OMNITTT: LINEAR COMPLEXITY MULTIMODAL VISUAL GENERATION WITH TEST-TIME TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal generative models jointly model text, images and videos within a shared context, enabling flexible visual generation but producing increasingly long token sequences dominated by visual tokens. Test-Time Training (TTT) offers a promising alternative to quadratic attention by compressing long contexts into an online-updated inner model with linear complexity. However, directly converting attention to TTT in pretrained generative models leads to substantial performance degradation under lightweight continual training. We attribute this difficulty to a modality asymmetry: language tokens are relatively few and inexpensive to retain with attention, yet difficult to compress without information loss, whereas visual tokens dominate computation and can be compressed effectively with the appropriate architectural design. Based on this observation, we introduce OmniTTT, a simple TTT-based architecture for efficient multimodal generation, which models visual interactions with TTT while retaining minimal attention for language. OmniTTT inherits weights from pretrained full-attention models and can be adapted through lightweight continual training without the need for training from scratch. Extensive experiments on text-to-image, text-to-video, and omnimodal reference-to-video generation demonstrate that OmniTTT substantially improves efficiency, achieving, for example, a 8.75$\times$ attention kernel and 2.87 $\times$ full model speedup on reference-to-video generation with comparable quality.

1 INTRODUCTION

Unified generative models that operate over omni-modal contexts comprising text, images and videos have emerged as a promising approach for state-of-the-art visual generation (Labs et al., 2025; Team et al., 2025; Seedance et al., 2026; MiniMax, 2026), unified multimodal (Deng et al., 2025), and world model (Agarwal et al., 2026; Team et al., 2026; Hu et al., 2026) systems. By jointly modeling heterogeneous modalities within a shared context, these models support flexible image and video generation, editing, and reference-conditioned synthesis. However, they also introduce increasingly long token sequences dominated by visual tokens, which can extend to 64K and beyond. The quadratic complexity of standard attention therefore becomes a critical computational bottleneck, particularly for high-resolution generation, long-video synthesis, and reference-rich generation, where large numbers of visual tokens must be jointly modeled over long ranges.

Test-Time Training (TTT) (Sun et al., 2025; Zhang et al., 2025e; Han et al., 2025) offers a promising approach to addressing this bottleneck. Rather than explicitly retaining the entire context for pairwise attention, TTT compresses token sequences into the parameters of an online-updated inner model that can subsequently be queried. This formulation replaces quadratic token-to-token interactions with state compression, enabling sequence modeling with linear complexity over context length. Moreover, given the prohibitive cost of training large-scale generative models from scratch with a new sequence-modeling architecture, converting existing pretrained full-attention models through lightweight continual training provides a practical pathway to adopting TTT.

However, naively converting pretrained attention layers to TTT leads to substantial performance degradation under lightweight continual training. Through systematic preliminary studies, we identify three observations that guide our design: (i) compressing language tokens incurs noticeable information loss while providing only limited computational savings; (ii) directly compressing visual tokens with vanilla MLP-based TTT adapts poorly under limited continual training; and (iii)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

incorporating local interactions via window attention into the visual TTT design substantially improves conversion. These observations reveal a clear modality asymmetry: language tokens are few and inexpensive to retain, yet difficult to compress without information loss, whereas visual tokens dominate computation and can be compressed more effectively under the proper architecture design.

Building on these observations, we introduce OmniTTT, a simple and efficient TTT-based architecture for multimodal visual generation that explicitly accounts for the asymmetry across modalities. OmniTTT retains minimal softmax attention for language tokens, while modeling the computationally dominant visual context with linear complexity through a global TTT module augmented by local tiled window attention. This enables TTT to be deployed throughout the transformer stack while preserving compatibility with pretrained weights. By inheriting pretrained weights, OmniTTT can be efficiently adapted through lightweight continual training without training from scratch, forming, to our knowledge, the first full-stack TTT-based architecture for complex multimodal generation.

We systematically evaluate OmniTTT across text-to-image, text-to-video, and omnimodal reference-to-video generation using three different backbones. Across these settings, OmniTTT substantially improves efficiency while preserving generation quality relative to full-attention baselines. At the kernel level, OmniTTT achieves an **8.75**$\times$ attention speedup for video reference-to-video generation. At the full-model level, it reduces latency by **7.30**$\times$ for 4K text-to-image generation, **2.56**$\times$ for 480p text-to-video generation, and **2.87**$\times$ for reference-to-video generation.

2 PRELIMINARY

Softmax Attention. Given queries $Q \in \mathbb{R}^{N_q \times d}$, keys $K \in \mathbb{R}^{N_k \times d}$, and values $V \in \mathbb{R}^{N_k \times d}$, softmax attention (Vaswani et al., 2017) computes

$$\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V. \tag{1}$$

Its computational complexity is $\mathcal{O}(N_q N_k d)$ due to the pairwise query-key interactions. In the self-attention setting with $N_q = N_k = N$, this reduces to $\mathcal{O}(N^2 d)$, resulting in quadratic scaling with sequence length and substantial computational overhead for long token sequences.

Test-Time Training. Test-Time Training (TTT) instead models sequence interactions through an input-conditioned inner model f_W (Sun et al., 2025; Han et al., 2025), typically parameterized as an MLP. Rather than explicitly retaining all key-value pairs, TTT compresses them into the parameters W of the inner model. Given keys K and values V, the inner model is updated by

$$W \leftarrow W - \eta \nabla_W \mathcal{L}\left(f_W(K), V\right), \tag{2}$$

where η is the inner-loop learning rate and $\mathcal{L}$ denotes the inner loss. Queries are then applied to the updated inner model to obtain the final output:

$$\mathrm{TTT}(Q, K, V) = f_W(Q). \tag{3}$$

With a fixed-size inner model, TTT replaces explicit pairwise attention with parameterized sequence compression, yielding linear complexity with respect to sequence length.

3 APPROACH

3.1 MOTIVATION: A PRELIMINARY ABLATION

A key challenge in converting pretrained multimodal generative models to TTT is determining which modality interactions can be effectively compressed without compromising generation quality. In unified architectures such as MMDiT, however, language and visual tokens are jointly modeled within a shared sequence, coupling language and visual interactions within the same attention operation. This makes it difficult to isolate the effect of compressing language or visual context individually. We therefore begin with a controlled study on Wan2.1-T2V-1.3B (Wan et al., 2025), which adopts a decoupled CrossDiT architecture where visual dependencies are modeled through self-attention and language conditioning through cross-attention (Esser et al., 2024). This separation enables us to independently examine the effect of TTT on different modality interactions, providing a clean basis for extending to unified architectures.

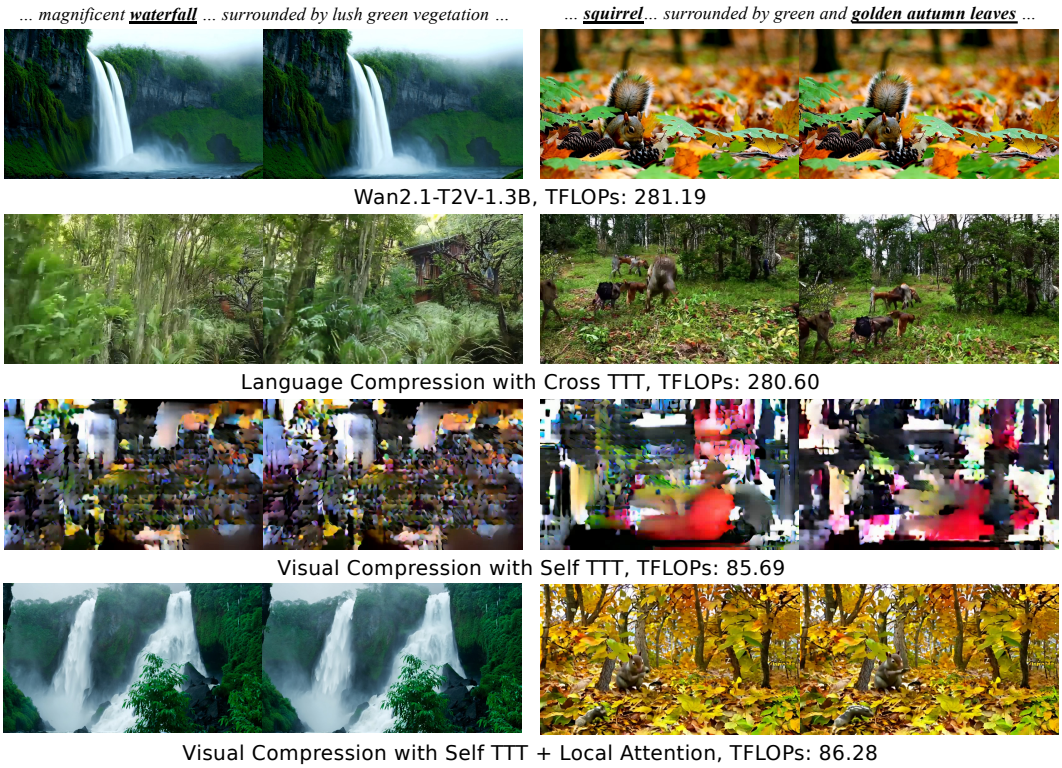


Figure 1: Qualitative video generation results from the preliminary study of different TTT replacement variants based on Wan2.1-T2V-1.3B.

Starting from the pretrained transformer, we retain all original weights and replace all selected attention modules with simplified SwiGLU-MLP TTT modules following (Han et al., 2025). All variants are continually trained on Koala-36M (Wang et al., 2025) for 2,000 steps at 480p resolution. We visualize the generated samples and report the computational cost of a single forward pass at 480p, 121 frames, and batch size 1 in Figure 1.

Observation 1: Compressing language tokens is lossy and brings limited computational savings. We first replace all cross-attention modules with the simplified SwiGLU-MLP TTT modules, which compress textual key-value representations into the inner model and subsequently queried by visual tokens. As visualized in Figure 1, although the converted model converges within the short continual-training stage, generation results exhibit substantially degraded instruction alignment. In particular, the left sample captures the overall scene with lush green vegetation but omits the key "waterfall" specified in the prompt, while the right sample reproduces the forest background but fails to capture the requested "golden autumn leaves". These examples suggest that compressing language representations can preserve coarse scene semantics while losing fine-grained semantic constraints that are critical for instruction following.

We attribute this to the high information density of language tokens: individual tokens often encode distinct objects, attributes, or relations that directly constrain different aspects of the generated content. Compressing these distinct constraints into a shared inner model may preserve the overall semantics while losing specific details, leading to the omission of prompt elements. Moreover, compressing language interactions yields limited computational savings. Cross-modal attention has complexity $\mathcal{O}(NLd)$, and since the number of language tokens L is typically much smaller than the number of visual tokens N, it constitutes only a small fraction of the overall computation. Replacing cross-attention with TTT therefore results in only a marginal reduction in FLOPs.

Observation 2: Directly compressing visual tokens with vanilla MLP-based TTT is difficult to adapt under a limited continual-training budget. We next replace all visual self-attention modules with simplified SwiGLU-MLP TTT modules, such that visual key-value representations are compressed. This configuration substantially reduces computation FLOPs, yet the converted model fails to converge within the limited continual-training budget, with generated outputs remaining

dominated by noise. This suggests that MLP-based TTT requires a substantially larger training budget to converge than the lightweight continual-training stage provides.

Observation 3: Combining visual compression with local information from window attention enables efficient conversion. Prior studies on linear-complexity attention (Han et al., 2023; 2024) suggest that linear modules can be less effective at preserving fine-grained local interactions. Motivated by this observation, we augment MLP-based visual TTT with lightweight tiled local window attention (Zhang et al., 2025d), retaining explicit local interactions while using TTT to model global context. As shown in Figure 1, this design enables the converted model to generate coherent visual content within the same continual-training budget. Notably, the generated samples also preserve the key semantics that are omitted when compressing language tokens. Additionally, the additional computational cost of local window attention is marginal, increasing from 85.69 to only 86.28 TFLOPs for a single forward pass. These results demonstrate that visual tokens can be effectively compressed under a lightweight conversion budget when augmented with explicit local interactions.

3.2 OMNITTT

Motivated by these observations, we instantiate OmniTTT using two complementary pathways, with lightweight attention modeling language interactions and TTT modeling visual interactions. The overall design is conceptually illustrated in Figure 2.

Let X_{T} and X_{V} denote the hidden states of language and visual tokens at the input of an OmniTTT layer. We first project them into queries, keys, and values:

$$Q_m = X_m W_Q, \quad K_m = X_m W_K, \quad V_m = X_m W_V, \qquad m \in \{\mathrm{T}, \mathrm{V}\}, \tag{4}$$

where W_Q, W_K, and W_V denote the QKV projection matrices.

Language Attention. For language tokens, we retain softmax attention over the textual context:

$$O_{\mathrm{T}}^{\mathrm{Attn}} = \mathrm{Attn}\left(Q_{\mathrm{T}}, K_{\mathrm{T}}, V_{\mathrm{T}}\right). \tag{5}$$

Since the language sequence is relatively short, this preserves explicit interactions among language tokens with little computational overhead.

Visual TTT. The visual pathway uses TTT to model global visual dependencies with linear complexity. Following (Han et al., 2025), visual keys and values are compressed into an online-updated inner model, which is queried by both language and visual tokens:

$$[O_{\mathrm{T}}^{\mathrm{TTT}}; O_{\mathrm{V}}^{\mathrm{TTT}}] = \mathrm{TTT}\left([Q_{\mathrm{T}}; Q_{\mathrm{V}}], K_{\mathrm{V}}, V_{\mathrm{V}}\right). \tag{6}$$

We adopt the simplified SwiGLU inner model with a dot-product inner loss objective:

$$f_W(x) = \mathrm{FC}(x) \odot \mathrm{SiLU}\left(\mathrm{FC}(x)\right). \tag{7}$$

The resulting TTT memory provides both language and visual tokens with access to global visual context with linear complexity over the visual sequence.

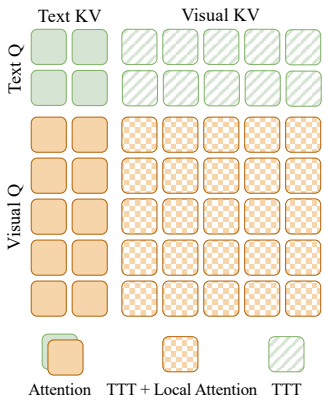


Figure 2: Conceptual illustration of OmniTTT.

To complement the global context modeled by TTT with fine-grained local visual interactions, we further incorporate tiled local window attention following (Zhang et al., 2025d). We partition the visual tokens into non-overlapping spatial or spatiotemporal tiles $\{\mathcal{G}_m\}_{m=1}^{M}$ with fixed size. For each tile $\mathcal{G}_m$, we construct a local window containing all neighboring tiles within a predefined radius r:

$$\mathcal{W}_r(m) = \{\mathcal{G}_n \mid \mathrm{dist}(\mathcal{G}_m, \mathcal{G}_n) \le r\}. \tag{8}$$

For a visual token $i \in \mathcal{G}_m$, its local neighborhood therefore consists of all visual tokens contained in these tiles:

$$\mathcal{N}(i) = \bigcup_{\mathcal{G}_n \in \mathcal{W}_r(m)} \mathcal{G}_n. \tag{9}$$

All queries within the same tile share the same local neighborhood. This tile-wise organization yields a regular block-structured attention pattern, allowing queries within each tile to be processed

Table 1: **Quantitative results on HPSv3.** We report the overall preference score and scores across semantic categories. Ani.: Animals. Arc.: Architecture. Cha.: Character. Des.: Design. Sce.: Natural Scenery. Oth.: Other. Sci.: Science. Pla.: Plants. Pro.: Products. Tra.: Transportation.

Model	TFLOPs↓	All↑	Ani.	Arc.	Arts	Cha.	Des.	Food	Sce.	Oth.	Sci.	Pla.	Pro.	Tra.
SD3.5-M	10.58	8.53	9.09	8.92	8.08	8.99	7.50	9.24	8.26	8.57	8.88	8.60	7.34	8.90
+SFT	10.58	8.74	8.95	9.30	8.17	9.64	7.39	9.59	8.43	8.88	9.09	9.07	7.14	9.24
T^5	8.72	8.48	8.79	9.02	7.80	9.59	6.84	9.26	8.32	8.76	8.90	8.95	6.61	8.93
OmniTTT	8.92	**8.73**	8.83	9.44	8.03	9.67	6.99	9.40	8.75	8.88	9.33	8.92	7.15	9.39
OmniTTT-Fast	**7.77**	8.60	8.80	9.48	8.09	9.53	6.89	9.17	8.78	8.63	9.15	8.45	7.03	9.20

Table 2: **FLOPs and Latency** for text-to-image generation under different resolutions.

Resolution	SD3.5-M		T^5		OmniTTT-fast	
	TFLOPs↓	Latency/ms↓	TFLOPs↓	Latency/ms↓	TFLOPs↓	Latency/ms↓
1024^2	10.58	99.21	8.72	72.16	7.77 (1.36×)	62.24 (1.59×)
2048^2	87.79	593.64	58.12	403.84	30.74 (2.86×)	216.84 (2.74×)
4096^2	1083.12	6237.74	607.56	3484.77	122.60 (8.83×)	854.03 (7.30×)

jointly and mapped efficiently to parallel block-wise kernels. Since both the tile size and the window radius are fixed, the number of local tokens attended to by each query remains bounded as the visual sequence grows, maintaining linear complexity with respect to the overall visual sequence length.

Fused Multimodal Attention. For simplicity, we implement language conditioning and local visual aggregation into a fused single softmax attention. Specifically, each visual query attends to all language tokens together with the visual tokens in its local neighborhood:

$$O^{\mathrm{Attn}}_{\mathrm{V},i} = \sum_{j\in\mathcal{T}\cup\mathcal{N}(i)} \frac{\exp\left(q^{\top}_{\mathrm{V},i}k_j/\sqrt{d}\right)}{\sum_{\ell\in\mathcal{T}\cup\mathcal{N}(i)}\exp\left(q^{\top}_{\mathrm{V},i}k_\ell/\sqrt{d}\right)} v_j, \tag{10}$$

where $\mathcal{T}$ indexes all language tokens.

Gated Aggregation. Finally, we combine the attention and TTT outputs using modality-specific token-wise gates:

$$[\alpha_{\mathrm{m}};\beta_{\mathrm{m}}] = \sigma(X_{\mathrm{m}}W^{\mathrm{m}}_{G}), \qquad m\in\{\mathrm{T},\mathrm{V}\}. \tag{11}$$

$$O_{\mathrm{m}} = \alpha_{\mathrm{m}}\odot O^{\mathrm{Attn}}_{\mathrm{m}} + \beta_{\mathrm{m}}\odot O^{\mathrm{TTT}}_{\mathrm{m}}, \qquad m\in\{\mathrm{T},\mathrm{V}\}. \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The resulting representations combine language and local visual interactions from attention with the compressed global visual context modeled by TTT.

4 Experiments

4.1 Experimental Setup

Training. For text-to-image generation, we initialize from the the pretrained StableDiffusion3.5-Medium (Esser et al., 2024) model model and train for 6K iterations with a global batch size of 128 on 1.75M FLUX-generated synthetic text-image pairs (Le, 2025). The pretrained backbone contains 24 MMDiT blocks, the first 13 of which include an additional image-only self-attention layer. We evaluate two conversion variants: OmniTTT replaces the joint multimodal attention modules, whereas OmniTTT-Fast additionally replaces the image-only self-attention layers. This comparison isolates the effect of extending conversion beyond joint multimodal attention.

For text-to-video generation, we initialize from HunyuanVideo1.5-T2V-8B (Wu et al., 2025a) and train on Koala-36M (Wang et al., 2025) for 10K iterations with a global batch size of 64 at 480p resolution and 5s length. For multimodal reference-to-video generation, we train MiniMax-H3-Ref2VA (MiniMax, 2026) with a mixture of text-to-video, video-reference, and multi-image reference-to-video training data. Details for the pretrained model architecture and additional implementation details are described in Section A.1 and Section A.2 respectively.



Figure 3: Qualitative text-to-video generation results based on HunyuanVideo1.5-8B.

Table 3: **Quantitative text-to-video generation results** on VBench and VideoAlign. IQ: Imaging Quality, OC: Overall Consistency, AQ: Aesthetic Quality, SC: Subject Consistency, BC: Background Consistency, MS: Motion Smoothness. VQ: Visual Quality, MQ: Motion Quality.

Model	TFLOPs↓	Latency/ms↓	VBench							VideoAlign	
			Ovr.↑	IQ	OC	AQ	SC	BC	MS	VQ	MQ
HunyuanVideo1.5	1319.21	2421	80.38	65.10	24.98	60.24	93.24	90.07	98.67	2.30	0.38
+SFT	1319.21	2421	81.54	67.79	25.10	57.43	93.86	92.35	98.19	2.13	0.41
OmniTTT	336.52	947	81.90	67.63	25.58	57.61	93.21	93.14	98.50	2.29	0.39
Δ	3.92 ×	2.56 ×	+0.36	-0.16	+0.48	+0.18	-0.65	+0.79	+0.31	+0.16	-0.02

Evaluation Protocol. For text-to-image generation, we report HPSv3 (Ma et al., 2025) preference scores on the HPD prompt set with four random seeds per prompt for robustness and report latency measured on a single NVIDIA A100 GPU. For text-to-video generation, we evaluate on VBench (Huang et al., 2024) and report VideoAlign (Liu et al., 2025) scores using prompts from the MovieGen (Polyak et al., 2024) evaluation set. Following the default HunyuanVideo1.5-T2V-T2V-8B protocol, we rewrite video prompts with the Qwen3-235B-A22B-Thinking-2507 (Yang et al., 2025a) prompt enhancement model. We report full transformer computational cost in TFLOPs and measure inference latency for a single forward with batchsize 1 on a single NVIDIA H100 GPU.

4.2 Main Results

Text-to-Image Generation. We evaluate whether OmniTTT can achieve compatible generation quality while improving efficiency, and compare its quality-efficiency trade-off with existing TTT-based methods. We compare against the pretrained model, its full-attention counterpart fine-tuned under the same setting (+SFT), and T^5 (Li et al., 2026), which converts approximately half of the attention modules with MLP-based TTT. As shown in Table 1, OmniTTT closely matches the full-attention SFT baseline while outperforming T^5 at comparable computational cost. Extending TTT to the image-only attention layers yields OmniTTT-Fast, which further reduces computation while outperforming both the pretrained model and T^5. Table 2 further shows that OmniTTT-Fast consistently reduces FLOPs and latency, with larger gains at higher resolutions. Compared with the full-attention backbone, its latency speedup increases from $1.59\times$ at 1024^2 to $7.30\times$ at 4096^2, where it also reduces FLOPs by $8.83\times$.

Text-to-Video Generation. We further evaluate whether OmniTTT can be effectively adapted to text-to-video generation. As shown in Table 3, OmniTTT achieves comparable performance compared to the full-attention SFT baseline on both VBench and VideoAlign, while reducing full model FLOPs by $3.92\times$ and latency by $2.56\times$. These results indicate that OmniTTT extends effectively to video generation, substantially reducing computational cost while maintaining comparable quality.



Figure 4: Qualitative multi-image reference-to-video results based on MiniMax-H3 Ref2VA.



Figure 5: Qualitative video reference-to-video generation results based on MiniMax-H3 Ref2VA.

Table 4: Kernel efficiency for different reference-to-video generation tasks.

Model	Multi-Image		Video	
	TFLOPs	Latency/ms	TFLOPs	Latency/ms
FlashAttn3	51.56	74.67	144.54	213.53
OmniTTT	5.70	14.44	9.37	24.41
	(9.04×)	(5.17×)	(15.42×)	(8.75×)

Table 5: Full model efficiency for different reference-to-video generation tasks.

Model	Multi-Image		Video	
	PFLOPs	Latency/s	PFLOPs	Latency/s
FlashAttn3	4.22	7.11	9.97	16.29
OmniTTT	1.92	3.41	3.21	5.67
	(2.19×)	(2.09×)	(3.11×)	(2.87×)

Omnimodal Reference-to-Video Generation. We evaluate the efficiency of OmniTTT for both multi-image and video reference-to-video generation at the attention-kernel and full transformer level in Table 4 and Table 5, respectively. OmniTTT achieves $5.17\times$–$8.75\times$ speedup at the attention-kernel level and $2.09\times$–$2.87\times$ full transformer level acceleration. Qualitative comparisons in Figure 4 and Figure 5 further show that OmniTTT maintains comparable generation quality to the pretrained and SFT full-attention models across both reference settings.

4.3 Ablation Studies and Analysis

Effect of TTT and Local Attention. Table 6 decomposes the roles of TTT and local attention in visual modeling. TTT alone yields an HPSv3 score of only 0.07, consistent with our preliminary finding that directly converting visual attention to TTT is difficult to adapt under a limited continual-training budget. Local attention achieves meaningful generation quality, but remains substantially below the joint configuration, supporting the complementary roles of local attention in preserving explicit local interactions and TTT in modeling global visual context.

Comparison with Linear Attention. We compare TTT with linear attention (Katharopoulos et al., 2020) and DeltaNet (Yang et al., 2024) as alternative linear-complexity mechanisms for global visual modeling. As shown in Table 7, both alternatives achieve lower HPSv3 scores than OmniTTT. This result suggests that TTT is better suited to the lightweight conversion setting, providing stronger generation quality while retaining linear complexity.

Gated Aggregation. Table 8 studies different gating mechanisms for combining local-attention and TTT outputs. Token-wise gating improves HPSv3 over direct aggregation with marginal increased computational cost, whereas channel-wise gating provides no further gain despite higher FLOPs. We therefore adopt token-wise gating as the default configuration.

Table 6: Effect of TTT and local attention for processing visual tokens.

TTT	Local Attention	TFLOPs↓	HPSv3↑
✓	✗	8.08	0.07
✗	✓	8.23	7.25
✓	✓	8.92	8.73

Table 7: Comparison for TTT with alternative linear-complexity attention variants.

Model	TFLOPs↓	HPSv3↑
Linear Attention	8.29	7.86
DeltaNet	8.38	8.09
OmniTTT	8.92	8.73

Table 8: Effect of different gating mechanism for combining attention and TTT outputs.

Gating	TFLOPs↓	HPSv3↑
No	8.91	8.63
Channel-wise	9.85	8.69
Token-wise	8.92	8.73

Table 9: Effect of different memory capacity size for visual TTT.

Memory Bandwidth	TFLOPs↓	HPSv3↑
1×	8.80	8.31
2×	8.92	8.73
4×	9.15	8.44

Table 10: Comparison with different sparse attention approaches on HunyuanVideo1.5-8B.

Model	Sparsity↑	TFLOPs↓	Latency/ms↓	VBench Ovr.↑	VA VQ↑	VA MQ↑
FlashAttention3	0%	1319.21	2421	80.38	2.30	0.38
VSA	90.00%	411.31	1301	78.02	2.07	0.14
SSTA	66.98%	627.56	1633	79.36	2.17	0.24
SSTA	81.87%	473.79	1281	78.94	2.07	0.17
SSTA	89.31%	396.90	1100	78.81	2.00	0.12
SLA	89.31%	401.01	1177	79.27	1.98	0.15
OmniTTT	97.24%	336.52	947	81.90	2.29	0.39

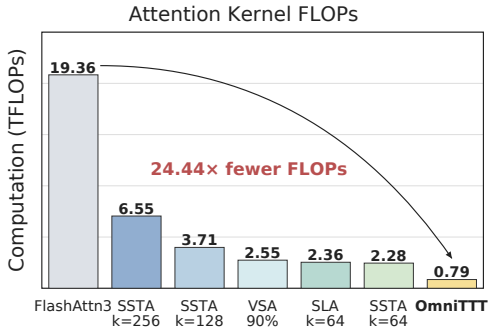


Figure 6: Computational cost of attention kernels in TFLOPs.

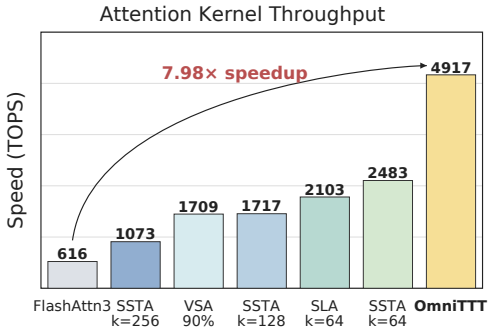


Figure 7: Dense-equivalent throughput of attention kernels in TOPS.

Memory Capacity. We further study the capacity of the visual TTT memory by varying its memory bandwidth. For an MLP-based inner model $f_W : \mathbb{R}^d \to \mathbb{R}^d$, we define the memory bandwidth as the number of parameters of the inner model and progressively increase it from $1\times$ to $4\times$. As reported in Table 9, increasing the bandwidth from $1\times$ to $2\times$ substantially improves generation quality, whereas further increasing it to $4\times$ provides no additional benefit.

Comparison with Attention Acceleration Methods. We further compare OmniTTT with representative attention acceleration methods, including VSA (Zhang et al., 2025c), SSTA (Wu et al., 2025a), and SLA (Zhang et al., 2025a), with all methods trained on the same data. We evaluate VSA at 90% sparsity and SSTA with top-$k \in \{64, 128, 256\}$. For OmniTTT, we also report the equivalent sparsity that refers to the fraction of interactions omitted by its fixed local attention component. As shown in Table 10, OmniTTT achieves stronger VBench and VideoAlign performance while also attaining the lowest full model FLOPs and latency among the evaluated methods.

Additionally, we benchmark the efficiency of attention kernel, in Figure 6 and Figure 7. We report theoretical efficiency in FLOPs and practical efficiency in TOPS (trillion operations per second), where TOPS is computed as the dense-equivalent operation count of the complete attention module divided by its measured latency. As reported, OmniTTT reduces kernel computation FLOPs by

$24.44\times$ and improves throughput by $7.98\times$ relative to the dense FlashAttention-3 baseline, while also surpassing all evaluated sparse-attention baselines.

5 Related Works

5.1 Multimodal Visual Generation

Visual generation has evolved from text-conditioned synthesis (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; BlackForest, 2024; Wan et al., 2025) toward unified omnimodal generation under rich contextual inputs, where models synthesize visual content conditioned on text, reference images and video frames. Recent systems such as FLUX-Kontext (Labs et al., 2025), Qwen-Image-2.0 (Zhao et al., 2026), Kling-Omni (Team et al., 2025), Seedance 2.0 (Seedance et al., 2026) and MiniMax-H3 (MiniMax, 2026) demonstrate the flexibility of multimodal context for controllable image and video creation. More broadly, unified multimodal models and world-model systems extend this paradigm toward interleaved perception, generation, and action across omnimodal inputs (Deng et al., 2025; Agarwal et al., 2026; Team et al., 2026; Hu et al., 2026).

5.2 Test-Time Training

Test-time training (TTT) (Sun et al., 2025) has emerged as a promising paradigm for efficient sequence modeling. Instead of explicitly storing and attending to all tokens, TTT maintains an online-updated inner model, which serves as a compact and dynamically updated memory of the preceding context. This formulation enables linear-complexity inference over long sequences while providing a more expressive memory mechanism than conventional recurrent states. Recent works have extended TTT along multiple directions, including scalable training and inference mechanisms (Zhang et al., 2025e; Tandon et al., 2025), long video generation (Dalal et al., 2025), spatial and 3D perception (Liu et al., 2026; Chen et al., 2025b; Xie et al., 2026), and test-time discovery for complex problem solving (Yuksekgonul et al., 2026). Most relevant to our work, ViT^3 (Han et al., 2025) and T^5 (Li et al., 2026) investigate TTT as a linear-complexity replacement for attention in vision Transformers.

5.3 Attention Acceleration for Visual Generation

Attention acceleration is central to scaling visual generation. A major line of work exploits sparse attention approaches in diffusion Transformers, including training-free sparse inference methods such as Sparse VideoGen (Xi et al., 2025), Sparse VideoGen2 (Yang et al., 2025b), SpargeAttention (Zhang et al., 2025b), and sliding-tile attention (Zhang et al., 2025d). Other works make sparse attention trainable or adaptive during model optimization, including VSA (Zhang et al., 2025c), SpargeAttention2 (Zhang et al., 2026a), and VMoBA (Wu et al., 2025b). SLA (Zhang et al., 2025a) and SLA2 (Zhang et al., 2026b), combine sparse and linear attention to better preserve generation quality during acceleration. Another line directly replaces quadratic attention with linear-complexity architectures, as explored in SANA (Xie et al., 2024) and SANA-Video (Chen et al., 2025a).

6 Conclusion

In this paper, we present OmniTTT, a simple TTT-based architecture for efficient multimodal visual generation. Our analysis reveals a modality asymmetry inherent in unified generation models: language interactions are better preserved with explicit attention, whereas visual interactions can be effectively compressed with the appropriate architectural design. Based on this insight, OmniTTT combines global visual TTT with lightweight language attention, enabling full-stack conversion under lightweight continual training. Across text-to-image, text-to-video, and omnimodal reference-to-video generation, OmniTTT substantially reduces computation while maintaining comparable generation quality, achieving up to an $8.75\times$ attention kernel and $2.87\times$ full model speedup for reference-to-video generation. These results establish OmniTTT as a practical new approach for processing long visual contexts in unified multimodal generative models.

AI USE STATEMENT

In this work, we used generative AI tools for polishing paper writing and in assisting the profiling and optimization of triton-based attention/TTT kernels. We have not used generative AI tools for retrieval and discovery, research ideation, drafting sections of the paper, generating synthetic datasets and proving mathematical claims. We have reviewed all AI-assisted work. LLM-generated code was verified manually and under a series of automated unit tests involving comparison of bit-wise output precision and controlled latency evaluations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

BlackForest. Black forest labs; frontier ai lab, 2024. URL `https://blackforestlabs.ai/`.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025b.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17702–17711, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5961–5971, 2023.

Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. *Advances in Neural Information Processing Systems*, 37:79221–79245, 2024.

Dongchen Han, Yining Li, Tianyu Li, Zixuan Cao, Ziming Wang, Jun Song, Yu Cheng, Bo Zheng, and Gao Huang. Vit 3: Unlocking test-time training in vision. *arXiv preprint arXiv:2512.01643*, 2025.

Yucheng Hu, Jianke Zhang, Yuanfei Luo, Yanjiang Guo, Xiaoyu Chen, Xinshu Sun, Kun Feng, Qingzhou Lu, Sheng Chen, Yangang Zhang, et al. Bagelvla: Enhancing long-horizon manipulation via interleaved vision-language-action generation. *arXiv preprint arXiv:2602.09849*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Duong H Le. Flux generated dataset, 2025. URL `https://huggingface.co/datasets/lehduong/flux_generated`.

Yining Li, Dongchen Han, Zeyu Liu, Hanyi Wang, Yulin Wang, and Gao Huang. Linearizing vision transformer with test-time training. *arXiv preprint arXiv:2605.02772*, 2026.

Fangfu Liu, Diankun Wu, Jiawei Chi, Yimo Cai, Yi-Hsin Hung, Xumin Yu, Hao Li, Han Hu, Yongming Rao, and Yueqi Duan. Spatial-ttt: Streaming visual-based spatial intelligence with test-time training. *arXiv preprint arXiv:2603.12255*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *European Conference on Computer Vision*, pp. 361–377. Springer, 2024.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

MiniMax. Minimax h3, 2026. URL `https://github.com/MiniMax-AI/MiniMax-H3`.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. In *International Conference on Machine Learning*, pp. 57503–57522. PMLR, 2025.

Arnuv Tandon, Karan Dalal, Xinhao Li, Daniel Koceja, Marcel Rød, Sam Buchanan, Xiaolong Wang, Jure Leskovec, Sanmi Koyejo, Tatsunori Hashimoto, et al. End-to-end test-time training for long context. *arXiv preprint arXiv:2512.23675*, 2025.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8428–8437, 2025.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025a.

Jianzong Wu, Liang Hou, Haotian Yang, Xin Tao, Ye Tian, Pengfei Wan, Di Zhang, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models. *arXiv preprint arXiv:2506.23858*, 2025b.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Tao Xie, Peishan Yang, Yudong Jin, Yingfeng Cai, Wei Yin, Weiqiang Ren, Qian Zhang, Wei Hua, Sida Peng, Xiaoyang Guo, et al. Scal3r: Scalable test-time training for large-scale 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21760–21771, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025b.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Mert Yuksekgonul, Daniel Koceja, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, et al. Learning to discover at test time. *arXiv preprint arXiv:2601.16175*, 2026.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, et al. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. *arXiv preprint arXiv:2509.24006*, 2025a.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In *International Conference on Machine Learning*, pp. 76397–76413. PMLR, 2025b.

Jintao Zhang, Kai Jiang, Chendong Xiang, Weiqi Feng, Yuezhou Hu, Haocheng Xi, Jianfei Chen, and Jun Zhu. Spargeattention2: Trainable sparse attention via hybrid top-k+ top-p masking and distillation fine-tuning. *arXiv preprint arXiv:2602.13515*, 2026a.

Jintao Zhang, Haoxu Wang, Kai Jiang, Kaiwen Zheng, Youhe Jiang, Ion Stoica, Jianfei Chen, Jun Zhu, and Joseph E Gonzalez. Sla2: Sparse-linear attention with learnable routing and qat. *arXiv preprint arXiv:2602.12675*, 2026b.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025c.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025d.

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025e.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

A Appendix

A.1 Pretrained Model Architecture

We summarize the architecture of the three pretrained backbones used in our experiments.

StableDiffusion3.5-Medium. For text-to-image generation, we use StableDiffusion3.5-Medium (Esser et al., 2024), which comprises 24 MMDiT-X blocks. Each block maintains separate image and text streams with joint multimodal attention. The first 13 blocks additionally include image-only self-attention. Text conditioning is provided by CLIP-L/14, OpenCLIP bigG/14, and T5-XXL, with pooled CLIP embeddings supplying extra global condition. The backbone uses adaptive LayerNorm, query-key RMSNorm, and GELU feed-forward activations.

HunyuanVideo1.5-T2V-8B. For text-to-video generation, we use HunyuanVideo-1.5-T2V-8B (Wu et al., 2025a), which comprises 54 dual-stream MMDiT blocks with separate video and text branches coupled through joint attention. A causal 3D VAE encodes video inputs into latent tokens. HunyuanVideo-1.5-T2V-8B uses Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) as the semantic encoder and Glyph-ByT5-v2 (Liu et al., 2024) as the auxiliary glyph encoder. The backbone also comprises of timestep-conditioned adaptive LayerNorm, query-key RMSNorm, and GELU-tanh feed-forward activations.

MiniMax-H3 Ref2VA. For multimodal reference-to-video generation, we use MiniMax-H3-Ref2VA (MiniMax, 2026). Its Omni-Transformer comprises 50 single-stream blocks that apply shared attention and feed-forward layers to a packed sequence of text, visual, and audio tokens. The H3-Encoder, based on Qwen3-VL-32B (Bai et al., 2025a), encodes text and visual references, while separate visual and audio VAEs provide modality-specific latent representations. The backbone uses RMSNorm with timestep- and modality-specific adaptive modulation, query-key RMSNorm, and SwiGLU feed-forward activations.

A.2 Training Details

We present the detailed training hyper-parameters for different generation tasks in Table 11.

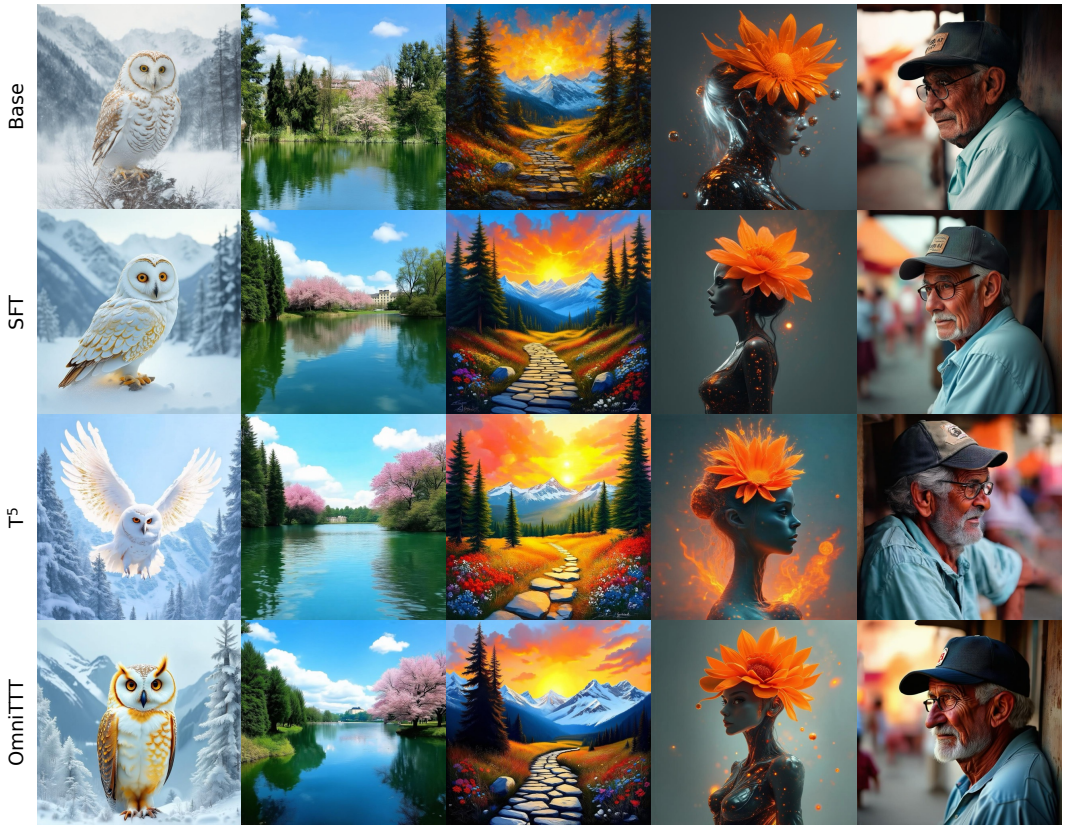


Figure 8: Qualitative text-to-image generation results based on StableDiffusion3.5-Medium.

Table 11: Detailed training hyperparameters for text-to-image (T2I), text-to-video (T2V), and reference-to-video (R2V) generation.

Hyper-parameter	T2I	T2V	R2V
Iterations	6K	10K	5K
Batch Size	128	64	64
Learning Rate	1e-4	1e-4	1e-5
Schedule	Constant	Constant	Constant
Warmup Steps	0	100	100
Optimizer	AdamW	Muon	AdamW
Weight Decay	1e-4	1e-2	1e-2
Grad Norm Clip	1.0	1.0	1.0
Window Attention Tile Size	(2, 2)	(4, 4, 4)	(4, 4, 4)
Window Attention Radius	1	1	1
TTT Inner Learning Rate	1.0	1.0	1.0

A.3 Additional Visualization Comparisons

We present qualitative visualization comparisons on text-to-image generation against the pretrained model, SFT full attention baseline and T^5 in Figure 8. Results indicate that OmniTTT achieves compatible generation quality with full attention baselines.

A.4 Limitations and Future Work

Given limited computational resources, our study focuses on architectures that support efficient conversion of pretrained full-attention models through lightweight continual training. This emphasis on pretrained-weight compatibility and rapid adaptation may leave unexplored designs that require

larger training budgets but offer better quality–efficiency trade-offs. Future work will extend this study to larger-scale training and native linear-complexity generative models trained from scratch, enabling architectural choices to be explored beyond the constraints of lightweight conversion.