

Peer Review

Review of: "Accelerating Vision Diffusion Transformers with Skip Branches"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper starts from a practical observation. Feature caching across denoising steps works well for U-Net diffusion models because the long skip connections let a model recompute only shallow layers and reuse deep features, while caching methods ported to Diffusion Transformers give smaller speedups at a larger quality cost. The authors attribute this to what they call "feature smoothness": in a vanilla DiT, the block outputs change sharply between timesteps, so there is little to reuse. Their fix has two parts. **Skip-DiT** adds long skip branches from block i to block $L+1-i$ (§3.3, aggregation by `Linear(Norm(concatenation))`). **Skip-Cache** then, at "local" timesteps, runs only the first and last blocks and feeds the cached output of block $L-1$ from the previous "global" timestep through the first skip branch (§3.4, Figure 3c).

The evidence covers four settings: class-to-video on FaceForensics, SkyTimelapse, UCF101, and Taichi-HD with Latte (Table 1, Table 2); text-to-video with Latte continually trained on 330k Vimeo-2M clips (Table 3); text-to-image with Hunyuan-DiT, which already has skip branches (Table 4); and class-to-image with DiT-XL/2 retrained with skips (Appendix §6, Tables 7 and 8). The headline figures in the abstract are a 1.5× speedup "almost for free" and a 2.2× speedup with a minor metric loss. Code is linked on GitHub.

2. Whether the conclusion holds on the data

I like the core idea a lot. It is simple, it reuses a mechanism that already works for U-Nets (the paper is candid in Appendix §8 that Skip-Cache "shares significant similarities with DeepCache"), and the authors test it across video and image, class- and text-conditioned generation, and four backbones. The Hunyuan-DiT experiment is especially convincing to me because there the skip branches come with the

pre-trained model and no retraining muddies the comparison. The text of §4.3 reports a 2.09× speedup with no FID degradation on that model. The ablation in Table 5 is also well designed: it asks the right question (does caching in the smooth region matter?) and the answer is clear. Caching over 700→50 gives PSNR 29.52, LPIPS 0.06, SSIM 0.89, while 950→300 gives 20.58, 0.23, and 0.73 (Table 5, columns PSNR, LPIPS, SSIM). That is a large, interpretable gap.

I checked what I could, and several things need tightening before the conclusions match the evidence.

Most tables are not legible in the Qeios version. As rendered on Qeios, Tables 1, 2, 3, 4, 6, and 8 appear only as captions; their cells did not come through. I therefore could not verify the per-task FVD/FID values behind §4.2, the VBench/PSNR/LPIPS/SSIM rows behind the text-to-video claims, or the Hunyuan-DiT and class-to-image rows. Every number I quote below for those experiments comes from the running text, and I would ask the authors to make sure the tables are uploaded as text rather than images, so a reader can check them.

Training-free versus trained. This is the point I would most like the authors to address. §4.2 says the 2.20× text-to-video speedup is "the highest speedup among current training-free DiT acceleration works." But Skip-DiT is not training-free. The class-to-video models are trained "from scratch without any initialization" (§4.1, Training Details), and the text-to-video model needs roughly one day of skip-branch training plus about five more days of overall training on 8 H100 GPUs (the same section; §4.2 summarises this as "six days of continual pre-training on 330k training samples"). Skip-Cache itself is training-free once Skip-DiT exists. The baselines (T-GATE, Δ-DiT, FORA, PAB) are all run on the original Latte (§4.1, last sentence of "Implementation Details of Other Caching Methods"). So the comparison is between a retrained model plus a cache and an untrained cache on the original model. That is a fair thing to report. It should just be framed as such, with the training cost next to the speedup.

The quality cost of conversion. In the same vein, the text-to-video comparison reports that Skip-DiT reaches a VBench score of 75.60 against 76.14 for the original Latte (§4.2, "Text-to-video Generation"). The sentence attributes this to Skip-Cache, but from context, it describes the uncached Skip-DiT. The 0.12% and 0.22% VBench losses for 1.49× and 1.72× are then measured relative to Skip-DiT, not to Latte. I checked the first one against Table 5: $75.60 \times (1 - 0.0012) = 75.51$, which matches Table 5, row 700→50, VBench. Relative to the original Latte, the loss is $(76.14 - 75.51) / 76.14 = 0.83\%$. I would state both reference points in the abstract because "almost for free" is true against Skip-DiT but less so against the model users actually start from.

Table 5 does not say which cache interval or speedup it uses. The text of §4.4 says the rows are compared "under equivalent throughput," but no column gives N, latency, or speedup. I also could not reconcile row 800→50 (VBench 75.36) with the text: §4.2 reports a 1.17% loss for the 800→50 setting at 2.20×, and $75.60 \times (1 - 0.0117) = 74.72$, not 75.36. Most likely, Table 5 uses a smaller N than the 2.20× configuration. Please add N and speedup to Table 5 so the two can be matched.

The link between Figure 2 and caching is argued, not shown. Figure 2 perturbs the *parameters* along two random directions (§3.2, Eq. 3) and plots the change in predicted noise. Caching, however, depends on how features change *across timesteps* with fixed parameters. The two are related but not the same quantity. Figure 4 (cosine similarity of same-layer features across timesteps) is the direct evidence, and I found it more persuasive. I would move Figure 4 forward, present Figure 2 as supporting intuition, and report Figure 2 on more than one dataset and seed of random directions, since at present it rests on Taichi and two timesteps (10 and 250).

No variability is reported. None of the FVD, FID, or VBench numbers carry a standard deviation or an indication of how many samples or seeds were used. Some of the gaps the paper calls meaningful are very small. In Table 7, row *cfg*=1.5, FID is 2.30 for DiT-XL/2 and 2.29 for Skip-DiT, and Precision and Recall are identical (0.83 and 0.58). The Appendix §6 text calls this "significantly outperforms" and "far exceeds." At *cfg*=1.0, the gap is larger (FID 9.49 vs 8.37, IS 122.49 vs 127.63, Table 7), and I think that row supports a modest claim of better training efficiency. At *cfg*=1.5, I would say "matches." Similarly, the FID loss of 0.02 reported for Skip-Cache at 1.46× (Appendix §6, Accelerating Evaluation) is likely within run-to-run noise, which is good news for the method but should be presented with an estimate of that noise.

An arithmetic mismatch in the training-cost claim. Appendix §6 states that Skip-DiT is trained for 2,900,000 steps against 7,000,000 for DiT-XL/2 (Table 7, column Steps: 2900k and 7000k) and that this is "only 38% of its training steps." $2,900 / 7,000 = 0.414$, so the figure is 41.4%, not 38%. If the 38% refers to wall-clock or GPU-hours (skip branches add some compute per step, so this would go the other way), please say so and give the measurement.

Class-to-video numbers are internally consistent where I could check. §4.2 reports Taichi FVD 163.06 → 167.54 under Skip-Cache, a loss of 4.48, which fits comfortably with the stated average FVD loss of 2.37 across the four tasks. I could not check the other three tasks without Table 1. One small wording point: the 2.19× result with N=3 is said to cover "the other three class-to-video tasks," which excludes Taichi; it would help to say why.

Novelty framing. §4.1 describes Hunyuan-DiT as "the first text-to-image DiT model with skip branches". Long skip connections between shallow and deep transformer blocks in diffusion models appeared earlier in U-ViT (Bao et al., CVPR 2023). The paper does cite U-ViT (reference [24] in §2), but only as one of the DiT models for image generation, not as a prior design with long skips. The contribution of this paper stands without that claim, since the novelty lies in using the skips for cross-timestep caching. I would soften the sentence and mention U-ViT at that point.

3. Reproducibility

The authors have done the most useful thing already: the code is public (abstract, GitHub link). The baseline configurations are also documented in Table 9 (Δ -DiT) and Table 10 (T-GATE, $m = 20$, $k = 2$ for both Latte t2v and Hunyuan-DiT t2i), and Appendix §8 explains the change made to FORA (no initialization step). That level of care about baselines is not common, and I appreciated it. In Table 9, the header symbols after "Diffusion steps" did not survive the Qeios rendering, so I cannot tell which of the numbers (e.g., 12, 28, 21 for Latte t2v) is the boundary b and which is the cached block count.

What I could not find in the text:

- Training hyperparameters for Skip-DiT: learning rate, batch size, optimizer, number of steps per stage. §4.1 gives durations only ("approximately one day", "approximately two days", "an additional three days").
- Checkpoint identifiers for the starting models. Which Latte text-to-video release was used, and which Hunyuan-DiT version? Both have had several public releases.
- The list of 330k Vimeo-2M clips sampled for training (§4.1). Without it, the text-to-video model cannot be retrained exactly.
- Evaluation sample counts and seeds: how many videos per class-to-video task for FVD, how many VBench prompts and samples per prompt. For MS COCO, the count (30,000 images) is given, which is good.
- Latency protocol: batch size, warm-up, number of timed runs, precision (fp16/bf16). Speedups are measured on an A100 for Tables 1, 3, and 4, but on an H100 with batch size 8 for Table 8, so the two sets are not directly comparable.
- For Table 1, whether the Latte baselines are the official checkpoints or models retrained by the authors under the same schedule as Skip-DiT. This decides whether "Skip-DiT consistently outperforms Latte in terms of FVD" (§4.2) is an architecture effect or a training effect.

I would also resolve one inconsistency about the Figure 2 model. §1 describes adding skip branches "to pre-trained DiT model" with "minimal continuous pre-training", while §3.3 says the Taichi model with skip branches was trained "from scratch".

4. *What specifically to fix*

1. Upload Tables 1, 2, 3, 4, 6, and 8 in a form Qeios renders as text. At present, only their captions are visible.
2. **Abstract and §4.2, last paragraph of "Text-to-video Generation"**: drop "training-free" for the 2.20× claim, or qualify it clearly. Put the Skip-DiT training cost (about six days on 8 H100 GPUs, §4.1) in the same sentence as the speedup.
3. §4.2: correct "Skip-Cache achieves a comparable VBench score (75.60 vs. 76.14)" to Skip-DiT if that is what is meant, and report the VBench loss of each cached configuration against the original Latte as well as against Skip-DiT (for 700→50 this is 0.83% versus 0.12%).
4. Table 5: add N and measured speedup per row, and explain why row 800→50 (75.36) differs from the 1.17% loss at 2.20× in §4.2.
5. Appendix §6: 2,900k / 7,000k steps is 41.4%, not 38%. Either fix the number or define the cost it refers to.
6. Appendix §6 and Table 7 caption: soften "significantly outperforms," "far exceeds" (text), and "significantly exceeds" (caption) for cfg=1.5, where FID differs by 0.01 and Precision/Recall are equal.
7. Report a spread (standard deviation over seeds, or a bootstrap interval) for FVD, FID, and VBench in at least the main configurations of Tables 1, 3, and 4.
8. §3.2–3.3: say explicitly that Figure 2 measures sensitivity to parameter perturbation, and point to Figure 4 as the direct timestep-wise evidence. Repeating Figure 2 with a second dataset would help.
9. §4.1, first paragraph: soften "the first text-to-image DiT model with skip branches" and point to U-ViT (already reference [24]) as earlier work with long skips.
10. State whether the Latte baselines in Table 1 are official weights or retrained.
11. Small corrections I noticed. §4.1, Evaluation Details, says "For class-to-image tasks, we evaluate ... using FVD"; this should read class-to-video. §4.3 says "All videos are generated at a resolution of 1024" for the text-to-image experiment. "PAB469" appears in §4.2 and again in the Figure 5 caption, which looks like a leftover reference marker. The Figure 8 caption names Hunyuan-DiT as the backbone for class-to-image, whereas Appendix §6 uses DiT-XL/2.

12. List training hyperparameters, checkpoint versions, and the latency protocol (see section 3 above) in an appendix.

5. Recommendation

Minor revisions. The method is sound, simple to adopt, released as code, and supported by a well-chosen ablation (Table 5) and a clean test on a model that already has skips (Hunyuan-DiT). What needs work is the framing rather than the experiments: the training cost of Skip-DiT belongs next to every "training-free" and "almost for free" claim, the reference point for quality losses should include the original model, and a few numbers (41.4% rather than 38%, the Table 5 versus §4.2 mismatch) need correcting. With the tables made legible and variability reported, I would be glad to see this accepted.

Declarations

Potential competing interests: No potential competing interests to declare.