Diffusion Noise Optimization for Synthetic VLM Training

Supplementary Material

Table 4. CLIP pre-training settings.

config	value
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

A. CLIP training details

We trained the models according to Table 4. All CLIP models used in this study are based on the OpenCLIP implementation [21], and all evaluations are conducted using CLIP Benchmark [27].

B. Detailed Zero-Shot Tasks Results

We provide the detailed per-dataset results of zero-shot retrieval and zero-shot classification in Table 8 and Table 9, respectively. In zero-shot retrieval, the overall performance decreases as the proportion of real images is reduced; however, with our method, the degradation becomes less severe compared to the original mini-SD and SD-Turbo. A similar tendency can also be observed in zero-shot classification.

C. Detailed Linear Probing Results

Corresponding to the averaged results presented in Section 5.3, Table 7 shows the per-dataset results on downstream tasks. For all models, accuracy tends to decrease as the proportion of synthetic images increases. Nevertheless, mini-SD+ours and SD-Turbo+ours generally outperform their respective baselines (mini-SD and SD-Turbo), mitigating the performance degradation. On the other hand, under the 100% synthetic setting for CIFAR-10 and Aircraft, the improvements for mini-SD are limited, whereas SD-Turbo exhibits clear gains. These findings suggest that the effectiveness of the proposed method depends on both the characteristics of the downstream dataset and the choice of generative model.

D. Detailed Scaling Results

In the main text, we reported the averaged zero-shot retrieval and classification performance under scaling settings. For completeness, we provide the detailed per-dataset results for each task in Table 8 and Table 9. The results are consistent with the averaged scores: performance generally improves as the proportion of synthetic images increases, and our method further enhances both mini-SD and SD-Turbo across datasets.

E. Effect of Time Step on the Denoising Process

To investigate the effect of denoising steps, we compared our proposed method with mini-SD using an increased number of denoising steps, under a setting where 80% of the dataset consisted of real images and 20% of synthetic images.

The results of the zero-shot retrieval experiment are shown in Table 10. When DDIM was used with 50 denoising steps, the average R@1 was 23.19%, while DPM++ with 20 steps yielded 22.87%. In contrast, our proposed method achieved 23.69%, outperforming DDIM (50 steps) by 0.50% and DPM++ (20 steps) by 0.82%. Moreover, when examining individual metrics, our method surpassed mini-SD with increased denoising steps across all evaluations except IR@1 on Flickr30k. These results demonstrate that, in zero-shot retrieval, optimizing the initial noise enables effective learning of high-quality images even with fewer denoising steps.

The results of zero-shot classification are presented in Table 11. In this case, mini-SD with DDIM (50 steps) outperformed our method on all datasets except ImageNet-R. However, since our method supports variable denoising time steps, increasing the number of steps may further improve its performance in classification tasks as well.

F. Distinction from Post-hoc Filtering

As discussed in Section 3.2 and Section 6.2, similarity scores from a single embedding model can be biased toward specific words or visual patterns, and post-hoc filtering based on a single model (*e.g.* CLIP filtering using image-text similarity) often results in unstable or overly selective data retention. While one could conceptually apply post-hoc filtering with multiple embedding models, the chance that an already-generated synthetic image simultaneously satisfies all embedding models decreases rapidly as the number of models increases, making multiple-model filtering impractical at scale.

In contrast, our noise optimization explicitly searches for an initial noise whose generated image aligns with all embedding models before the image is produced. This could lead to steering the generator toward regions of the synthetic distribution that jointly satisfy multiple semantic con-

straints, which post-hoc filtering generally cannot achieve. In this sense, our method is not merely a more complex filtering strategy, but a mechanism that reshapes the generative process itself so that high-quality images emerge naturally without discarding samples.

Table 5. Detailed results of zero-shot retrieval.

Method	Ratio of dataset construction		Flickr8k		Flickr30k		MSCOCO	
	CC3M	Synthetic Image	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1
	100%	✗	22.95	29.10	21.55	28.60	11.96	15.63
mini-SD	80%	20%	23.03	29.39	23.34	29.89	12.12	16.41
	60%	40%	23.01	30.50	21.34	28.60	12.18	15.53
	40%	60%	21.08	27.70	20.62	27.30	10.69	13.74
	20%	80%	17.08	24.09	16.45	21.99	8.11	10.44
	✗	100%	15.37	18.50	14.33	15.19	6.74	7.00
mini-SD+ours	80%	20%	24.14	33.39	22.80	32.19	12.51	17.12
	60%	40%	23.45	30.89	23.47	31.40	11.53	16.04
	40%	60%	23.36	30.39	21.11	26.49	11.41	14.82
	20%	80%	19.56	28.09	20.44	24.79	10.91	14.15
	✗	100%	17.52	19.59	16.74	17.20	7.68	8.35
SD-Turbo	80%	20%	23.63	31.20	22.38	30.70	11.84	16.32
	60%	40%	22.12	31.40	21.53	29.10	11.41	14.93
	40%	60%	19.09	26.19	18.58	23.89	9.27	12.57
	20%	80%	18.19	25.00	17.26	24.69	9.11	12.42
	✗	100%	11.64	15.89	9.70	10.89	5.08	6.10
SD-Turbo+ours	80%	20%	23.96	31.99	22.64	30.39	12.39	15.97
	60%	40%	22.98	30.50	22.54	32.40	12.27	16.06
	40%	60%	21.84	29.89	22.15	30.00	11.46	15.39
	20%	80%	21.19	28.40	20.31	27.50	10.70	15.16
	✗	100%	16.17	21.79	15.09	19.20	7.52	10.05

Table 6. Detailed results of zero-shot classification.

Method	Ratio of dataset construction		ImageNet-1k	ImageNet-A	ImageNet-O	ImageNet-R
	CC3M	Synthetic Image	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.
	100%	✗	17.18	3.93	21.80	19.27
mini-SD	80%	20%	17.65	4.25	19.48	23.30
	60%	40%	16.93	3.90	19.84	22.25
	40%	60%	15.39	3.41	17.47	19.65
	20%	80%	12.47	2.80	13.43	17.59
	✗	100%	8.52	2.02	10.68	12.05
mini-SD+ours	80%	20%	17.97	4.08	20.34	21.60
	60%	40%	17.18	3.89	19.84	21.90
	40%	60%	15.79	3.84	18.19	21.55
	20%	80%	15.73	3.74	18.20	20.15
	✗	100%	8.93	2.42	12.26	13.50
SD-Turbo	80%	20%	17.47	4.16	19.42	23.45
	60%	40%	16.10	3.77	18.52	19.30
	40%	60%	13.02	2.74	13.75	18.14
	20%	80%	13.18	2.86	14.27	19.90
	✗	100%	5.46	1.90	6.47	10.80
SD-Turbo+ours	80%	20%	17.60	4.20	19.56	22.40
	60%	40%	16.86	4.01	19.86	22.10
	40%	60%	16.00	3.18	18.21	21.40
	20%	80%	14.72	3.20	16.73	19.95
	✗	100%	8.53	2.14	10.51	12.90

Table 7. Detailed results of linear-probing.

Method	Ratio of dataset construction		CIFAR-10	CIFAR-100	Aircraft	DTD	Flowers	SUN397	Caltech101	Food101
	CC3M	Synthetic Image	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.
mini-SD	80%	20%	80.50	57.75	20.46	52.82	71.65	68.27	92.03	48.19
	60%	40%	79.41	57.28	20.19	51.33	69.54	66.22	92.75	47.87
	40%	60%	77.17	54.72	19.32	51.33	68.92	64.61	90.75	45.35
	20%	80%	71.43	48.96	16.98	48.56	63.21	60.24	91.71	40.04
	✗	100%	73.12	50.89	18.39	48.46	65.17	58.78	92.47	38.14
mini-SD+ours	80%	20%	78.29	55.85	20.70	54.26	71.05	67.78	91.92	48.11
	60%	40%	79.85	57.75	19.02	54.36	69.56	65.48	93.65	47.60
	40%	60%	78.56	57.38	19.77	53.30	68.42	64.19	96.03	45.96
	20%	80%	77.01	54.73	19.86	52.18	67.13	63.79	95.57	44.01
	✗	100%	75.20	53.50	19.74	50.27	68.69	60.39	93.43	39.61
SD-Turbo	80%	20%	78.43	56.05	14.46	51.33	63.52	67.76	86.64	47.91
	60%	40%	77.27	54.77	14.19	50.85	61.88	65.66	85.06	45.77
	40%	60%	72.55	47.36	12.45	47.23	54.82	60.74	79.07	41.33
	20%	80%	72.80	49.42	14.25	46.86	55.44	60.20	81.23	41.19
	✗	100%	64.67	40.45	13.08	40.00	44.36	52.99	74.29	36.27
SD-Turbo+ours	80%	20%	79.44	57.26	14.13	51.91	62.68	67.65	86.18	47.58
	60%	40%	78.86	56.22	14.76	51.01	60.34	67.14	85.43	47.29
	40%	60%	75.53	52.26	14.61	50.16	61.49	65.31	84.35	45.95
	20%	80%	74.88	51.72	15.75	47.87	60.99	63.26	83.40	44.43
	✗	100%	70.30	46.41	16.50	44.95	51.00	59.21	78.32	39.78

Table 8. Detailed scaling results of zero-shot retrieval.

Method	Ratio of dataset construction		Flickr8k		Flickr30k		MSCOCO	
	CC3M	Synthetic Image	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1
	100%	✗	22.95	29.10	21.55	28.60	11.96	15.63
mini-SD	100%	20%	24.86	33.39	24.71	31.29	13.54	17.57
	100%	40%	26.60	33.79	26.53	35.40	14.35	19.38
	100%	60%	27.14	35.10	26.91	36.50	14.61	19.40
	100%	80%	27.91	36.30	28.72	36.70	14.89	20.26
	100%	100%	27.82	36.30	27.86	37.70	15.14	20.65
mini-SD+ours	100%	20%	23.05	30.50	21.53	28.40	11.23	14.69
	100%	40%	27.34	35.69	27.50	36.59	14.79	19.74
	100%	60%	28.13	38.60	28.06	36.00	15.17	20.52
	100%	80%	28.94	35.60	29.84	38.60	16.08	22.40
	100%	100%	28.47	36.50	29.30	39.59	15.93	21.76
SD-Turbo	100%	20%	23.73	31.40	21.92	30.39	11.68	15.13
	100%	40%	27.57	35.60	26.64	35.10	13.88	19.66
	100%	60%	27.63	36.30	27.57	37.40	14.40	19.61
	100%	80%	27.50	39.50	26.85	37.59	14.93	21.06
	100%	100%	26.73	36.70	27.09	37.00	14.01	19.85
SD-Turbo+ours	100%	20%	25.20	34.99	25.20	32.89	13.49	17.83
	100%	40%	27.32	34.79	27.70	36.19	14.67	20.18
	100%	60%	28.51	37.79	28.08	38.29	15.45	21.14
	100%	80%	29.80	38.99	29.51	38.69	15.84	21.09
	100%	100%	29.51	39.19	30.19	39.19	16.02	22.25

Table 9. Detailed scaling results of zero-shot classification.

Method	Ratio of dataset construction		ImageNet-1k	ImageNet-A	ImageNet-O	ImageNet-R
	CC3M	Synthetic Image	Top1-Acc.	Top1-Acc.	Top1-Acc.	Top1-Acc.
	100%	✗	17.18	3.93	19.27	21.80
mini-SD	100%	20%	18.71	4.55	21.71	22.85
	100%	40%	20.03	4.79	23.44	24.90
	100%	60%	20.52	4.68	24.11	24.80
	100%	80%	20.65	4.76	24.43	24.65
	100%	100%	19.82	4.13	23.02	24.05
mini-SD+ours	100%	20%	16.88	3.83	18.81	19.90
	100%	40%	20.56	5.03	25.04	25.50
	100%	60%	20.78	4.93	25.26	25.40
	100%	80%	21.29	5.25	25.67	24.65
	100%	100%	20.75	4.73	25.37	25.45
SD-Turbo	100%	20%	17.05	3.89	19.10	21.60
	100%	40%	19.57	4.80	22.85	24.55
	100%	60%	19.41	4.48	22.85	24.05
	100%	80%	20.12	4.29	23.31	25.45
	100%	100%	19.72	3.99	22.74	26.20
SD-Turbo+ours	100%	20%	19.23	4.65	22.06	24.20
	100%	40%	19.45	4.33	22.95	23.90
	100%	60%	21.03	4.80	25.24	25.25
	100%	80%	21.08	4.65	25.47	26.55
	100%	100%	21.02	4.59	25.10	26.30

Table 10. Results of the zero-shot retrieval when changing the sampling method and the number of denoising steps.

Method	sampler	Denoising steps	Flickr8k		Flickr30k		MSCOCO		Ave.
			IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	R@1
mini-SD+ours	DPM++	10	24.14	33.39	22.80	32.19	12.51	17.12	23.69
mini-SD	DDIM	50	23.47	31.79	22.92	31.29	12.21	15.53	22.87
mini-SD	DPM++	20	23.00	32.49	22.75	31.99	12.35	16.53	23.19

Table 11. Results of the zero-shot classification when changing the sampling method and the number of denoising steps.

Method	Sampler	Denoising steps	ImageNet-1k	ImageNet-A	ImageNet-O	ImageNet-R	Ave.
mini-SD+ours	DPM++	10	17.97	4.08	21.60	20.35	16.00
mini-SD	DDIM	50	18.08	4.40	22.60	19.65	16.18
mini-SD	DPM++	20	17.89	4.08	22.50	19.33	15.95