Appendix

Appendix Contents

A. Appendix

B. Baseline Details

DiT setup. We follow the standard DiT block design and incorporate recent improvements including QK normalization and rotary position embeddings (RoPE). Training hyperparameters match those of Peebles and Xie [45]: batch size = 256, 12 transformer blocks for DiT-B and 28 for DiT-XL. We use patch size of 2×2 for all experiments. We train all baseline using rectified flow matching objective and use logit-normal distribution for sampling the timesteps.

U-ViT setup. U-ViT mirrors DiT but adds U-Net–style residual (skip) connections. To isolate architectural effects, we use the same transformer blocks and training hyperparameters as DiT, differing only in the inclusion of these residual connections.

HDiT setup. HDiT follows DiT but applies PixelShuffle/PixelUnshuffle to reduce the token count while increasing channel dimensionality. We adopt this token–channel trade-off on the same transformer blocks as baselines. We use a single downsampling/upsampling operation after blocks 6 and 22. We also exclude local attention and instead use full self-attention. We train with the same hyperparameters as the other baselines.

Qwen-Image setup. We add ELIT *Read/Write* layers at blocks 8 and 52 of the 60-layer Qwen-Image backbone. Training uses a weighted sum of RF and distillation losses. The distillation term is scaled by $20\times$ to match the magnitude of the RF loss. We train for 60k steps at 512px with a global batch size of 1536, followed by 60k steps at 1024px with a global batch of 384. We sample timesteps from a logit-normal distribution and use time shifting of 2.22 during training and 2.0 during inference, following [56]. We do not apply any timestep-aware loss re-weighting. The training dataset is a combination of internal real images with synthetic samples generated by FLUX.1-Schnell and Stable Diffusion-XL (with 50/50 ratio). We found that the model converges quickly, but we observe a style bias toward the synthetic data (reduced detail and more saturation relative to original Qwen-Image). For sampling, we use the Euler ODE sampler with 40 steps and use CFG value of 6.0

TeaCache setup. TeaCache proposes two strategies for deciding when to reuse (cache) the previous step's prediction: (1) using *timestep-modulated tensor relative error* between current and previous step to predict the accumulative error of caching the current step. (2) using *timestep-embedding relative error*, which measures the relative change of the timestep embedding itself across steps.

The original paper reports that strategy (1) generally works better. In text-to-image models (e.g., FLUX [34]), input tensors are modulated by the timestep embedding, providing access to the timestep-modulated input tensor. In our class-conditional image and video setting, those tensors are additionally modulated by the class signal, preventing access to timestep-modulated tensor. Empirically, on DiT for class-conditional ImageNet, we found that using class-timestep modulated input tensor following strategy (1) does not provide good estimate for the caching error and leads to degraded quality, underperforming the second strategy. Consequently, we adopt the timestep-embedding relative error (strategy 2) for all TeaCache experiments in this work.

C. Method Details

Adapting ELIT to baselines. Aside from adding the Read/Write operations, we leave each baseline's architecture and training unchanged. Unless noted, we place the *Read* at block 4 and the *Write* at block 24 for XL-size models across all baseliness (DiT, U-ViT, HDiT), as motivated by our ablations in Table 4.

Multi-budget training setup. We use 16 spatial groups per image in all main experiments. On ImageNet-1K, each group contains 16 latent tokens at 256px and 64 at 512px. Unless otherwise noted, during training we set $J_{\max}$ to the per-group maximum (64 at 512px; 16 at 256px) and $J_{\min}$ to 1 for 256px and 4 for 512px, yielding 16 distinct inference budgets at 256px and 60 at 512px. At each training iteration, $\tilde{J}$ is sampled once and broadcast to all GPUs, ensuring

	Spat. Blocks	Lat. Blocks	Read	Write
Attn. Proj.	$8Nd^2$	$8JGd^2$	$d^2(4N+4JG)$	$d^2(4N+4JG)$
Attn. Mat.	$2N^2d$	$2J^2G^2d$	$2JNd$	$2JNd$
FF	$16Nd^2$	$16JGd^2$	$4JGd^2$	$4Nd^2$

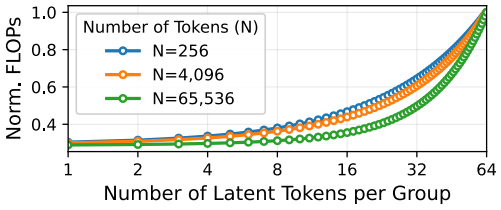

Figure 10. (left) FLOPs for spatial blocks, latent blocks, and Read/Write layers as a function of input tokens N, groups count G, latent tokens per group J, and hidden size d. (right) Relationship between latent tokens per group and model FLOPs for a DiT-XL with 8 spatial blocks, 20 latent core blocks, and $N/64$ groups, varying input tokens N and latent tokens per group $\tilde{J}$. FLOPs are shown relative to 64 tokens per group.

synchronized compute with no added overhead. To account for the reduced compute, we increase the batch size from 256 (baselines) to 384 to match training FLOPs.

Kinetics-700 setup. We train at 256px on 29 frames sampled at 24 fps. The encoder produces 8 latent frames of shape $8\times32\times32$. We use a patch size of $1\times2\times2$, this yields 2,048 tokens. We use a group size of $2\times4\times4$, giving 64 groups per video. Kinetics-700 is trained with a single compute budget without multi-budget training.

Inference setup. We use the Euler ODE sampler with 40 steps for all experiments. Image experiments are evaluated on the ImageNet-1K validation split, and video experiments on the Kinetics-700 validation split.

D. Compute-quality Tradeoff Efficiency

Increasing the training image resolution scales the required compute quadratically, making higher resolution training expensive. To control the compute while keeping model configuration the same, DiT proposed to increase the patch size to cut token count, while HDiT inserts a downsampling stage that reduces tokens but increases parameters count. We instead propose to cap the number of latent tokens per group during training, reducing training compute while keeping both patch size and model size constant.

To evaluate compute–quality trade-offs, we train low/high-compute variants for each baseline: DiT (larger patch size for the low variant), HDiT (model size matching other baselines), and ELIT-DiT (fewer latent tokens). Intuitively, given a similar reduction in compute between the two versions, the architecture with least performance degradation is more desirable.

To measure this, we define a degradation metric $\rho = ((\text{Metric Ratio})/\text{FLOPs Ratio})$, where "Metric Ratio" represents metric degradation caused by the low-compute model and "FLOPs Ratio" represents the corresponding reduction in FLOPs. As shown in Table 5, not only our method outperforms baselines at similar training compute, but also shows consistently lower ρ indicating it can more efficiently make use of its comupte if constrained, a capability we attribute to the latent interface's focus on the most important information in the input.

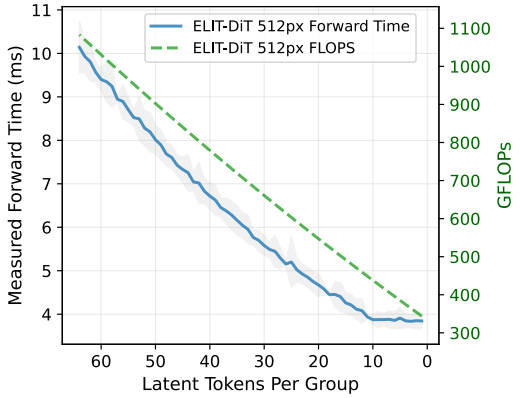

Figure 11. Lowering inference budget by using fewer latent tokens per group yields correlated reductions in forward time and FLOPs.

E. Ablations on Read/Write Strategies.

In Table 6, we compare alternative Read/Write designs and find that a single cross-attention Read layer outperforms both a Q-Former–style Read layer [35] and full self-attention. Additionally, stacking two cross-attention layers in the Read yields no measurable gain, suggesting one layer suffices. However, adding a second cross-attention layer in the Write or expanding the FFN hidden dimension by $\times4$ (as in the DiT block) offers improvements at the cost of additional FLOPs. To keep overhead at a minimum, we adopt a single Read/Write layer.

F. Compatibility with Distillation Methods.

We evaluate the compatibility of ELIT with the distillation technique such as grafting [7], which distills a base model into a smaller version of itself. We apply grafting on 100% ELIT MLPs with expansion ratio $r = 3$, obtaining 12.6% degradation in FID and 8.9% in IS, consistent with the original paper's reported degradation of 17.2% FID and 9.4%

Table 5. **Compute-quality tradeoff efficiency of baselines on ImageNet-1K 512px.** $\rho = (\text{Metric Ratio})/(\text{FLOPs Ratio})$ indicates the model degradation with respect to change in FLOPs between the low- and high-compute variants.

Baseline	Params	TFLOPs	$\text{FID}_{50K}\downarrow$ ($\rho\downarrow$)		$\text{FDD}_{50K}\downarrow$ ($\rho\downarrow$)		IS$\uparrow$ ($\rho\downarrow$)	
			–G	+G	–G	+G	–G	+G
DiT	675M	806	18.8 (1.00)	9.5 (1.00)	339.2 (1.00)	233.6 (1.00)	53.0 (1.00)	86.4 (1.00)
⌞ Patch size 2x4	675M	377	22.5 (0.56)	12.3 (0.61)	434.0 (0.60)	317.9 (0.74)	45.7 (0.54)	73.8 (0.55)
HDiT	1.4B	776	13.0 (1.00)	6.0 (1.00)	260.3 (1.00)	170.5 (1.00)	69.4 (1.00)	114.2 (1.00)
⌞ Smaller backbone	703M	392	22.2 (0.85)	11.5 (0.96)	435.2 (0.83)	315.4 (0.93)	48.8 (0.71)	80.0 (0.71)
ELIT-DiT	698M	831	11.1 (1.00)	4.9 (1.00)	175.6 (1.00)	106.1 (1.00)	80.0 (1.00)	134.1 (1.00)
⌞ 25% Tok.	698M	386	12.5 (**0.52**)	5.7 (**0.54**)	217.7 (**0.57**)	137.8 (**0.60**)	75.7 (**0.49**)	124.5 (**0.50**)

Baseline	$\text{FID}_{10K}\downarrow$	$\text{FDD}_{10K}\downarrow$	IS$\uparrow$
ELIT-DIT	26.53	531.8	45.95
Qformer Read	30.49	589.9	41.10
Self-Attn Read	28.38	602.5	40.12
Self-Attn Read/Write	29.46	631.1	38.49
↑ Read Capacity	27.45	540.7	45.40
↑ Write Capacity	25.23	516.9	47.59
↑ FFN Capacity	**24.80**	**507.7**	**48.22**

Table 6. **Architectural ablations on DiT-B/2.** Using cross-attn in Read/Write is superior to alternatives. Increasing the model capacity is only beneficial in Write and FFN.

Table 7. Budget scheduling across noise levels. ImageNet 512px, ELIT-DiT-XL/2. 50%_100%: uses 50% of tokens for high-noise steps ($t < 0.5$), 100% otherwise.

Method	**$\text{FID}_{10K}\downarrow$**	**IS$\uparrow$**	$\text{Iter}_{\text{FLOPs}}$
100%_100%	**11.60**	86.68	188
50%_100%	11.98	**90.18**	154

IS. This confirms that ELIT remains compatible with orthogonal efficiency methods such as network pruning and distillation.

G. Budget scheduling Across Noise Levels.

We explore allocating different token budgets across noise levels. As a proof of concept, we train ELIT on ImageNet 512px (DiT-XL/2) with 50% of tokens for high-noise steps ($t < 0.5$) and 100% for the remaining steps (50%_100%). As shown in Table 7, despite lower per-iteration compute (154 vs. 188 TFLOPs), performance remains comparable, suggesting that high-noise steps may require fewer tokens. We leave a principled study of noise-level-aware budget scheduling for training and inference as future work.

H. Joint vs. Independent Budget Training.

We compare our joint multi-budget model against independently trained single-budget ELIT models on ImageNet 512px (DiT-XL/2). As shown in Table 8, the joint model consistently outperforms independently trained models across all token budgets (100%, 50%, 25%) and all metrics. This demonstrates that multi-budget training acts as a regularizer, and that a single ELIT model natively supporting multiple budgets eliminates the need to train separate models for each budget.

Table 8. Joint multi-budget vs. independently trained single-budget models. When tested on ImageNet 512px, ELIT-DiT-XL/2, joint multi-budget models consistently outperform single budget models.

Method	**$\text{FID}_{10K}\downarrow$**	**$\text{FDD}_{10K}\downarrow$**	**IS$\uparrow$**
Indep. (100% tok.)	13.60	205.23	80.90
Joint (100% tok.)	**12.00**	**189.50**	**90.29**
Indep. (50% tok.)	14.14	222.43	77.99
Joint (50% tok.)	**12.95**	**203.58**	**85.18**
Indep. (25% tok.)	15.36	247.77	74.04
Joint (25% tok.)	**14.21**	**228.08**	**79.60**

First Two Latent Tokens Last Two Latent Tokens

Figure 12. Read attention masks averaged over noise levels. Early latent tokens attend to broad, semantically important image regions, while later tokens exhibit sparser attention focusing on fine-grained details.

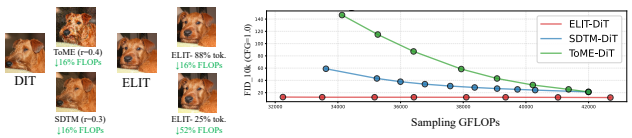

Figure 13. ELIT vs. token merging methods on ImageNet 512px (DiT-XL/2). Training-free methods (ToMe, SDTM) are bounded by DiT quality, while ELIT surpasses it even at 25% tokens.

I. Latent Token Importance Visualization.

We visualize the attention mask of the Read operation, averaged over noise levels, in Figure 12. Early latent tokens attend to broad image regions covering both background structure and the main object, whereas later tokens exhibit sparser attention patterns, often concentrating on fine-grained texture details. This confirms the importance ordering learned through tail dropping and is consistent with the observation that increasing the token count for Qwen-Image primarily improves high-frequency texture details while preserving overall structure (Figure 1).

J. Comparison with Token Merging Methods.

Token-merging methods [4, 8, 16, 42, 46, 57] can provide a knob to control inference budget. They are often training-free or require lightweight finetuning [8, 53] We compare ELIT against training-free token merging approaches (ToMe [4], SDTM [16]) on ImageNet 512px (DiT-XL/2). As shown in Figure 13, both training-free methods trade compute for quality less favorably than ELIT. ELIT improves over the base DiT even when using only 25% of the tokens ($\text{FID}_{10K} = 14.2$), while training-free methods are upper bounded by DiT quality ($\text{FID}_{10K} = 20.9$).

K. Compute Analysis of ELIT

We analyze the theoretical computation requirement for ELIT-DiT in comparison with standard DiT design. Figure 10 (left) shows the relation between main architecture hyperparameters and FLOPs for the blocks employed by our architecture. When the number of core blocks is large with respect to spatial blocks, computation is focused on the latent core blocks and the Read and Write operations' cost is minimal with respect to the model cost. Figure 10 (right) exemplifies the case of a DiT-XL/2 architecture for varying input sequence lengths. The latent interface is particularly effective at reducing FLOPs with large sequence lengths (*e.g.* training on higher resolutions) due to the dominant self attention cost that is quadratically reduced with $\tilde{J}$.

FLOPs vs latency in ELIT. Figure 11 reports FLOPs and wall-clock forward time for ELIT–DiT on ImageNet-1k at 512px as we vary the number of latent tokens per group.

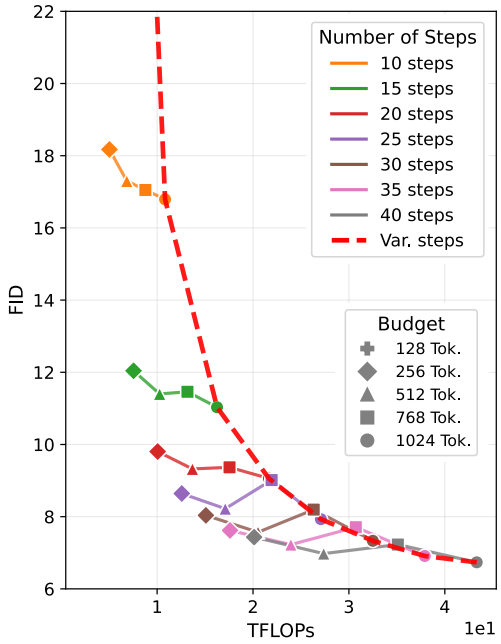

Figure 14. When tested with CFG 0.25, ELIT provides better quality-compute tradeoff than reducing the number of sampling steps.

Forward time drops monotonically with token count and closely follow the FLOPs reduction, showing that budget control yields real speedups. At higher budgets, the correlation weakens slightly due to fixed overheads (e.g., I/O and kernel launch), but the overall trend remains strongly aligned.

L. Additional Results

Compute-quality tradeoff. To verify the advantage of our method over simply reducing the number of sampling steps, we show in Figure 14 that our multi-budget model achieves a more favorable quality–compute tradeoff compared to varying the number of sampling steps.

Comparison to baselines. We show in Figure 18 additional qualitative results comparing our method to baselines on ImageNet-1K 512px. ELIT variants show less structural artifacts while allowing for per-step selection of inference budget and enabling autoguidance and cheap classifier-free guidance out of the box for cheaper and higher quality sampling.

Varying inference budget. In Figure 16, we evaluate the effects of varying the number of tokens in the latent interface for ELIT-DiT trained on ImageNet-1K 512px. As the model FLOPs decrease with the number of latent tokens, the model is able to preserve image structure while changing less noticeable details.

Comparison of guidance methods. We qualitatively eval-

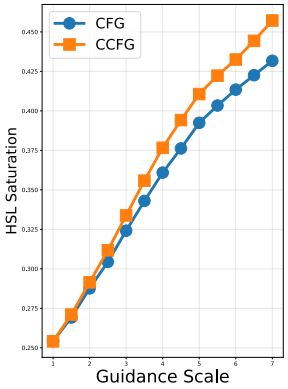

Figure 15. HSL saturation comparison between CFG and CCFG, across guidance scales on ImageNet 512px (DiT-XL/2). CCFG exhibits slightly higher saturation than CFG, attributed to its autoguidance component.

uate the effects of classifier-free guidance (CFG), autoguidance [30], and the proposed cheap classifier-free guidance (CCFG) (see Figure 17). We notice that AG produces results with most variation, including wider ranges of camera poses, compositions with multiple subjects and objects occlusion. By comparing results across different weights, we notice that AG remains most closely aligned with low guidance weight results, avoiding the mode collapse effect visible for CFG and CCFG that pushes samples towards more object-centric representations for the given class. We attribute this observation to the lower Inception Scores obtained by AG in Figure 5. Both AG and CCFG produce improved results which are particularly noticeable in complex concepts such as humans. CCFG combines the object-centric behavior of CFG, while reaping improved generation of complex objects from AG.

CCFG saturation analysis. We quantitatively analyze the saturation behavior of CCFG compared to CFG and AG. As shown in Figure 15, CCFG exhibits slightly higher HSL saturation across guidance scales, which we attribute to the stronger guiding effect contributed by its autoguidance component. The qualitative comparisons in Figure 17 shows that CCFG tends to saturate at larger guidance scales. Thus, we recommend using lower guidance scales with CCFG to mitigate this effect.

Additional Qwen-Image Results. We provide in Figure 19 additional qualitative comparison for ELIT-Qwen-Image against the original model. Thanks to CCFG, our model performs sampling with 69% of the FLOPs with respect to Qwen-Image and is able to produce a smooth trade-off between sample quality and model FLOPs by varying the amount of tokens in the latent interface. In the cheapest shown configuration, ELIT-Qwen-Image uses only 35% of the FLOPs with respect to the original model. As the number of latent tokens is decreased, the model preserves structural details, prioritizing changes in the least prominent image details.

Additional ImageNet-1k 512px Results. We provide in Figure 20, Figure 21 and Figure 22 additional qualitative comparison on ImageNet-1k 512px where we compare baseline DiT method with ELIT-DiT using CFG and CCFG. Class Ids and samples were randomly selected.

M. Failed Experiments

Spatial token masking for flexible inference computation. We explored ideas from masked diffusion transformers [17, 33, 65] as a way to obtain variable inference budget by dropping tokens in the spatial domains. We found token dropping in the spatial domain not to produce satisfactory results when applied at inference time and attribute its lower performance to the unrecoverable information loss in the spatial regions corresponding to dropped tokens.

Per group latent tokens count. We experiment with automatic per-group budget assignment, i.e. making $\tilde{J}$ different for each group rather than uniform across groups, with the aim of assigning more tokens to groups with more complex content, further improving compute reallocation. To achieve this, we use the loss map to supervise an additional DiT block positioned at the beginning of the DiT which predicts importance score for every group according to the loss map. Given a desired total number of tokens, we automatically distribute latent tokens to different groups, assigning more tokens to groups with higher importance score. We find this variant to increase model and implementation complexity while matching the performance of ELIT. We hypothesize that our read operation is already tailored to read more from spatial tokens with higher loss as shown in Figure 2.

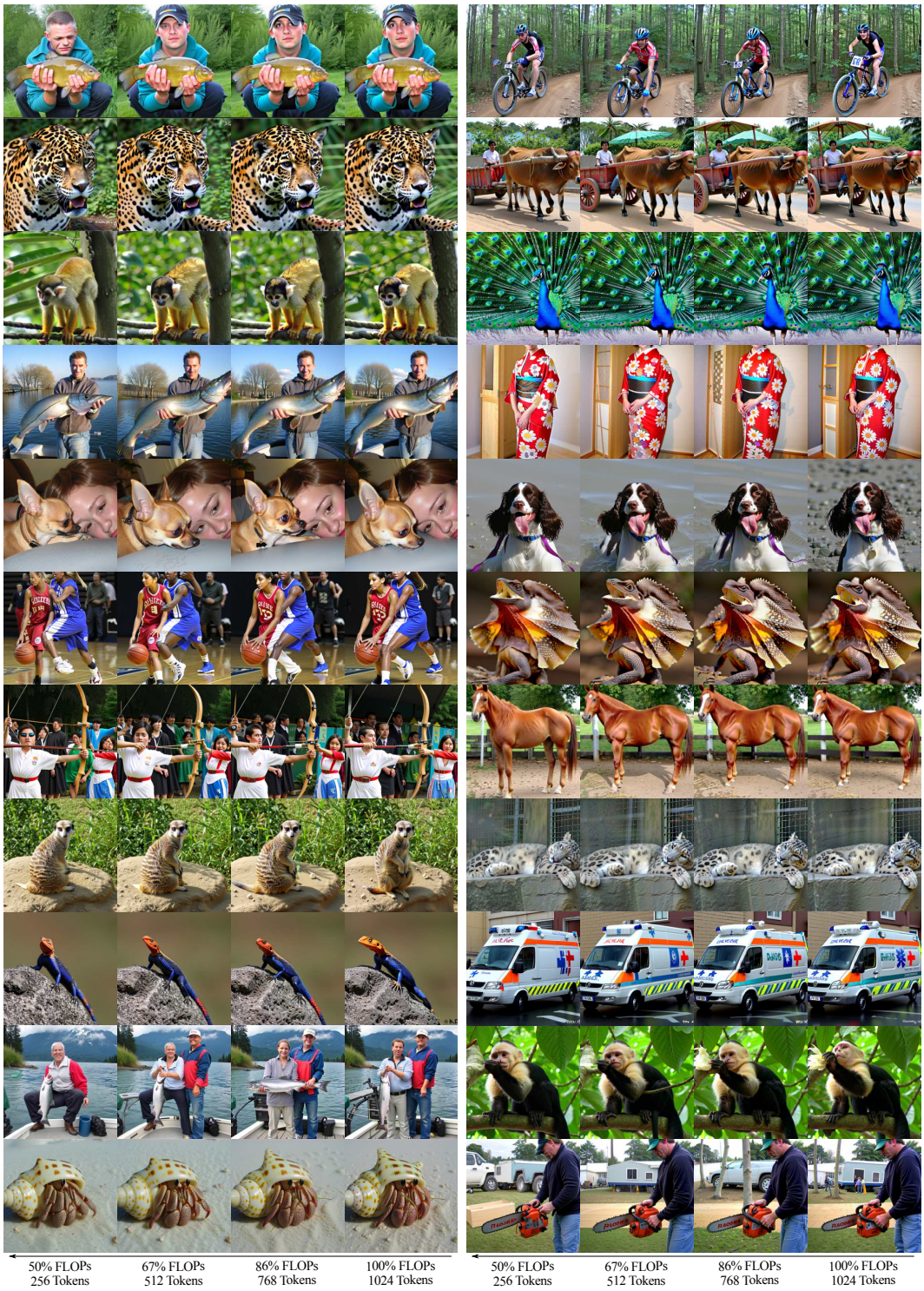

Figure 16. Qualitative results produced by ELIT-DiT on ImageNet-1K 512px with CCFG 4.0 for varying number of tokens in the latent interface. As the tokens and model FLOPs are reduced, the model preserves structure, while varying image details, producing gradual image changes. FLOPs are expressed relative to the model variant where no latent tokens are dropped.

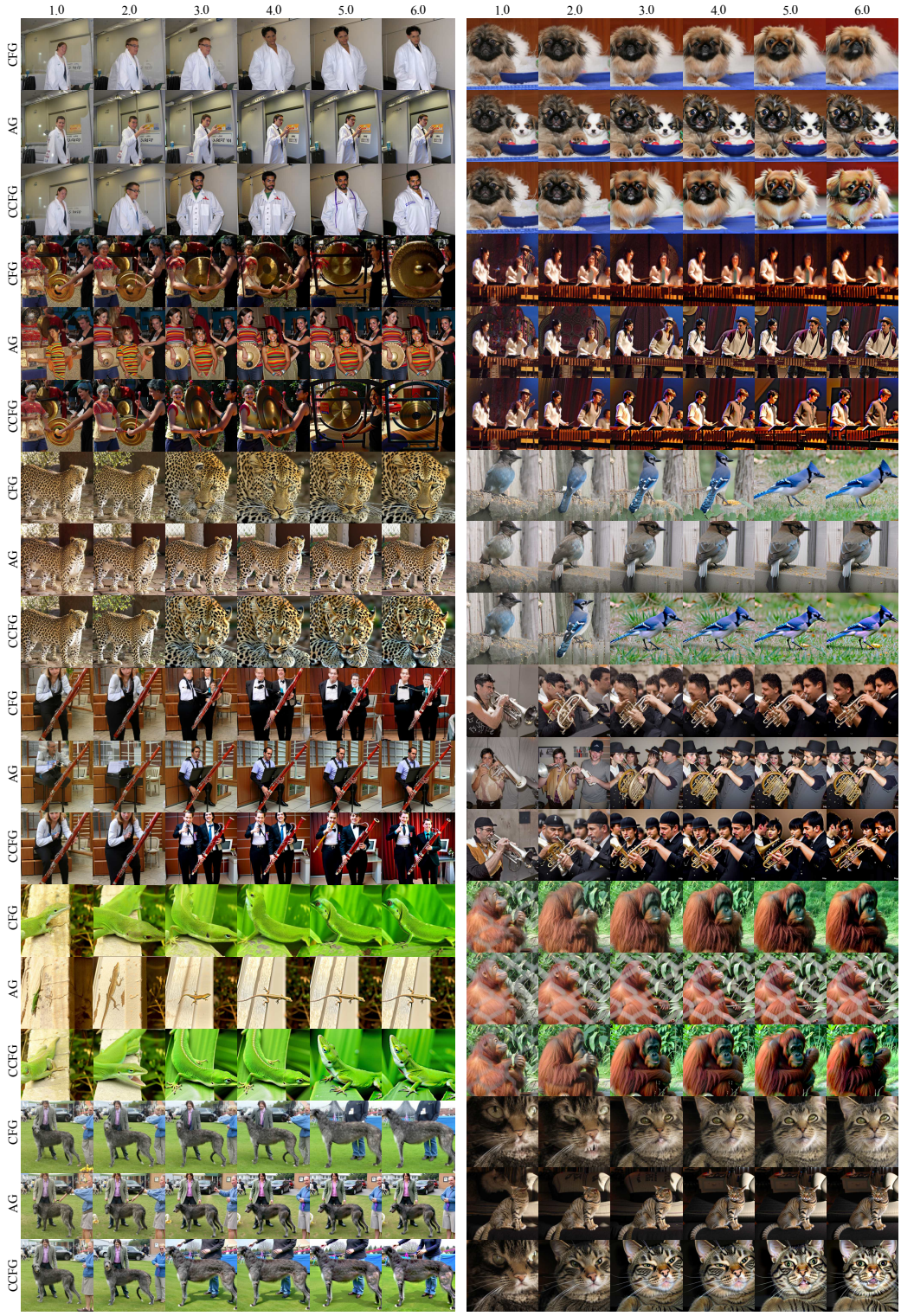

Figure 17. Qualitative comparison of classifier-free guidance (CFG), autoguidance [30] (AG), cheap classifier-free guidance (CCFG) with different weights, when applied to ELIT-DiT trained on the ImageNet-1K 512px dataset. AG produces the most varied samples, generating results with similar structure across guidance weights, as opposed to CFG and CCFG which favor object-centric generations. Both AG and CCFG produce better generations of complex concepts such as human faces.

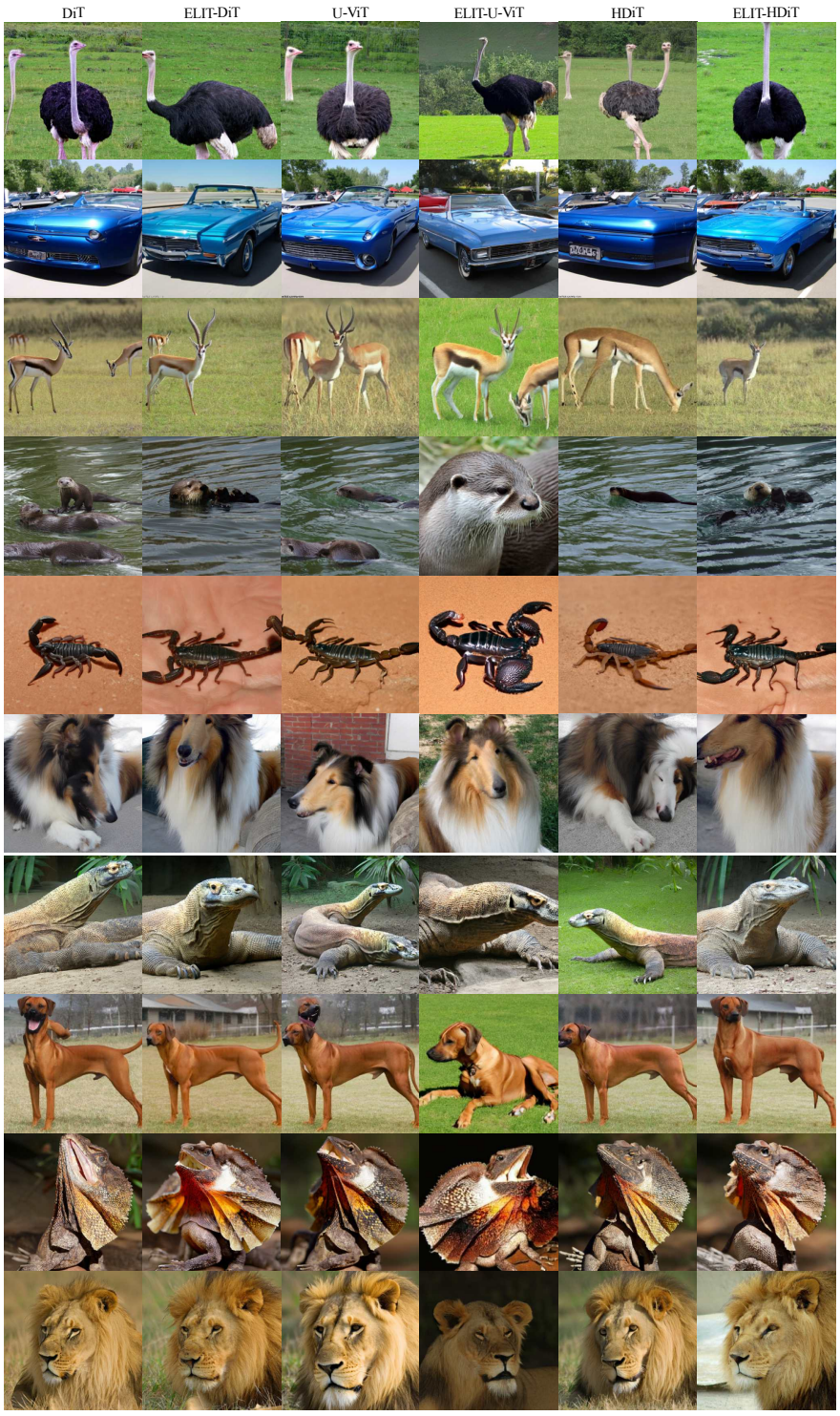

Figure 18. Qualitative comparison of ELIT against baselines on the ImageNet-1K 512px dataset. Results are produced using CFG with weight 4.0 for all methods.

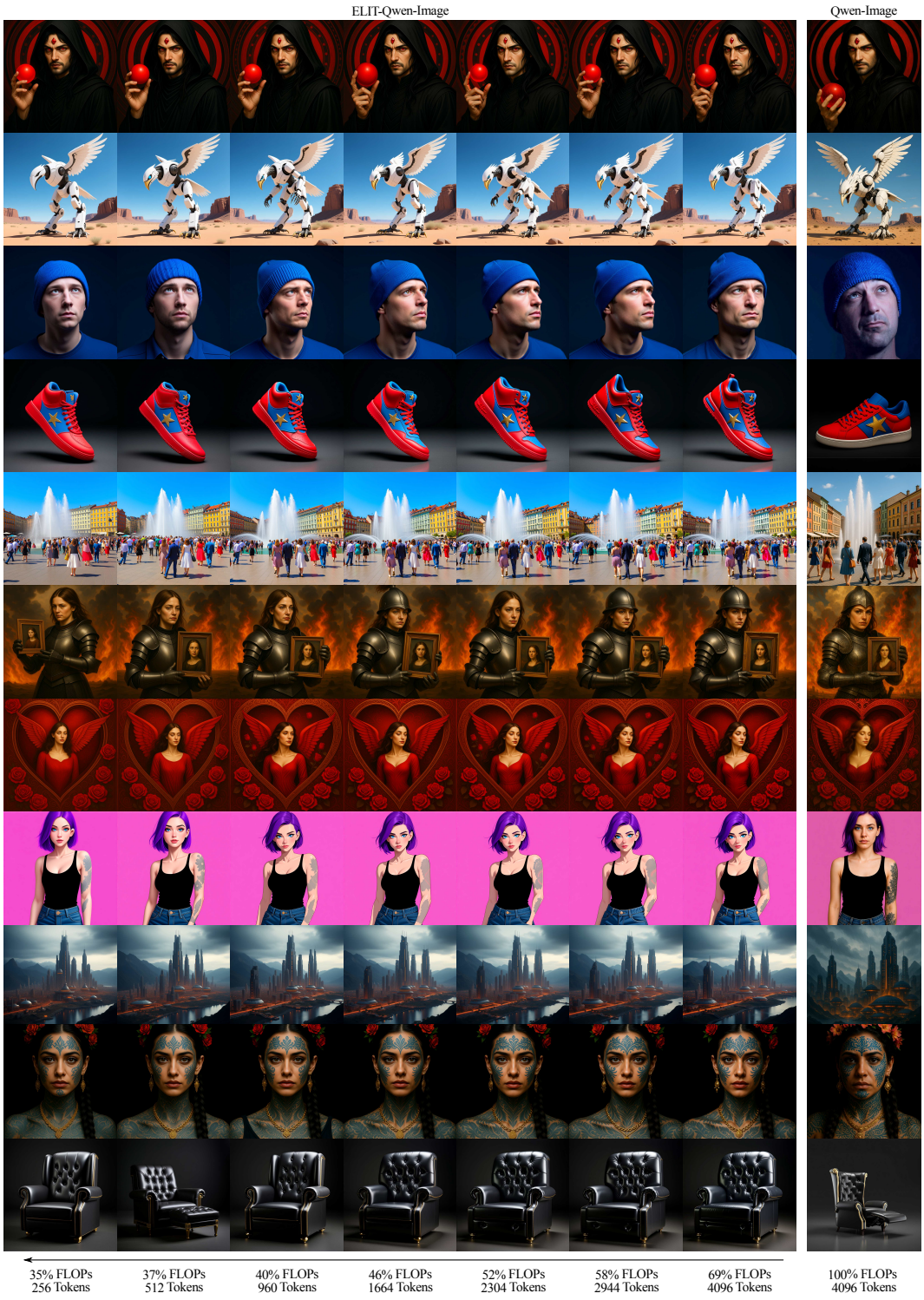

Figure 19. Qualitative results produced by ELIT-Qwen-Image for varying number of tokens in the latent interface. As the number of tokens is decreased and model FLOPs are reduced, our method can preserve structural details, while prioritizing changes in image details, preserving perceptual quality. Reported FLOPs are expressed relative to the original Qwen-Image and account for both the sampling FLOPs reductions brought by CCFG and the reduction in the number of tokens in the latent interface.

Figure 20. Uncurated Qualitative samples comparing DiT with ELIT-DiT using CFG and CCFG on ImageNet-1k 512px. Results are produced using CFG with weight 4.0 for all methods.

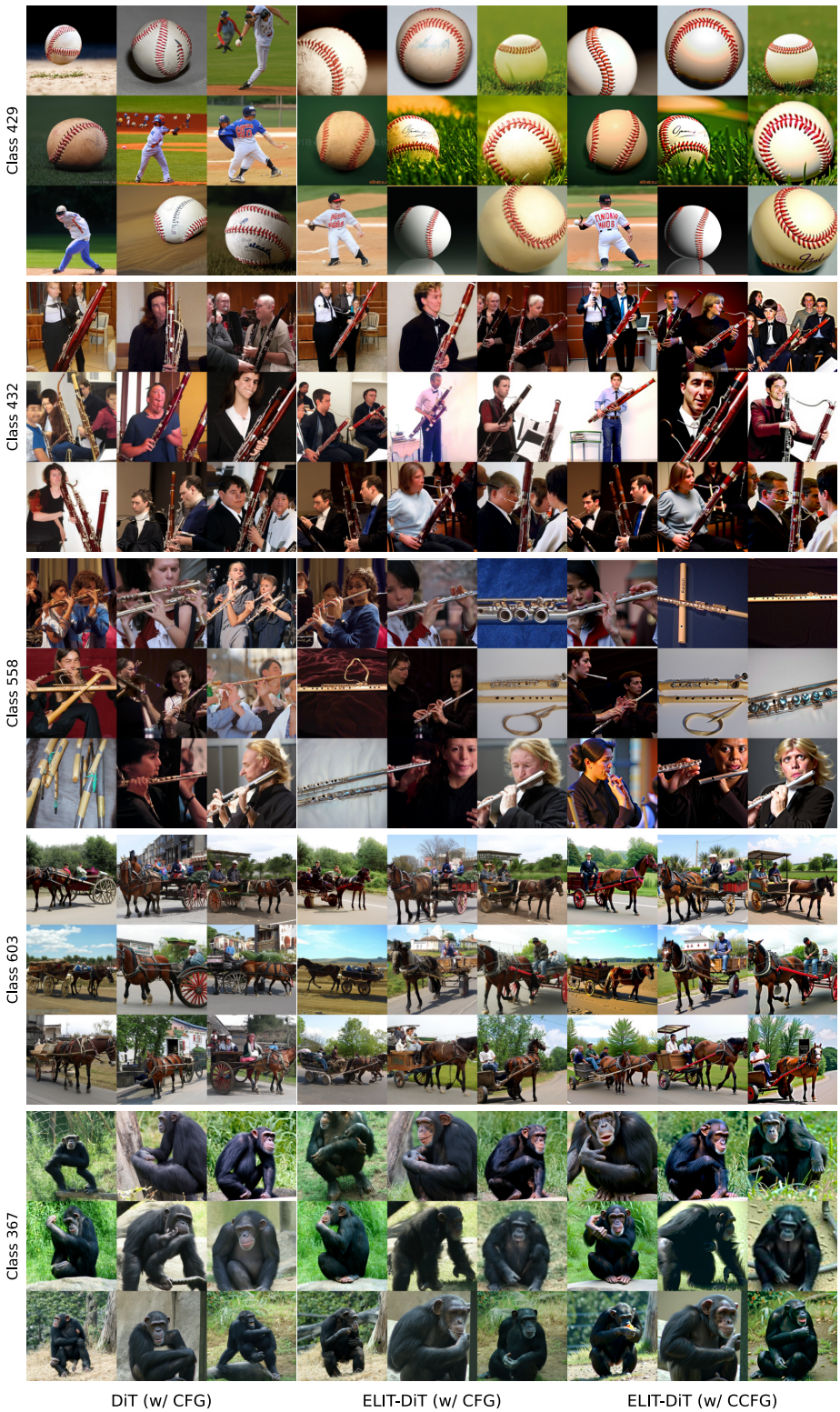

Figure 21. Uncurated Qualitative samples comparing DiT with ELIT-DiT using CFG and CCFG on ImageNet-1k 512px. Results are produced using CFG with weight 4.0 for all methods.

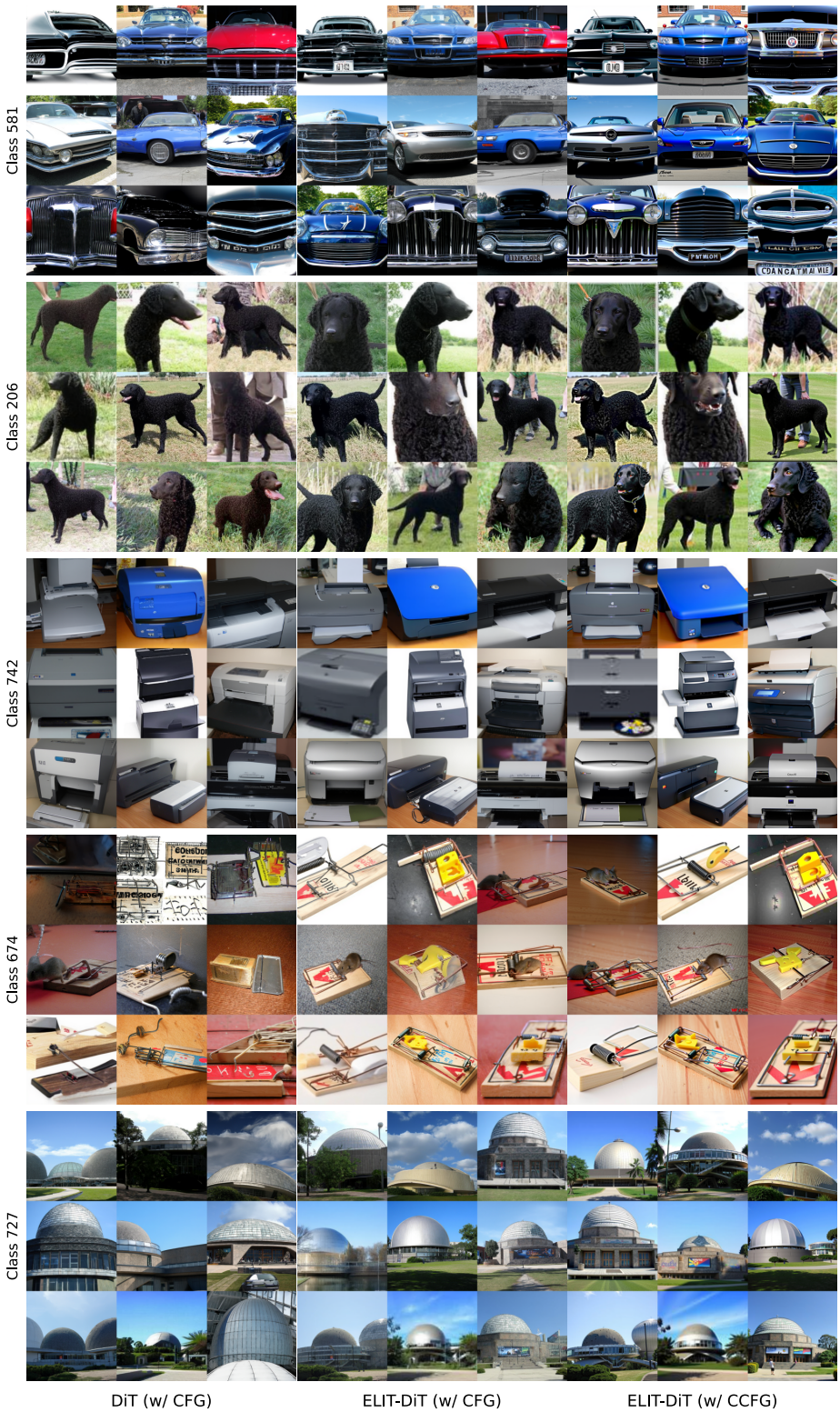

Figure 22. Uncurated Qualitative samples comparing DiT with ELIT-DiT using CFG and CCFG on ImageNet-1k 512px. Results are produced using CFG with weight 4.0 for all methods.

Figure	Prompt
Figure 1	*"The image portrays a woman with dark skin wearing a gold headpiece adorned with a blue jewel. Her gaze is directed towards something off-camera, giving her a focused expression. The background appears to be blurred, drawing attention to her face and headpiece."*
Figure 1	*"The image features actor Liev Schreiber in a snowy scene from a movie or TV show. He is dressed in black tactical gear, including a vest with "ARCTIC OCEAN" written on it, and a helmet with goggles. The setting appears to be a bustling city street filled with people and vehicles, all covered in snow."*
Figure 1	*"The image features a woman walking down a city street at night. She is wearing a black leather jacket, a white crop top, and a short black skirt. The street is illuminated by neon signs and streetlights, creating a vibrant atmosphere. There are other people visible in the background, but they are not the main focus of the image."*
Figure 19	*"The image portrays a man with long black hair and red eyes, wearing a black hooded cloak. He has a red gem on his forehead and holds a red orb-like object in his hand. The background features a circular pattern with red and black colors."*
Figure 19	*"The image features a large, white robot-like creature with wings standing on a desert landscape. The creature has sharp claws and appears to be looking down at something. Its body structure resembles a fusion of humanoid and bird-like characteristics. The background consists of a clear blue sky and rocky terrain."*
Figure 19	*"The image features a man wearing a blue knit cap, looking upwards with a serious expression. The background is dark blue, creating a contrast with the man's face and hat."*
Figure 19	*"The image showcases a vibrant sneaker with a red upper and blue accents. The shoe features a gold star design on the side and has red laces. The background appears to be a dark gray or black surface, providing a stark contrast to the colorful sneaker."*
Figure 19	*"The image captures a lively scene in a city square where people are walking around a fountain that is spraying water into the air. The square is surrounded by colorful buildings, creating a vibrant atmosphere. People are dressed in various styles of clothing, including dresses and suits, indicating a diverse crowd. Some individuals are carrying handbags, suggesting they might be tourists or shoppers. The sky above is blue"*
Figure 19	*"The image portrays a woman dressed in full armor, holding a small picture frame with a portrait of another woman inside. The background features dramatic clouds and fire, adding intensity to the scene."*
Figure 19	*"The image portrays a woman inside a large, ornate heart with wings. The heart is surrounded by red roses and intricate designs, creating a fantastical and romantic atmosphere."*
Figure 19	*"The image portrays a woman with purple hair and tattoos on her arm. She has striking blue eyes and is wearing a black tank top and jeans. The background is a solid color, possibly pink or magenta."*
Figure 19	*"The image depicts a futuristic cityscape with tall buildings and domed structures illuminated by orange lights. The city is surrounded by mountains and is situated near a body of water. The sky above the city appears cloudy."*
Figure 19	*"The image features a woman with intricate blue tattoos on her face and neck. She has a serious expression and is adorned with gold jewelry, including earrings and a necklace. Her hair is styled in braids, and she wears a flower crown. The background is dark, which contrasts with her colorful appearance."*
Figure 19	*"The image features a luxurious black leather armchair with gold accents. The chair has a high backrest adorned with buttons and a footrest. It is positioned against a dark background, creating a dramatic effect."*

Table 9. Prompts used to produce the showcased qualitative results for Qwen-Image [56].