$
learning rate schedule	cosine decay
warmup epochs	1

B. Zero-shot Retrieval

The detailed results of the experiments in Section 4.3 are shown in Table 4 and Table 5. In the case without scaling, both the LDM-based method and our proposed method exhibit a decline in IR@1 and TR@1 scores across all benchmark datasets as the proportion of generated images increases. However, our method is more robust to performance degradation compared to the LDM-based approach. For instance, the LDM-based method shows a decrease of approximately 42.48%($\approx$ $(22.36 - 12.86)/22.36$) in Ave. R@1 from its peak performance, whereas our method shows a smaller drop of approximately 38.75($\approx$ $(23.69 - 14.51)/23.69$)%.

As shown in Table 5, when scaling up the CC3M dataset, the proposed method—using only 60% of the CC3M 100% generated images—achieves higher performance on all evaluation metrics except for TR@1 on Flickr30K, compared to using 100% of the CC3M and 100% of images generated by LDM.

Table 2. Zero-shot retrieval results of the ablation study on the effect of using multiple embedding models.

	DataComp1B	DFN2B	Ave.R@1
Ours	✓	✓	23.69
	✗	✓	22.33
LDM	✗	✗	22.36

Table 3. Zero-shot classification results of the ablation study on the effect of using multiple embedding models.

	DataComp1B	DFN2B	ImageNet1k
Ours	✓	✓	17.97
	✗	✓	17.76
LDM	✗	✗	17.65

C. Effect of Using Multiple Embedding Models

We examine the effect of using multiple embedding models. For zero-shot retrieval and classification, we used the same evaluation datasets and trained models using a dataset composed of 80% CC3M and 20% synthetic images, which yielded the best performance without dataset scaling. Results are shown in Tables 2 and 3.

As Table 2 shows, optimizing only with the DFN2B CLIP model resulted in an average R@1 of 22.43%, a 1.36% drop compared to using multiple models, and slightly below LDM. For classification, Table 3 shows a Top-1 accuracy of 17.76%, 0.2% lower than with multiple models, but still outperforming LDM.

These results suggest that a single embedding model can still yield strong performance, particularly in classification. However, using multiple models consistently improves zero-shot performance across tasks.

D. Qualitative evaluation of synthetic images

Upon a detailed examination of the synthetic images, we observed several notable improvements. For instance, in Figure 7(a)-(c), we found instances in which images that initially failed to align with the input text were effectively realigned through our optimization process. Additionally, Figure 7(d) illustrates cases where, even though the original image exhibited good text-image alignment, the proposed method produced an image with improved visual quality. However, we also encountered failure cases, such as those shown in Figure 7(e)-(f), where crucial semantic information from the initial noise was lost. Addressing this issue could further enhance training performance.

Table 4. Results of Zero-Shot Retrieval

Method	Ratio of dataset construction		Flickr8K		Flickr30K		MSCOCO		Ave.
	CC3M	Synthetic Image	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	R@1
	100%	✗	22.95	29.10	21.55	28.60	11.96	15.63	21.63
LDM	80%	20%	23.03	29.39	23.34	29.89	12.12	16.41	22.36
	60%	40%	23.01	30.50	21.34	28.60	12.18	15.53	21.86
	40%	60%	21.08	27.70	20.62	27.30	10.69	13.74	20.19
	20%	80%	17.08	24.09	16.45	21.99	8.11	10.44	16.38
	✗	100%	15.37	18.50	14.33	15.19	6.74	7.00	12.86
Ours	80%	20%	24.14	33.39	22.80	32.19	12.51	17.12	23.69
	60%	40%	23.45	30.89	23.47	31.40	11.53	16.04	22.80
	40%	60%	23.36	30.39	21.11	26.49	11.41	14.82	21.26
	20%	80%	19.56	28.09	20.44	24.79	10.91	14.15	19.66
	✗	100%	17.52	19.59	16.74	17.20	7.68	8.35	14.51

Table 5. Scaling Results of Zero-Shot Retrieval

Method	Ratio of dataset construction		Flickr8K		Flickr30K		MSCOCO		Ave.
	CC3M	Synthetic Image	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	R@1
	100%	✗	22.95	29.10	21.55	28.60	11.96	15.63	21.63
LDM	100%	20%	24.86	33.39	24.71	31.29	13.54	17.57	24.23
	100%	40%	26.60	33.79	26.53	35.40	14.35	19.38	26.01
	100%	60%	27.14	35.10	26.69	36.50	14.61	19.40	26.57
	100%	80%	27.91	36.30	28.72	36.70	14.89	20.26	27.46
	100%	100%	27.82	36.30	27.86	37.70	15.14	20.65	27.58
Ours	100%	20%	23.05	30.50	21.53	28.40	11.23	14.69	21.57
	100%	40%	27.34	35.69	27.50	36.59	14.79	19.74	26.94
	100%	60%	28.13	38.60	28.06	36.00	15.17	20.52	27.75
	100%	80%	28.94	35.60	29.84	38.60	16.08	22.40	28.58
	100%	100%	28.47	36.50	29.30	39.59	15.93	21.76	28.59

(a) people sitting on a bench in street

(b) beautiful blush pink and gray living room - christmas decorating ideas for the home

(c) car driving on snow and ice near the arctic circle

(d) a dog in the snow

(e) a compass lying on a topographic map

(f) national flag above the building

Figure 7. Images generated by the proposed method. The text below each image represents the input text used to generate the image. **Left**: Image generated from the initial noise. **Right**: Image generated from the optimized noise.