Let Triggers Control: Frequency-Aware Dropout for Effective Token Control

Supplementary Material

Junyoung Koh[1*] Hoyeon Moon[1] Dongha Kim[1] Seungmin Lee[1] Sanghyun Park[1] Min Song[1,2]
[1]Yonsei University [2]Onoma AI
solbon1212@yonsei.ac.kr

1. Training Hyperparameters

We summarize the training hyperparameters used in our experiments in Tab. 1 for SD 1.5 and SDXL, and in Tab. 2 for FLUX and Qwen-Image. For the SD 1.5 and SDXL experiments, we train U-Net LoRA using a batch size of 8 and 64 gradient accumulation steps on a single RTX 3090 GPU. We adopt the AdamW optimizer [1] with weight decay $\lambda = 0.01$ and $\beta = (0.9, 0.999)$, and set the learning rate to 5×10^{-5}. For the DiT-based backbones (FLUX and Qwen-Image), we follow their official LoRA training recipes.

Hyperparameter	Value
Train batch size	8
Gradient accumulation steps	64
Max train steps	1500
Min adaptive dropout	0.35
Max adaptive dropout	1.0
Step dropout start	0.1
Step dropout end	0.8
Step dropout warmup ratio	0.1

Table 1. Training hyperparameters for SD 1.5 and SDXL.

Hyperparameter	FLUX	Qwen
Base model	FLUX.1-dev	Qwen-Image
LoRA rank	32	32
Learning rate	1×10^{-4}	1×10^{-4}
Num epochs	2	2
Dataset repeat	50	50
Grad. accum. steps	1	1

Table 2. Training hyperparameters for FLUX and Qwen-Image.

2. Effect of Trigger Token on Inference Performance

Although improving the generation capacity of LoRA models is crucial, it is equally important to prevent catastrophic forgetting of the inherent knowledge of the base model. To investigate this, we conduct an ablation experiment in which we deliberately exclude the trigger token during inference. By omitting the trigger token, we aim to observe whether the model's ability to reproduce the target identity or style diminishes, thereby assessing how tightly the learned features are bound to the trigger token. The quantitative results are reported in Tabs. 3 to 6.

Interestingly, for the Normal Dropout setting, the scores without the trigger token are often similar to—or occasionally even higher than—those with the trigger token. This suggests that character-related features are diffusely distributed across all tokens, rather than being strongly bound to the trigger token. In contrast, for FAD and sFAD, the scores

Normal Dropout sFAD
Prompt: `2 object, no humans, masterpiece, right is character- pikachu, left is character- pochacco, swimming, sunset`

Normal Dropout sFAD
Prompt: `2 object, no humans, masterpiece, right is character- pochacco, left is character- pikachu, astronaut, black hole, rocket`

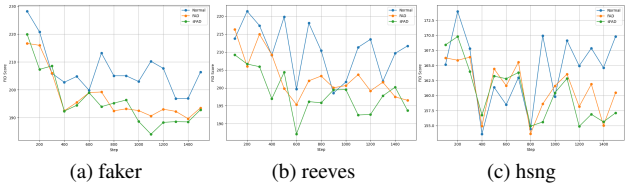
Figure 1. Multi-concept Training: Normal Dropout (left) vs. sFAD (right). The model trained with sFAD successfully reproduces both characters, whereas Normal Dropout results in noticeable deformations.

*Corresponding author

drop drastically when the trigger token is omitted, demonstrating that our method successfully concentrates the critical character features into the trigger token itself. This confirms that FAD and sFAD effectively strengthen the association between the trigger token and the intended identity or style, while Normal Dropout does not.

3. Detailed Example of GPT-4.1 Evaluation

For this evaluation, we follow the evaluation prompt from [2] and the representative images are selected from the model outputs at the 1300th training step. Tab. 8 shows the complete result used for the evaluation, and Tab. 9 presents the complete evaluation result corresponding to Tab. 8.



(a) faker (b) reeves (c) hsng

(d) mbst (e) pikachu (f) pochacco

Figure 2. Comparison of FID (↓) scores of three methods across all steps for six datasets, extracted from **SD 1.5**.

4. Dropout Scheduling Strategies

In Tab. 7, the proposed sFAD method employs an exponential scheduling function to gradually increase the dropout rate during training, ranging from 0.1 to 0.8. To further validate the robustness of our approach, we also conduct experiments using a linear scheduling strategy.

Beyond the choice of scheduling function, we investigate how the magnitude of dropout probabilities should vary across diffusion timesteps. Specifically, we define two distinct dropout ranges: (1) $0.1 \rightarrow 0.8$ and (2) $0.8 \rightarrow 0.1$. In the first setting ($0.1 \rightarrow 0.8$), training begins with a lower dropout rate to prioritize generalization in the early timesteps, and the rate gradually increases to promote disentanglement between the trigger token and surrounding tokens, following the principles of FAD. Conversely, in the second setting ($0.8 \rightarrow 0.1$), we apply a higher dropout rate in the early timesteps to encourage token disentanglement, and then gradually reduce it in later steps to stabilize learning and preserve fine-grained visual fidelity.

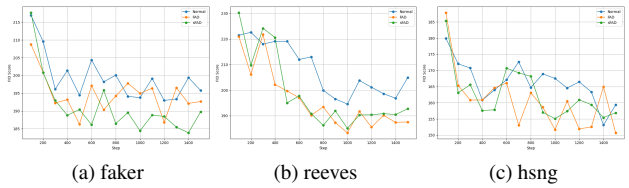
As part of future work, we plan to explore additional scheduling strategies, such as cyclical or adaptive schedules, to further improve the balance between generalization and visual fidelity.

5. Multi-concept Training

We also train LoRA using Normal Dropout and sFAD on multiple datasets for SDXL, specifically **pikachu** and **pochacco**. As shown in Fig. 1, the model trained with sFAD successfully reproduces both characters, whereas the model trained with Normal Dropout fails to do so, resulting in noticeable character deformations.

6. Score Comparison

We present graphs comparing the scores of Normal Dropout, FAD, and sFAD across all datasets over 100 to 1500 training steps for the metrics FID, DINO, InsightFace and CCIP. Overall, FAD and sFAD consistently achieve better scores than Normal Dropout, although the degree of improvement varies across datasets. The curves exhibit significant fluctuations depending on the dataset, and there are intervals where the ranking of scores swap or score drops, indicating potential overfitting during training. Graphs are shown in Figs. 2 and 3 for FID, Figs. 4 and 5 for DINO, Figs. 6 and 8 for InsightFace, and Figs. 7 and 9 for CCIP. However, CCIP scores appear to be saturated and less discriminative because the CCIP model already has strong prior knowledge of characters such as Pikachu. This makes CCIP less reliable for evaluating subtle differences in character reproduction for these datasets.



(a) faker (b) reeves (c) hsng

(d) mbst (e) pikachu (f) pochacco

Figure 3. Comparison of FID (↓) scores of three methods across all steps for six datasets, extracted from **SDXL**.

References

[1] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1

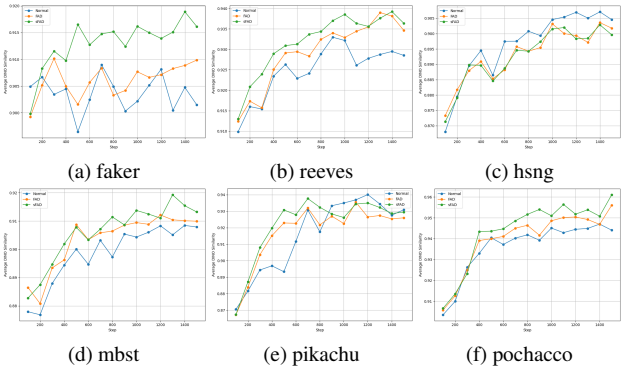
[2] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation, 2024. 2

Table 3. DINO (↑) scores with and without trigger token. Bold indicates large change of Δ (x - o).

Model	Method	faker			reeves			hsng			mbst			pikachu			pochacco		
		o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ
SD 1.5	Normal	0.897	0.899	+0.002	0.924	0.924	+0.000	0.896	0.898	+0.002	0.898	0.898	+0.000	0.915	0.919	+0.004	0.936	0.937	+0.001
	FAD	0.907	0.902	**-0.005**	0.929	0.926	**-0.003**	0.893	0.893	**+0.000**	0.903	0.902	**-0.001**	0.917	0.918	**+0.001**	0.940	0.940	**+0.000**
	sFAD	0.910	0.906	-0.004	0.932	0.930	-0.002	0.893	0.893	**+0.000**	0.907	0.906	**-0.001**	0.922	0.924	+0.002	0.944	0.944	**+0.000**
SDXL	Normal	0.9037	0.9029	-0.0008	0.9238	0.9224	-0.0014	0.9013	0.8960	**-0.0053**	0.9025	0.9017	-0.0008	0.8509	0.8592	+0.0083	0.8582	0.8561	-0.0024
	FAD	0.9061	0.9058	-0.0003	0.9322	0.9296	**-0.0026**	0.9005	0.9028	+0.0023	0.9115	0.9094	**-0.0021**	0.8628	0.8522	**-0.0106**	0.8828	0.8671	**-0.0157**
	sFAD	0.9107	0.9096	**-0.0011**	0.9325	0.9307	-0.0018	0.9034	0.9035	+0.0001	0.9124	0.9111	-0.0013	0.8671	0.8585	-0.0086	0.8668	0.8673	+0.0005

Table 4. FID (↓) scores with and without trigger token. Bold indicates the largest change of Δ (x - o).

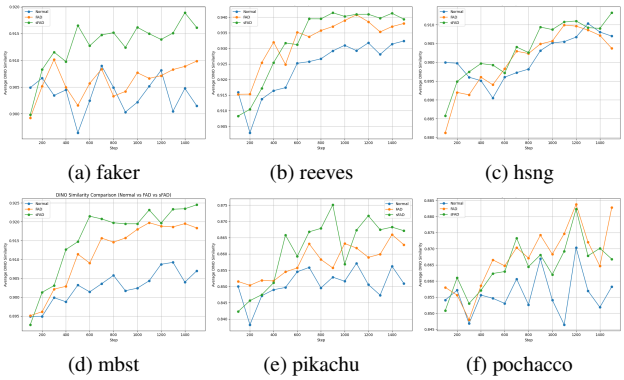
Model	Method	faker			reeves			hsng			mbst			pikachu			pochacco		
		o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ
SD 1.5	Normal	207.103	206.449	-0.654	210.532	210.463	-0.069	164.259	163.539	-0.720	137.473	136.017	-1.456	137.473	135.702	-1.771	136.017	135.404	-0.613
	Adaptive	197.462	204.428	**+6.966**	203.077	204.407	+1.330	161.200	162.812	+1.612	137.760	133.723	**-4.037**	137.760	138.203	+0.443	133.723	137.205	**+3.482**
	Step	195.863	197.097	+1.234	198.524	200.642	**+2.118**	160.478	161.665	**+1.187**	136.746	131.513	**-5.233**	136.746	137.327	**+0.581**	131.513	134.811	+3.298
SDXL	Normal	199.310	199.970	+0.660	208.110	209.600	**+1.490**	166.350	166.390	**+0.040**	236.945	234.35	-2.595	170.046	164.853	-5.194	193.026	194.358	+1.332
	Adaptive	194.690	199.730	**+5.040**	196.300	196.150	-0.150	160.800	159.750	-1.050	216.498	218.28	**+1.782**	160.946	165.081	**+4.134**	182.566	193.153	**+10.587**
	Step	197.760	197.720	-0.040	199.090	197.320	-1.770	162.650	161.010	-1.640	217.34	216.95	-0.39	164.224	165.339	+1.115	184.087	193.974	+9.887



(a) faker (b) reeves (c) hsng

(d) mbst (e) pikachu (f) pochacco

Figure 4. Comparison of DINO (↑) scores of three methods across all steps for six datasets, extracted from **SD 1.5**.



(a) faker (b) reeves (c) hsng

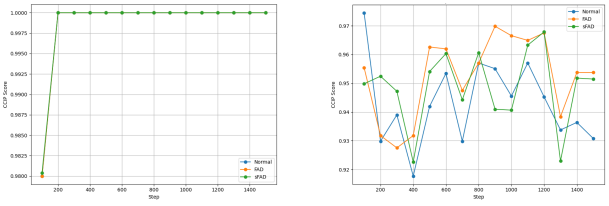
(d) mbst (e) pikachu (f) pochacco

Figure 5. Comparison of DINO (↑) scores of three methods across all steps for six datasets, extracted from **SDXL**.

(a) faker (b) reeves

(c) hsng (d) mbst

Figure 6. Comparison of InsightFace (↓) scores of three methods across all steps for four human datasets, extracted from **SD 1.5**.



(a) pikachu (b) pochacco

Figure 7. Comparison of CCIP (↑) scores of three methods across all steps for two character datasets, extracted from **SD 1.5**.

Table 5. InsightFace (↓) scores with and without trigger token. Bold indicates the largest increase in Δ (x - o).

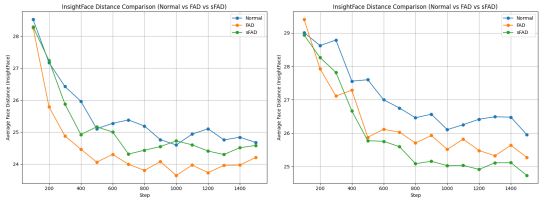
Model	Method	faker			reeves			hsng			mbst		
		o	x	Δ	o	x	Δ	o	x	Δ	o	x	Δ
SD 1.5	Normal	28.064	28.202	+0.138	30.619	30.999	+0.380	27.792	27.655	-0.137	28.015	28.087	+0.072
	Adaptive	26.997	27.826	**+0.829**	30.236	31.131	**+0.895**	27.511	27.596	+0.085	28.058	28.098	+0.040
	Step	26.954	27.655	+0.701	30.192	30.922	+0.730	27.433	27.522	**+0.089**	27.835	27.908	**+0.073**
SDXL	Normal	25.510	25.560	+0.050	27.070	26.570	-0.500	24.960	24.990	+0.030	27.280	27.300	**+0.020**
	Adaptive	24.470	25.320	**+0.850**	26.290	26.960	**+0.670**	24.410	24.820	**+0.410**	25.440	25.410	-0.030
	Step	25.130	25.160	+0.030	25.930	26.250	+0.320	24.530	24.490	-0.040	24.550	24.570	**+0.020**

Table 6. CCIP (↑) scores with and without trigger token. Bold indicates the largest decrease in Δ (x - o).

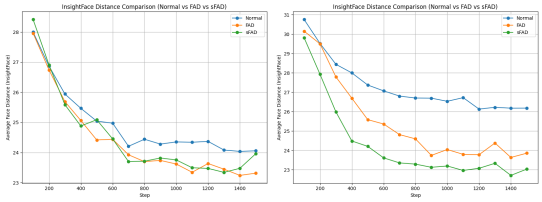
Model	Method	Pikachu			Pochacco		
		o	x	Δ	o	x	Δ
SD 1.5	Normal	0.9987	0.9987	+0.0000	0.9431	0.9350	**-0.0081**
	Adaptive	0.9987	0.9987	+0.0000	0.9528	0.9487	-0.0041
	Step	0.9987	0.9986	**-0.0001**	0.9487	0.9428	-0.0059
SDXL	Normal	0.9992	0.9972	-0.0020	0.8996	0.8341	-0.0655
	Adaptive	0.9996	0.9451	-0.0545	0.9091	0.7514	-0.1577
	Step	1.0000	0.9412	**-0.0588**	0.9295	0.7708	**-0.1587**

Table 7. Comparison of SD 1.5 and SDXL for Step (0.1~0.8) vs Step (0.8~0.1) using Linear / Exp_up methods.

Char.	Step	Metric	SD 1.5		SDXL	
			Linear	Exp_up	Linear	Exp_up
Faker	0.1∼0.8	FID (↓)	196.119	195.863	**191.29**	197.76
		DINO (↑)	0.9100	0.9100	**0.9131**	0.9107
		InsightFace (↓)	26.976	**26.954**	**24.51**	25.13
	0.8∼0.1	FID (↓)	195.411	**194.655**	196.85	**194.93**
		DINO (↑)	**0.9107**	0.9100	0.9087	**0.9120**
		InsightFace (↓)	26.967	26.984	**25.23**	25.87
Reeves	0.1∼0.8	FID (↓)	199.778	**198.524**	**194.17**	199.09
		DINO (↑)	**0.9320**	**0.9320**	**0.9336**	0.9325
		InsightFace (↓)	30.038	30.192	**25.85**	25.93
	0.8∼0.1	FID (↓)	201.095	201.838	**195.26**	198.06
		DINO (↑)	**0.9320**	0.9310	**0.9342**	0.9322
		InsightFace (↓)	**29.950**	29.995	**25.53**	26.22
hsng	0.1∼0.8	FID (↓)	161.579	**160.478**	**162.49**	162.65
		DINO (↑)	**0.8930**	**0.8930**	**0.9062**	0.9035
		InsightFace (↓)	27.503	**27.433**	**24.25**	24.53
	0.8∼0.1	FID (↓)	160.742	161.246	162.92	**158.16**
		DINO (↑)	0.8940	**0.8930**	**0.9069**	0.9034
		InsightFace (↓)	27.451	27.479	24.59	**24.36**
mbst	0.1∼0.8	FID (↓)	212.874	212.874	**210.85**	217.34
		DINO (↑)	0.9060	**0.9070**	**0.9159**	0.9124
		InsightFace (↓)	27.955	**27.835**	**24.27**	24.55
	0.8∼0.1	FID (↓)	**212.153**	**212.153**	217.08	**213.92**
		DINO (↑)	0.9040	0.9050	**0.9099**	0.8945
		InsightFace (↓)	27.898	27.862	25.67	**24.31**

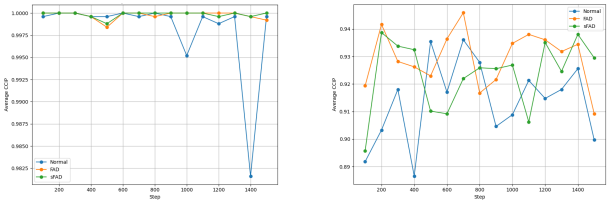


(a) faker (b) reeves



(c) hsng (d) mbst

Figure 8. Comparison of InsightFace (↓) scores of three methods across all steps for four human datasets, extracted from **SD XL**.



(a) pikachu (b) pochacco

Figure 9. Comparison of CCIP (↑) scores of three methods across all steps for two character datasets, extracted from **SDXL**.

Table 8. The full version of evaluation prompts for comparative evaluation with GPT-4.1

Evaluation Prompt:
You will be shown TWO images:
- **Image 1**: the ground truth character image (reference only)
- **Image 2**: a generated image that should depict the same character.

Your task is to critically evaluate how well **Image 2** resembles **Image 1**, considering the character identity and overall image quality.

Ground Truth Character key features (from Image 1):
Character (mbst): mbst, beard, brown hair, short hair

Evaluation Dimensions
Character Similarity:
- How closely does the generated character (Image 2) resemble the reference character (Image 1)?
- Consider race, facial structure, hairstyle, eye color, and other distinctive features.

Composition & Image Quality:
Assess technical and artistic quality of Image 2, including:
- Composition coherence
- Deformities (e.g., extra limbs, distorted face)
- Texture, lighting, color accuracy, clarity

Scoring Criteria
Character Similarity (10-point scale):
10 = highly similar, 1 = completely different
- Deduct 2 points for incorrect race or major facial mismatch
- Deduct 1 point for missing or incorrect key features (e.g., hair/eye color)

Composition & Image Quality (10-point scale):
10 = excellent, 1 = very poor
- Deduct 2 points for major deformities
- Deduct 1 point for lighting/color/texture issues or minor artifacts

Output Format (strict)
Image2 Evaluation:
- [Concise bullet points listing observed issues or mismatches]

Scores:
Character Similarity: [score] / 10
Composition & Image Quality: [score] / 10

Notes
- Do **not** evaluate Image 1.
- Be critical and specific.
- Start your answer immediately after this instruction.

Table 9. The full version of evaluation result for comparative evaluation with GPT-4.1


Image 2 Evaluation:

Character Similarity:

- Hairstyle is similar (short, brown hair, styled to the side), but appears slightly neater and more stylized than in Image 1 (**-1 points**).
- Beard is present, but the style is somewhat different (**-1 points**).
- General facial structure is fairly close but the face in Image 2 appears slimmer and more angular (**-1 points**).

Composition & Image Quality:

- No major issues with race or overall appearance.
- Image 2 shows the person holding a microphone, altering focus compared to the original (**-1 points**).
- No major deformities or artifacts detected.
- Lighting and texture are clean, but the image feels more artificial and softened, losing some realism compared to the reference.

Scores:

Character Similarity: **7 / 10** Composition & Image Quality: **9 / 10**