ReCAST: Probing Sparse Reference Use in In-Context Image Generation

Yeon Gyu Han[1,*] dusrb37@gmail.com
Junah Jung[2,*] goldfish0907@gmail.com
Dongheon Lee[3,4,†] dhlee.jubilee@gmail.com

Abstract

Personalized image generation in Diffusion Transformers (DiTs) increasingly relies on *in-context conditioning*, where a reference image is tokenized and concatenated with the denoising sequence. While effective, this design introduces a substantial computational burden: in FLUX.1 Kontext, 4,096 additional reference tokens increase the quadratic attention cost by up to 3.6×. We observe that this cost is largely redundant. Generation tokens attend overwhelmingly to a small subset of reference tokens aligned with the foreground subject, while many background tokens receive negligible attention. Based on this finding, we propose ReCAST, a training-free sparse reference conditioning method that ranks reference tokens using cross-image attention and selects a timestep-adaptive subset during denoising. On FLUX.1 Kontext, ReCAST reduces the average reference budget from 4,096 to approximately 600 tokens, achieving a 2.4× wall-clock speedup with less than 3% degradation in DINO identity score. Beyond acceleration, our results provide a compact diagnostic of how in-context DiTs use reference images for personalized generation.

1. Introduction

Personalized image generation synthesizes novel images that preserve the visual identity of a reference subject while following a new text prompt (Ruiz et al., 2023; Gal et al., 2023; Ye et al., 2023). Recent Diffusion Transformers (DiTs) (Peebles & Xie, 2023; Esser et al., 2024) enable a simple and powerful form of personalization: the reference image is converted into tokens and concatenated directly with text and generation tokens. FLUX.1 Kontext (Black Forest Labs et al., 2025) exemplifies this in-context paradigm, preserving identity without per-subject fine-tuning, external adapters, or optimization at inference time.

The simplicity of in-context personalization comes with a significant computational cost. A standard text-to-image generation pass processes roughly 4,608 tokens, including text and generation tokens. Appending a reference image adds another 4,096 tokens, expanding the sequence to 8,704 tokens and increasing the quadratic attention cost from about 21M to 76M entries, or approximately 3.6×. Across many transformer blocks and denoising steps, this overhead becomes a practical bottleneck.

This motivates a natural question: *are all reference tokens necessary for identity preservation?* In a typical reference image, many tokens correspond to background or prompt-irrelevant regions, while only a subset captures identity-defining evidence such as shape, texture, color pattern, or facial structure. We find that FLUX.1 Kontext's own attention patterns support this intuition: generation tokens concentrate over 80% of their reference attention mass on the top 25% of reference tokens. The high-score tokens are also spatially aligned with foreground subject regions.

We introduce ReCAST (**Re**ference-**C**onditioned **A**daptive **S**parse **T**okens), a training-free sparse reference conditioning method. ReCAST profiles the full sequence once, ranks reference tokens by *cross-image attention*, and denoises with a timestep-adaptive subset that uses fewer tokens in coarse early steps and more tokens in detail-sensitive late steps. It requires no retraining, architectural changes, or additional modules.

[*]Equal contribution. [1] Department of Biomedical Engineering, Chungnam National University College of Medicine, Daejeon, Republic of Korea. [2] Interdisciplinary Program in Medical Informatics, Seoul National University College of Medicine, Seoul, Republic of Korea. [3] Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea. [4] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea. Correspondence to: Dongheon Lee <dhlee.jubilee@gmail.com>.
Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

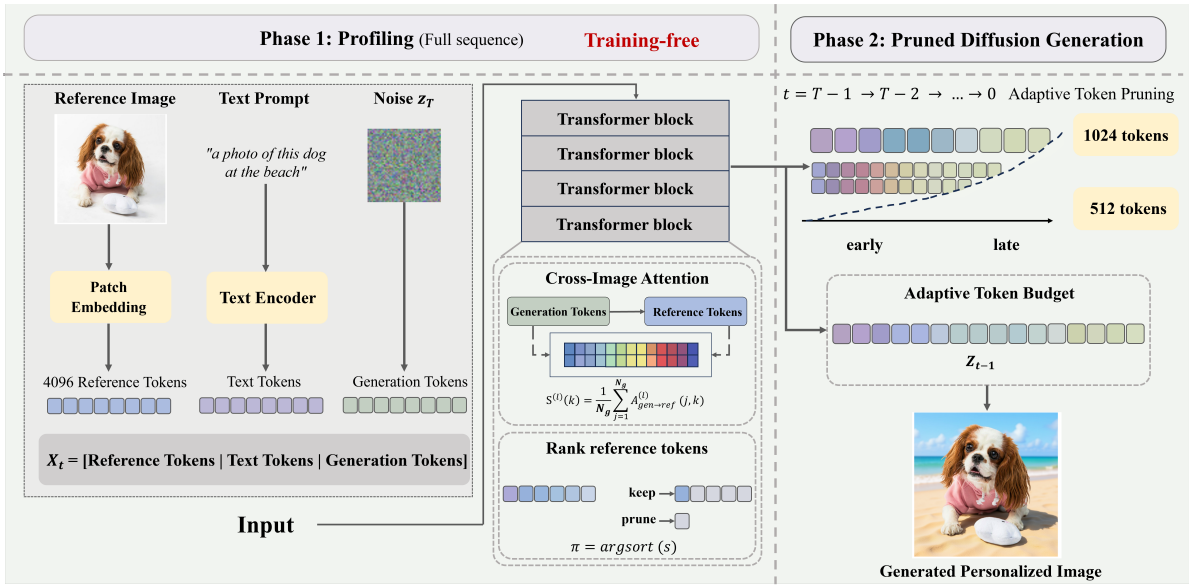

Figure 1. **System overview of ReCAST.** Phase 1 ranks reference tokens from one full-sequence profiling step using generation-to-reference attention already computed by the DiT. Phase 2 denoises with a timestep-adaptive sparse reference set. The method is training-free, requires no architectural modification, and uses $N_{\min} = 256$, $N_{\max} = 1{,}024$ (average $\sim$600 reference tokens) in our experiments.

Our goal is both methodological and diagnostic: to accelerate personalized DiT generation while revealing how in-context generative models use reference context. ReCAST achieves a 2.4$\times$ wall-clock speedup while retaining 97.3% of the full model's DINO identity score. The results suggest that identity transfer is driven by a sparse, foreground-aligned reference signal rather than uniform use of the full image.

2. Methods

2.1. System Overview

Figure 1 summarizes ReCAST. For a reference image, text prompt, and noisy generation tokens, FLUX.1 Kontext processes the concatenated sequence

$$\mathbf{X}_t = [\mathbf{x}_{\text{ref}};\ \mathbf{x}_{\text{txt}};\ \mathbf{x}_{\text{gen},t}], \tag{1}$$

where $\mathbf{x}_{\text{ref}} \in \mathbb{R}^{N_r \times d}$, $\mathbf{x}_{\text{txt}} \in \mathbb{R}^{N_t \times d}$, and $\mathbf{x}_{\text{gen},t} \in \mathbb{R}^{N_g \times d}$. In FLUX.1 Kontext, $N_r = N_g = 4{,}096$. ReCAST first ranks reference tokens during one full-sequence profiling step, then denoises with a timestep-adaptive top-$N'_r(t)$ sparse reference subset. Only reference tokens are pruned; text and generation tokens are kept intact.

2.2. Reference Token Redundancy

In the single-stream blocks, all tokens interact through joint self-attention. Let $\mathbf{A}^{(l)} \in \mathbb{R}^{N \times N}$ be the attention matrix at block l, with $N = N_r + N_t + N_g$, and let $\mathbf{A}^{(l)}_{\text{gen}\to\text{ref}} \in \mathbb{R}^{N_g \times N_r}$ denote the generation-to-reference submatrix. We score reference token k by the average attention it receives from generation tokens:

$$s^{(l)}(k) = \frac{1}{N_g}\sum_{j=1}^{N_g} \mathbf{A}^{(l)}_{\text{gen}\to\text{ref}}(j,k). \tag{2}$$

The resulting distribution is highly skewed: the top 25% of reference tokens receive over 80% of the total attention mass. As shown in Figure 2, high-importance tokens spatially coincide with the foreground subject, while many background tokens receive near-zero attention. This non-uniform usage pattern indicates that most reference tokens are redundant for identity conditioning and motivates pruning by generation-conditioned relevance rather than by generic visual saliency.

2.3. Cross-Image Attention Scoring

Given a reference image, we run one profiling step with the full reference token set and read out the already-computed attention weights $\mathbf{A}^{(l)}_{\text{gen}\to\text{ref}}$. For each reference–prompt pair, the ranking is computed once during profiling and reused for the remaining denoising steps. We aggregate scores across the single-stream blocks $\mathcal{L}$ of FLUX.1 Kontext:

$$s(k) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} s^{(l)}(k), \tag{3}$$

where $\mathcal{L}$ corresponds to blocks 20–56, which expose direct cross-image interactions. Reference tokens are ranked by $s(k)$ in descending order, and the top-N'_r tokens are retained.

This score is not generic image saliency. A visually salient background patch can be irrelevant to preserving the requested subject, while a small eye, logo, or fur marking can receive high generation-to-reference attention because it is repeatedly queried during denoising. This distinction explains why ToMe (Bolya et al., 2023), which merges visually similar tokens within an image, and FastV-style pruning (Chen et al., 2024), which uses self-attention saliency, underperform in our ablations.

2.4. Timestep-Adaptive Reference Budget

The score in Equation (3) determines *which* reference tokens to retain. A complementary question is *how many* tokens to retain at each denoising step. A fixed budget is suboptimal because the diffusion trajectory is coarse-to-fine (Ho et al., 2020; Rombach et al., 2022): early timesteps establish global structure, while later timesteps refine textures and identity-specific details. We empirically observe the same trend in cross-image attention. The entropy of the normalized reference-token attention distribution decreases from roughly 11.2 bits at high noise to 8.4 bits near the final step, indicating that late denoising depends on a more focused set of identity-relevant tokens.

We therefore use a monotonic budget schedule

$$N'_r(t) = N_{\min} + (N_{\max} - N_{\min}) \left(1 - \frac{t}{T}\right)^{\gamma}, \tag{4}$$

where $t = T$ is the initial high-noise step and $t = 0$ is the final step. Unless otherwise stated, we set $N_{\min} = 256$, $N_{\max} = 1{,}024$, and $\gamma = 2$, yielding approximately 600 retained reference tokens per step on average. With this budget, the average sequence length becomes $600 + 512 + 4{,}096 = 5{,}208$ instead of $4{,}096 + 512 + 4{,}096 = 8{,}704$, reducing the theoretical attention cost from 75.8M to 27.1M entries. Accounting for profiling and non-attention overheads, the measured wall-clock speedup is 2.4×.

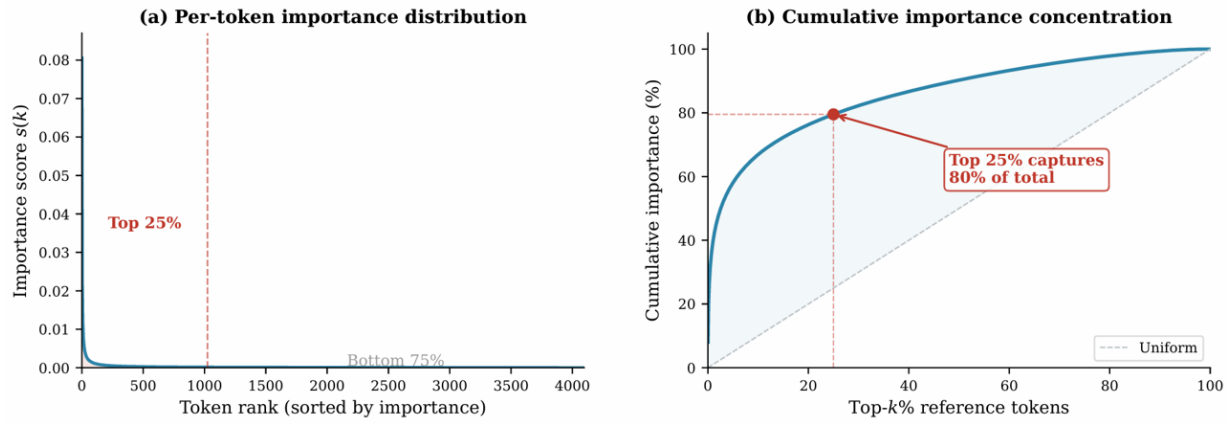

Figure 2. **Reference token attention distribution in FLUX.1 Kontext.** Per-block importance scores $s^{(l)}(k)$ from Equation (2), aggregated over 100 reference-prompt pairs, show that the top 25% of reference tokens receive over 80% of generation-to-reference attention mass. High-importance tokens consistently overlap with foreground subject regions.

Table 1. **Main comparison averaged over KontextBench and DreamBench++.** All sparse-reference methods use an average reference budget of ∼600 tokens per step except Full. Best results among sparse-reference methods are in **bold**.

Method	DINO ↑	CLIP-I ↑	AuraFace ↑	CLIP-T ↑	rFID ↓	Time (s) ↓	Speedup
Full (4096)	0.782	0.861	0.724	0.314	–	14.2	1.0×
Random	0.618	0.753	0.534	0.308	28.7	5.9	2.4×
ToMe (Bolya et al., 2023)	0.691	0.812	0.612	0.311	19.4	6.1	2.3×
Attn-Prune (Chen et al., 2024)	0.703	0.819	0.631	0.310	17.1	5.9	2.4×
ReCAST (fixed)	0.748	0.843	0.689	0.312	9.6	6.0	2.4×
ReCAST (adaptive)	**0.761**	**0.851**	**0.701**	**0.313**	**7.2**	**5.9**	**2.4**×

Table 2. **Component ablation on KontextBench.** All variants use an average budget of ∼600 reference tokens per step.

Configuration	DINO ↑	CLIP-T ↑	rFID ↓
ReCAST	**0.761**	**0.313**	**7.2**
Fixed budget ($N_r' = 600$)	0.748	0.312	9.6
Single-block score	0.739	0.312	10.8
Self-attention score	0.703	0.310	17.1
Cosine similarity score	0.695	0.311	18.7
Random selection	0.618	0.308	28.7

3. Experiments

3.1. Setup and Baselines

We evaluate FLUX.1 Kontext-dev (Black Forest Labs et al., 2025) at $1{,}024 \times 1{,}024$ resolution with 28 denoising steps on a single NVIDIA A100 GPU (80GB). We use KontextBench (Black Forest Labs et al., 2025), comprising 1,026 image-prompt pairs across subject preservation, style transfer, background change, and related editing tasks, and DreamBench++ (Peng et al., 2024), containing 150 subjects with 9 prompts each. We report DINOv2 feature similarity (Oquab et al., 2024), CLIP image-image similarity (CLIP-I) and CLIP text-image similarity (CLIP-T) (Radford et al., 2021), AuraFace face-embedding similarity for face subjects (Black Forest Labs et al., 2025; isidentical, 2024; Deng et al., 2019), reference FID (rFID) (Heusel et al., 2017) computed between pruned and full-model output sets, and wall-clock time.

Full keeps all 4,096 reference tokens. Random uniformly samples reference tokens. ToMe (Bolya et al., 2023) merges similar reference tokens by cosine similarity. Attn-Prune adapts FastV (Chen et al., 2024) by ranking reference tokens using intra-reference self-attention. All pruning methods use a matched average budget of approximately 600 reference tokens per step.

3.2. Main Results

Table 1 shows that the choice of reference tokens is crucial. Random selection reduces DINO from 0.782 to 0.618, indicating that identity evidence is concentrated in specific tokens rather than uniformly distributed over the reference image. ToMe and Attn-Prune improve over Random but still lose substantial fidelity because they do not measure generation-conditioned reference relevance. ReCAST with the adaptive schedule retains 97.3% of the full model's DINO score, keeps CLIP-I and CLIP-T nearly unchanged, and reduces per-image wall-clock time from 14.2s to 5.9s. The gap between ReCAST and Attn-Prune is especially important: both methods use attention and the same token budget, but only cross-image attention identifies reference tokens that are actually used by the generation process.

Figure 3 gives a qualitative view of the same effect. Random selection can produce a different subject because identity-defining tokens are removed indiscriminately. ToMe and Attn-Prune retain more visually salient content, but they can blur object boundaries or miss fine details such as fur texture and facial markings. By ranking tokens with generation-to-reference attention, ReCAST preserves identity-relevant features comparably to the full model across environment changes and accessory edits.

3.3. Ablation Study

Table 2 confirms that cross-image scoring is the dominant component. Replacing it with reference self-attention reduces DINO from 0.761 to 0.703, while cosine similarity gives 0.695. Removing the adaptive schedule gives 0.748, showing a smaller but consistent benefit. These results support ReCAST's central design choice: reference tokens should be ranked by how strongly generation tokens query them, not by intrinsic saliency or visual similarity.

3.4. Additional Analysis

Budget sensitivity. A fixed-budget sweep shows that identity preservation degrades gracefully as the reference budget is reduced from 2,048 to 512 tokens, then drops more sharply at 256 tokens. At 2,048, 1,024, 512, and 256 fixed tokens, DINO is 0.774, 0.758, 0.741, and 0.704, while speedup is 1.5×, 2.0×, 2.6×, and 2.9×, respectively. This supports our adaptive range of 256–1,024 tokens: the method operates in a high-efficiency regime while allocating more capacity to late, detail-sensitive denoising steps.

Ranking stability and amortization. When the same reference is paired with multiple prompts, importance rankings are stable but not identical: the mean Spearman correlation is 0.912 over 10 references and 15 prompts. Reusing a ranking reduces DINO by only 0.006 relative to per-prompt profiling and amortizes the profiling step, with speedup

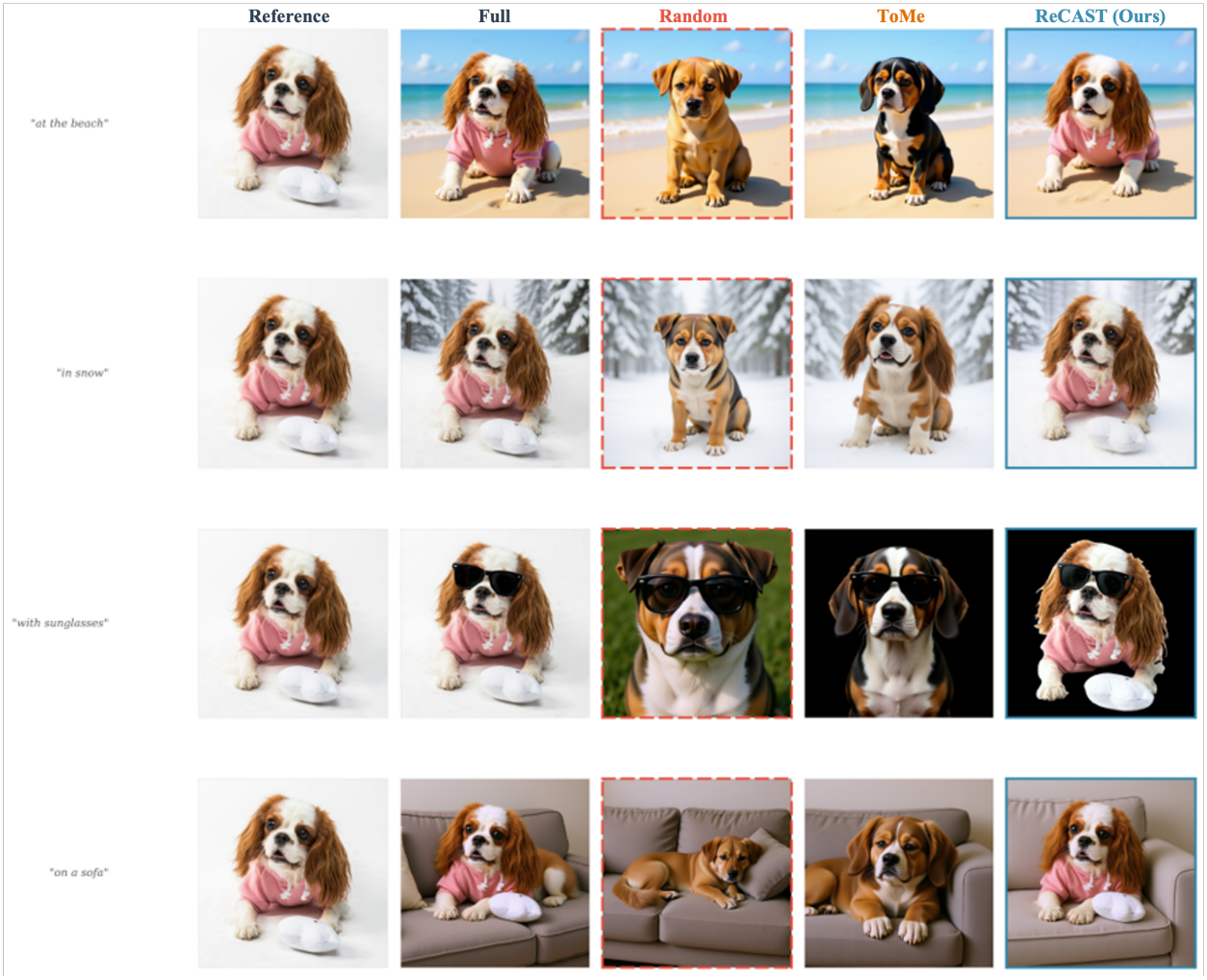

Figure 3. **Qualitative comparison.** Random selection often removes identity-defining details, and similarity-based selection can blur object boundaries. ReCAST preserves foreground identity features comparably to the full model while using far fewer reference tokens.

approaching 2.59× for 100 generations from the same reference. This suggests that foreground structure dominates the reference signal, while prompts still modulate fine identity cues.

Cross-model evidence and limitations. The sparse-conditioning pattern also appears in an IP-Adapter (Ye et al., 2023) FLUX.1 variant with 128 image-condition tokens: the top 25% of reference tokens capture about 72% of generation-to-reference attention mass, and pruning to 64 tokens retains 97.6% of the full DINO score. ReCAST is strongest for localized foreground subjects; style transfer has the largest DINO gap (-0.027) because style cues are spatially distributed. The ranking is computed once from a profiling step and reused, so recomputing it at every denoising step could improve fidelity but would reduce speedup.

Compatibility. ReCAST removes redundant context tokens within each step, whereas feature caching skips redundant computation across steps. Combined with ToCa (Zou et al., 2025), speedup increases from 2.41× to 3.38× with DINO 0.752, indicating that sparse reference conditioning is complementary to existing DiT acceleration techniques.

Overall, ReCAST is an efficient personalized-generation method whose sparse-conditioning behavior also reveals a useful model property: FLUX.1 Kontext can preserve identity with approximately 600 reference tokens because it relies on a sparse, foreground-aligned subset of the conditioning image rather than uniform access to all reference tokens.

Acknowledgements

This work was supported by the Artificial Intelligence Graduate School Program (Seoul National University); and by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

Black Forest Labs, Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., and Smith, L. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your ViT but faster. In *Int. Conf. Learn. Represent.*, 2023.

Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., and Chang, B. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Eur. Conf. Comput. Vis.*, 2024.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4690–4699, 2019.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorber, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Int. Conf. Mach. Learn.*, 2024.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Int. Conf. Learn. Represent.*, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, 2020.

isidentical. Introducing AuraFace: Open-source face recognition and identity preservation models. Hugging Face Blog, 2024. URL https://huggingface.co/blog/isidentical/auraface.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, 2023.

Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., and Xia, S.-T. DreamBench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22500–22510, 2023.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Zou, C., Liu, X., Liu, T., Huang, S., and Zhang, L. Accelerating diffusion transformers with token-wise feature caching. In *Int. Conf. Learn. Represent.*, 2025.