STEERING VIDEO DIFFUSION TRANSFORMERS WITH MASSIVE ACTIVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

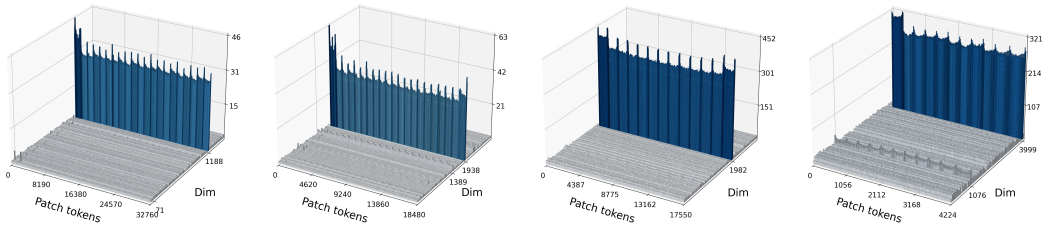
In this work, we study the role of Massive Activations (MAs), which are rare, high-magnitude spikes confined to a few fixed hidden dimensions in video diffusion transformers (DiTs). We uncover a structured positional hierarchy: MA magnitudes peak at first-frame tokens and recur at the spatial boundary tokens of latent frames, with this pattern being most pronounced during early denoising. We trace this organization to an encoding asymmetry of the video VAEs, whose causal temporal padding and zero spatial padding cause the first latent frame and frame borders to carry reduced content load. Elevated MAs consistently align with these lower-content structural positions. To understand their function, we analyze intermediate representations and find that MAs act as implicit rescalers of residual computation: enlarging MAs suppresses the corresponding self-attention and feed-forward updates, while erasing them amplifies these updates. Together, these observations suggest that MAs serve as a token-level rescaler of residual computation, which video DiTs deploy unevenly, placing the strongest damping at the encoding-asymmetric structural positions. Motivated by this native rescaling behavior, we propose **St**ructured **A**ctivation **S**teering (**STAS**), a training-free technique that steers MAs at the observed structural positions toward a scaled, model-derived reference during early denoising. STAS requires no additional forward passes and consistently improves video quality and temporal coherence across text-to-video models with negligible overhead.

1 INTRODUCTION

Recent advances in video generation have been driven by diffusion- and flow-based frameworks, evolving from U-Net based text-to-video models (Singer et al., 2023; Ho et al., 2022; Guo et al., 2024b; Blattmann et al., 2023; Chen et al., 2023; 2024; Bar-Tal et al., 2024; Zhou et al., 2022) to diffusion transformer (DiT) architectures (Yang et al., 2025b; HaCohen et al., 2024; Zheng et al., 2024; Lin et al., 2024; Jin et al., 2025b; Wan Team et al., 2025; Kong et al., 2024; Meituan LongCat Team et al., 2025; Ma et al., 2025) that scale better with model capacity and training data. Beyond architecture, video quality and temporal coherence are further improved through training-time objectives (Chefer et al., 2025; Jin et al., 2024; Zhang et al., 2025), post-training preference optimization (Liu et al., 2025; Cheng et al., 2025; Yang et al., 2025a; Bu et al., 2026), and inference-time guidance (Kim & Kim, 2026; Shaulov et al., 2025; Wu et al., 2024; Yuan et al., 2025; Bu et al., 2025; Xia et al., 2024; Chen et al., 2026; Hyung et al., 2025), though typically at the cost of extra training or inference computation.

In parallel, a separate line of research has uncovered striking activation phenomena inside transformers. In large language models, Massive Activations (MAs) are rare hidden-state activations with magnitudes orders of magnitude larger than typical values (Sun et al., 2024). They are often viewed as the counterpart of attention sinks (Xiao et al., 2024): where attention sinks concentrate on a few tokens, MAs concentrate in a few dimensions that stay persistently large across most tokens, which recent work terms residual sinks (Qiu et al., 2026). Similar phenomena have also been observed in vision transformers and image DiTs, where a small number of tokens or channels consistently carries disproportionate magnitude and plays an outsized role in model behavior (Darcet et al., 2024; Gan et al., 2026; Turri et al., 2026). Video DiTs, however, place tokens on a joint spatiotemporal grid, and it is unclear how far these observations carry over. We therefore ask: *(1)*

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.



(a) Wan2.1-1.3B (b) Wan2.2-5B (c) CogVideoX-5B (d) LTX2 (base)

Figure 1: **Video DiTs exhibit a structured positional hierarchy of Massive Activations.** Activation magnitudes across four video DiTs, averaged over 100 prompts and aggregated in bins of 60 consecutive tokens for visualization. Across all models, MA magnitudes peak at tokens of the first latent frame and recur at the spatial boundaries of subsequent latent frames.

how are MA magnitudes organized over the video-token layout, and (2) what function do they serve during generation?

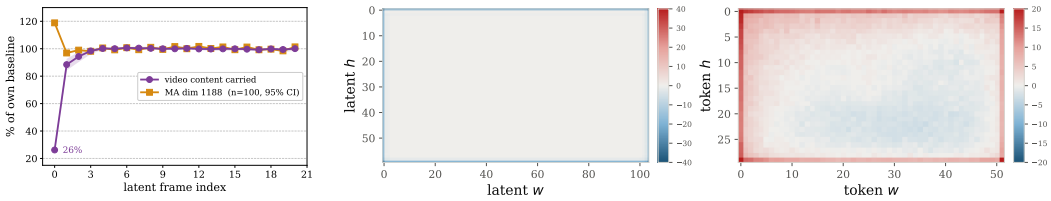
Taking a closer look at the internal activations of video DiTs, we find that MAs not only emerge, often exceeding $50\times$ the mean activation magnitude, but also exhibit a structured positional hierarchy (Figure 1). MA magnitudes are highest at tokens of the first latent frame, elevated at the spatial boundaries of subsequent latent frames, and comparatively moderate in their interiors, producing recurring peaks along the flattened token sequence. This pattern is consistent across Wan2.1-1.3B and Wan2.2-5B (Wan Team et al., 2025), CogVideoX (Yang et al., 2025b), and LTX-2 (HaCohen et al., 2026), spanning both standard DiT and MM-DiT designs as well as different latent compression ratios of $4\times8\times8$, $4\times16\times16$, and $8\times32\times32$.

We trace this positional hierarchy to the encoding asymmetries of the video VAE: causal temporal encoding leaves the first latent frame with only a fraction of the content load of a later frame, and zero spatial padding depresses frame borders, exactly where MAs are most elevated (Figure 2). We then probe the computational role of MAs with bidirectional interventions and find that raising MA magnitude weakens the corresponding self-attention and feed-forward residual updates, whereas zeroing it strengthens them. Read together, these findings expose a native behavior: video DiTs use MAs to damp residual writing precisely where their encoder supplies the least content. This echoes the broader tendency of transformers to place sinks on less informative tokens, such as delimiters in LLMs and background patches in ViTs (Sun et al., 2024; Darcet et al., 2024). In video DiTs, however, these positions are structural, fixed by the encoder's geometry rather than by content.

Motivated by this native rescaling behavior, we propose **St**ructured **A**ctivation **S**teering (**STAS**), a training-free technique that strengthens it, steering MAs at first-frame and boundary positions toward a scaled reference derived from the model's own activations during early denoising. STAS modifies no model parameters and requires no additional model evaluations. It involves only element-wise assignments at selected token positions, introducing negligible overhead of less than 0.1%. Across different baselines, STAS consistently improves video quality and temporal coherence.

Our contributions are summarized as follows:

- We systematically characterize MAs in video DiTs, revealing a positional hierarchy across token locations that is consistent across architectures and latent compression ratios, and we trace its origin to encoding asymmetries introduced by video VAEs.
- Through bidirectional interventions, we show that MA magnitude acts as an implicit rescaler of self-attention and feed-forward residual updates, with its strongest effects concentrated at the lower-content structural positions.
- We propose STAS, a training-free technique that exploits this structured rescaling signal, and demonstrate consistent improvements in temporal coherence and video quality across multiple text-to-video models, benchmark evaluations, and human preference studies, with negligible overhead and compatibility with other training-free methods.



(a) Per-frame content load vs. MA (b) Latent boundaries of the VAE (left) and the DiT MA map (right)

Figure 2: **The MA hierarchy aligns with video-VAE encoding asymmetries.** (a) Latent-frame ablation in Wan2.1 VAE shows that causal temporal padding leaves the first latent frame carrying less content than later frames, precisely where MA magnitudes peak. (b) A uniform video isolates spatial padding: the VAE latent has attenuated boundaries (left), and the DiT MA map (dimension 1188) is elevated at the corresponding boundary tokens (right; shown at token resolution, where each token aggregates a 2×2 patch of latent cells).

2 Related Work

Massive Activations. Massive activations have been reported across transformer families spanning language (Sun et al., 2024; Owen et al., 2025), vision (Darcet et al., 2024), and diffusion models (Gan et al., 2026; 2025; Shao et al., 2026; Fang et al., 2025). In Large Language Models (LLMs), such activations often emerge at fixed dimensions and are frequently associated with low-information tokens, and have been linked to contextual knowledge modeling in several analyses (Jin et al., 2025a). Similar patterns are observed in Vision Transformers (ViTs), where MAs are reported to arise in redundant or background-like tokens and correlate with global semantic encoding (Darcet et al., 2024; Yang et al., 2024). In DiTs, MAs have been discussed mainly due to their practical implications for acceleration techniques such as quantization (Shao et al., 2026; Liu & Zhang, 2024; Zhao et al., 2025) and distillation (Fang et al., 2025), and their presence can also affect extracted feature quality (Gan et al., 2025). Recent studies further examine their generative role in image DiTs, relating them to local-detail synthesis or prompt-conditioned semantic subspaces (Gan et al., 2026; Turri et al., 2026). However, how MA magnitudes are organized over spatiotemporal video tokens, and how they modulate residual computation across these positions, remain unexplored.

Improving Video Generation with Training-Free Methods. Prior works (Bu et al., 2025; Xia et al., 2024) improve temporal consistency in U-Net-based video generation models by manipulating attention maps. Others use external guidance from image generators (Guo et al., 2024a; Jagpal et al., 2025) or multimodal LLMs (Li et al., 2025). FreeInit (Wu et al., 2024) and FreqPrior (Yuan et al., 2025) refine noise priors to reduce spatiotemporal incoherence. Beyond external signals, some methods use the generator itself for guidance. Auto-guidance contrasts the original model with a weaker variant (Karras et al., 2024), obtained by skipping selected spatiotemporal layers (Hyung et al., 2025) or dropping DiT blocks (Chen et al., 2026). FlowMo (Shaulov et al., 2025) and TeDiO (Tursynbek et al., 2026) optimize intermediate latents using temporal variation and self-attention diagonals, respectively; SRVS (Jang et al., 2026) performs iterative predict-and-perturb updates, using prediction inconsistency to localize uncertain regions; and ANSE (Kim & Kim, 2026) uses attention-derived uncertainty to select initial noise. All, however, typically either require additional model evaluations or impose operations the model does not natively perform. Our method instead strengthens a regulation the model already implements, the MA-based rescaling of its least informative tokens, within existing forward passes and with negligible overhead.

3 Preliminaries: Video VAEs and DiTs

Modern video DiTs generate in the latent space of a spatiotemporal VAE. An input video is encoded causally: the first physical frame forms the first latent frame on its own, and every subsequent chunk of r_{temp} frames forms one more latent frame. The encoder pads asymmetrically, front-only in time and symmetric zeros in space. Consequently, the receptive fields of the first latent frame and of the outermost spatial positions contain more padding and less video content than those of later frames and interior positions. The latent is patchified (2×2 or 1×1 spatially) into video tokens

and denoised by a stack of transformer blocks conditioned on the text prompt and timesteps, trained under the flow-matching objective (Lipman et al., 2023). Finally, sampling proceeds iteratively over denoising steps $k = 0, 1, \ldots,$ from pure noise to the clean latent.

4 The Organization and Function of MAs in Video DiTs

Video DiTs inherit the MA properties reported for image DiTs (Gan et al., 2026): MAs persist across layers and model scales, concentrate on fixed channel dimensions, and remain largely invariant to text prompts (Figure 1). Beyond these, we study MAs along the token axis, answering the two questions posed in Sec. 1: how MA magnitudes are organized over the video-token layout, and what function they serve during generation.

Periodic MA Pattern Emerges along the Token Sequence. We record hidden states during generation over 100 prompts and examine MA magnitudes along the flattened token sequence (Figure 1, see Appendix A.1 for the measurement protocol). Each model typically exhibits a single dominant MA dimension. Their MA magnitudes follow a three-level hierarchy: highest at tokens of the first latent frame, elevated along the spatial boundaries of every subsequent latent frame, and moderate in frame interiors. Under row-major flattening, these boundaries appear as periodic head and tail bands. The hierarchy is robust across different video DiT designs and compression ratios.

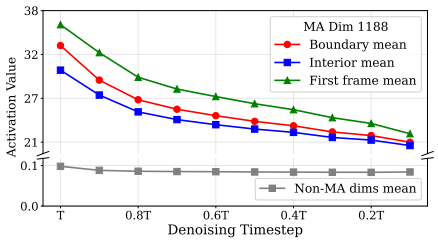


Figure 3: **The MA hierarchy is most pronounced during early denoising.**

The hierarchy has a characteristic timing: it is most pronounced during early denoising (Figure 3), the phase deciding global structure, and attenuates monotonically as sampling progresses, with the boundary-to-interior ratio falling $1.10 \rightarrow 1.02$. This motivates the early-step window of our method.

The Hierarchy Mirrors the VAE's Encoding Asymmetry. Why these positions? The padding facts from Sec. 3 predict exactly where the encoder supplies less content, and two controlled probes confirm that the MA hierarchy lands on those positions.

Temporally, we measure the content each latent frame carries with an ablate-and-decode probe. We encode natural videos, replace one latent frame at a time with its spatial mean, decode, and measure the total reconstruction damage on the physical frames it generates (formalized in Appendix B.3). At equal token count, the first latent frame, encoding a single physical frame, carries only about $1/4$ of a later frame's content load, and MA magnitudes peak exactly there (Figure 2a).

Spatially, we encode a video with no spatial structure, every frame a single constant color. The resulting latent still exhibits a depressed border ring (Figure 2b, left). Nothing in the input can cause it, so the ring must be the footprint of zero padding. The DiT MA map mirrors this ring (Figure 2b, right), and the correspondence is graded rather than binary: the four corners, padded twice over, show roughly twice the edge deficit and more than double the edge-level MA elevation. The same additivity holds across probes: border tokens of the first latent frame carry both shortfalls, and show the highest MA magnitudes in the sequence (Figure 1). Natural videos and re-encoded generations reproduce the ring (Appendix C), showing that this prior is present in the model's own outputs.

The hierarchy thus inherits the geometry of the video VAE encoder, concentrating MAs at positions with reduced content load. This echoes a broader tendency of transformers to offload onto less informative tokens, previously described as attention sinks (Sun et al., 2024; Darcet et al., 2024). Here it appears instead as a magnitude hierarchy on a single MA dimension, at positions fixed by the encoder rather than by the content of each input. The model locates them through RoPE (Su et al., 2024), and permuting the temporal or spatial indices relocates the MA peaks (Appendix D).

MA Magnitude Acts as a Token-Level Rescaler of Residual Writing. What do these elevated MAs do? In pre-normalized DiT blocks, a dominant MA occupies a disproportionate share of a token's normalized scale, attenuating the remaining content-bearing dimensions, consistent with the analysis in LLMs (Qiu et al., 2026). This attenuation directly enters the feed-forward and

attention-value pathways, whereas query/key normalization reduces scale variation but not directional changes. We give further discussions in Appendix E and hypothesize a two-fold rescaling effect: increasing a token's MA directly weakens its own residual writing and, through self-attention, can indirectly alter the attention updates received by other video tokens. In contrast, token-local computation at non-target positions and the unmodified text pathway should be less affected.

We test this hypothesis on Wan2.1-1.3B by steering the MA dimension at structural tokens, the first latent frame and the head-and-tail bands of later frames, which cover the contiguous sides of the border ring under flattening, toward α times its global maximum, sweeping α from 0 (erasure) to 5.5 over 100 paired generations (setup in Appendix B.3). In the following block, we measure post-normalization content and the self-attention and feed-forward residual writes. As shown in Figure 4, the structural tokens respond monotonically in both directions: MA erasure amplifies their writes to 1.3–1.5× the native values, whereas α=2.5 reduces their self-attention and feed-forward writes to 0.50× and 0.52×, respectively. The main non-target response appears in interior self-attention writes (1.07× under erasure and 0.89× at α=2.5), consistent with propagation through token mixing, while interior feed-forward writes remain near native. These results identify MA magnitude as a token-level rescaler of both self-attention and feed-forward writes, whose reach to other tokens runs through self-attention.

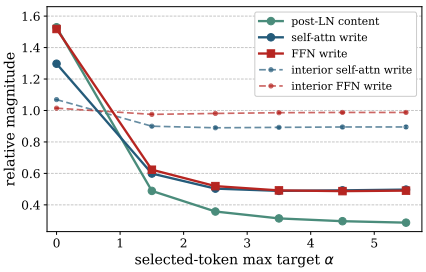


Figure 4: **MA magnitude is a token-level rescaler of residual updates.**

5 METHOD

The analysis in Sec. 4 reveals that MA magnitude acts as an implicit, token-level rescaler of residual computation, and that video DiTs deploy it unevenly, concentrating elevated MAs at VAE-aligned structural positions with reduced content load. We hypothesize that residual writing at these positions, when insufficiently damped, injects perturbations during early denoising that propagate through self-attention and persist along the denoising trajectory, surfacing as degraded first frames, incoherent motion, and abrupt object appearances and disappearances. This motivates us to strengthen the model's native rescaling at its existing structural locations.

We therefore propose STAS, a training-free method that specifies *what* to steer (MA dimensions), *where* to steer (VAE-aligned structural tokens), *how* to steer them (max-referenced calibration), and *when* to intervene (early denoising). Let $H^m_{k,i,d}$ denote the hidden dimension d of token i after DiT block m at sampling step k, where $k = 0$ denotes the first denoising timestep. STAS modifies only the model-specific MA dimensions $\mathcal{M}$ identified by activation profiling. Its target token set is

$$\mathcal{S} = \mathcal{F}_0 \cup \mathcal{B}(p), \tag{1}$$

where $\mathcal{F}_0$ contains all tokens in the first latent frame and $\mathcal{B}(p)$ contains the head and tail $p\%$ of the flattened tokens within each latent frame. In our default implementation, $\mathcal{B}(p)$ focuses on the horizontal boundary bands for two complementary reasons. Under row-major flattening, the head and tail rows form contiguous spans that capture the prominent recurring MA peaks, whereas the left and right columns are sparse and non-contiguous. Moreover, the tail band of one latent frame and the head band of the next lie on opposite sides of each latent frame-frame junction, making them a compact, ordering-aligned choice for steering around frame transitions.

Although structural positions exhibit higher MAs on average, their native magnitudes remain heterogeneous: some still fall within the interior-token range (as detailed in Appendix F). Multiplying each token by a fixed factor preserves this heterogeneity and may leave weak structural MAs comparatively small. STAS instead derives an adaptive reference from the current hidden state:

$$r^m_{k,d} = \max_{j \in \mathcal{V}} \left| H^m_{k,j,d} \right|, \tag{2}$$

where $\mathcal{V}$ denotes the video tokens. It then defines the steering target as

$$g^m_{k,i,d} = \alpha \cdot r^m_{k,d} \cdot \operatorname{sign}\left(H^m_{k,i,d}\right), \tag{3}$$

where α controls the steering strength. This max-referenced rule gives selected tokens a consistent target magnitude while adapting to each sample, layer, and denoising step, with their native polarity

preserved. As shown in Figure 5, MA erasure (zeroing out) produces a negative quality trend and fixed scaling yields only a modest improvement, whereas max-referenced calibration produces the clearest positive response.

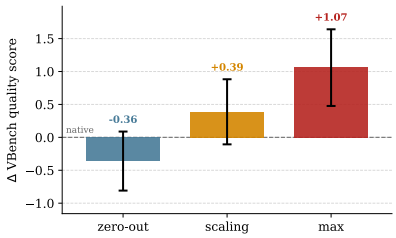


Figure 5: **Steering direction and calibration rule both matter.**

STAS is applied only during the first K denoising iterations, when the structural MA hierarchy is most pronounced and residual perturbations have more opportunity to affect the subsequent trajectory. Let $\mathcal{L}$ denote the selected DiT blocks. The complete update is

$$\widehat{H}^m_{k,i,d} = \begin{cases} g^m_{k,i,d}, & k < K,\ m \in \mathcal{L},\ i \in \mathcal{S},\ d \in \mathcal{M}, \\ H^m_{k,i,d}, & \text{otherwise}, \end{cases} \tag{4}$$

Eligible entries are therefore replaced by their calibrated targets, while all other activations follow the original denoiser unchanged. We provide the full algorithm in Appendix B.1. We also discuss our method through the lens of guidance in Appendix G.

In practice, we apply STAS on top of CFG (Ho & Salimans, 2021). By calibrating MA dimensions at the structural token positions (first-frame and boundary tokens), STAS complements CFG's global, prompt-level guidance with token-level regulation of residual writing.

6 Experiments

We evaluate STAS on four publicly available text-to-video models spanning different architectures and scales: Wan2.1-T2V-1.3B (Wan Team et al., 2025), Wan2.2-TI2V-5B (Wan Team et al., 2025), CogVideoX-5B (Yang et al., 2025b), and LTX-2 base model (HaCohen et al., 2026). All experiments use the officially released weights and default sampling configurations; unless otherwise specified, STAS is applied on top of the default CFG. Videos are generated at $81\times480\times832$ for Wan2.1, $49\times480\times720$ for CogVideoX-5B, $81\times704\times1280$ for Wan2.2, and $81\times512\times768$ for LTX-2. Implementation details are deferred to the Appendix B.2.

Evaluation Metrics. Our primary benchmark is VBench (Huang et al., 2024), which aggregates quality (seven dimensions) and semantic (nine dimensions) scores into a total score on a 0–100 scale. All VBench results are averaged over five random seeds, over 4,700 videos per model and setting. Since VBench emphasizes per-dimension appearance and consistency, we additionally report the Fréchet Video Motion Distance (FVMD) (Liu et al., 2024), which compares motion-feature statistics between generated and real videos.

Quantitative Comparisons. As shown in Table 1, STAS consistently improves the aggregate VBench scores across all four backbones. The improvements generalize from standard DiT architectures, including Wan2.1 and Wan2.2, to the MM-DiT-based CogVideoX and the more aggressively compressed LTX-2, demonstrating broad applicability across architectures, model scales, and latent geometries. Notably, on every backbone, the semantic score improves at least as much as the quality score, consistent with our finding that STAS regulates video-token residual writing while affecting text-conditioned computation less directly. STAS also preserves or improves every visual-quality dimension in nearly all settings, without training or additional denoiser evaluations. On LTX-2, the gains concentrate in semantic dimensions while visual quality is preserved, and the improved semantic and aggregate scores show that STAS remains effective under this distinct latent geometry with aggressive compression.

Following Kim & Kim (2026), we further evaluate motion fidelity using combined FVMD (Liu et al., 2024) on the MSR-VTT test set (Xu et al., 2016) in Table 2. FVMD represents tracked keypoint trajectories using velocity and acceleration descriptors. Its combined variant concatenates these first- and second-order motion features before computing the Fréchet distance between generated and real-video distributions, with lower values indicating better motion fidelity. We report the mean and 95% confidence interval over ten paired 200-video subsets. Our method consistently reduces FVMD across all four backbones. These results complement VBench with distribution-level evidence of improved motion fidelity across architectures and latent geometries.

Table 1: **VBench results across different video backbones.**

Backbone	Method	Subject Cons.↑	Background Cons.↑	Aesthetic Quality↑	Imaging Quality↑	Quality Score↑	Semantic Score↑	Total Score↑
Wan2.1-1.3B	Vanilla	94.63	95.81	61.91	68.14	81.81	79.70	81.39
	+Ours	**95.00**	**95.93**	**62.03**	**68.95**	**82.03**	**80.66**	**81.76**
CogVideoX-5B	Vanilla	93.40	95.29	59.98	64.62	79.78	77.59	79.34
	+Ours	**93.80**	**95.47**	**60.31**	**65.12**	**79.95**	**78.24**	**79.61**
Wan2.2-5B	Vanilla	95.13	96.63	61.67	69.02	81.75	81.68	81.74
	+Ours	**95.37**	**96.70**	**61.72**	**69.39**	**81.82**	**82.35**	**81.93**
LTX-2 (base)	Vanilla	92.03	94.73	**60.34**	**66.36**	81.11	76.83	80.26
	+Ours	**92.11**	**94.77**	**60.34**	66.35	**81.13**	**77.32**	**80.37**

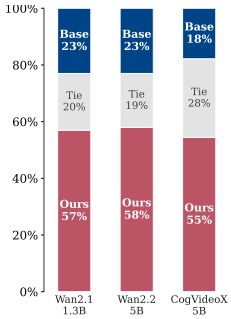


Figure 6: **User study.**

We additionally evaluate STAS on subsets of T2V-CompBench (Sun et al., 2025), which assess object–attribute consistency under static and temporally changing conditions. Due to limited computational resources, we conduct this evaluation on Wan2.1-1.3B. As shown in Table 3, STAS improves both metrics while leaving inference time virtually unchanged, further supporting its effectiveness.

Table 2: **FVMD metric comparisons.**

Backbone	Method	FVMD↓ [95% CI]
Wan2.1-1.3B	Vanilla	11111.8 [10578.6, 11645.0]
	+Ours	**10775.8 [10293.5, 11258.1]**
Wan2.2-5B	Vanilla	11524.0 [11083.8, 11964.2]
	+Ours	**11366.6 [10865.9, 11867.3]**
CogVideoX-5B	Vanilla	14647.2 [14027.3, 15267.1]
	+Ours	**13291.3 [12873.9, 13708.7]**
LTX-2 (base)	Vanilla	12586.0 [11906.9, 13265.1]
	+Ours	**12481.1 [11840.0, 13122.2]**

Table 3: **Results on T2V-CompBench.**

Backbone	Method	Consistent Attr.↑	Dynamic Attr.↑	Time (s)↓
Wan2.1-1.3B	Vanilla	76.27	2.72	234.47
	+Ours	**77.36**	**3.40**	234.53

Computational Cost. We report inference time on an NVIDIA RTX A6000 GPU, averaged over 5 runs after 100 warm-up denoising steps. As shown in Table 3, our method increases runtime from 234.47 s to 234.53 s, indicating negligible inference-time overhead. This is expected since STAS introduces no additional forward computation and only applies an in-place update to a small set of MA dimensions (only one dimension in practice).

Adding STAS to Existing Training-Free Methods. STAS is orthogonal and complementary to existing training-free methods for improving video generation and can be added on top of them to boost performance. We demonstrate this by integrating STAS with FlowMo (Shaulov et al., 2025) and CFG-Zero* (Fan et al., 2025) on Wan2.1-1.3B, and report the quality dimensions in Table 4. Adding STAS improves most quality dimensions and increases the overall quality score for both CFG-Zero* and FlowMo, with gains in subject/background consistency, temporal stability, and imaging quality (e.g., +0.99 and +0.78, respectively). While most dimensions improve, we observe a mild drop in dynamic degree. This is expected because dynamic degree primarily measures motion magnitude rather than temporal coherence, and incoherent motion such as flickering can artificially inflate this score. The net effect is a consistent improvement in generation quality, while the slight reduction in dynamic degree may be a natural byproduct of improved coherence.

Qualitative Comparisons. We present qualitative comparisons across four backbones in Figure 7. STAS improves temporal coherence while preserving visual quality: it prevents abrupt teddy-bear appearance on Wan2.1, maintains panda identity on CogVideoX, preserves car geometry on Wan2.2, and suppresses cat duplication on LTX-2. We further conduct a blinded user study with nine participants and 20 randomized video pairs per model. STAS is preferred in 57%, 58%, and 55% of comparisons on Wan2.1, Wan2.2, and CogVideoX, respectively, consistently exceeding the corresponding baselines (Figure 6). Additional results are provided in the Appendix H.

7 Ablation Study

We mainly ablate four design choices of STAS on Wan2.1-1.3B model: *what* to steer (MA dimension $\mathcal{M}$), *where* to steer (target token set $\mathcal{S}$), *when* to steer (timestep threshold K), and *how* to steer (the calibration rule g). More ablations are provided in Appendix H.2.

Table 4: **Compatibility with CFG-Zero* and FlowMo.**

Method	Subject Cons.↑	Background Cons.↑	Temporal Flicker↑	Motion Smooth.↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑	Quality Score↑
CFG-Zero*	93.43	94.88	97.23	98.43	**58.41**	62.17	67.03	81.81
+ STAS	**93.93**	**95.09**	**97.42**	**98.54**	57.69	**62.35**	**68.02**	**82.21**
FlowMo	93.96	96.40	97.38	98.50	**56.57**	65.51	67.90	82.83
+ STAS	**94.12**	**96.42**	**97.51**	**98.55**	55.78	**65.53**	**68.68**	**83.00**



Figure 7: **Qualitative comparison with and without STAS on different backbones.**

Effect of dimension choice $\mathcal{M}$. We first ablate the choice of the steered activation dimension. Table 5 shows that steering non-MA (dim 42, randomly selected, $\sim 1\times$ peak-to-mean ratio) or weak-MA[1] (dim 71, $\sim 6\times$) dimensions yields negligible or even negative changes, in contrast to the MA dimension (dim 1188, $59\times$). This confirms that meaningful activation spikes are essential for effective steering. In contrast, all variants operating on the MA dimension achieve notable improvements.

[1]We term dimensions with moderate peak-to-mean ratios (5–$20\times$) "weak-MA".

Table 5: **Ablation study of STAS on Wan2.1-1.3B.**

Variant	Dim	Target Tokens	K	Rule	Sub. Consist.	BG. Consist.	Aesthetic Quality	Imaging Quality	Quality Score
Vanilla	–	–	–	–	94.36	95.68	63.28	67.21	82.01
What (Dimension)									
Non-MA	42	first+boundary	20	max	94.33	95.70	63.24	67.23	81.97
Weak-MA	71	first+boundary	20	max	94.24	95.61	63.15	67.22	81.99
Where (Token Position)									
First	1188	first only	20	max	94.25	95.52	63.24	67.58	82.04
Boundary	1188	boundary only	20	max	**94.79**	**95.84**	<u>63.45</u>	67.96	82.09
Interior	1188	first+interior	20	max	94.72	95.73	63.26	68.08	82.09
When (Timestep)									
$K = 5$	1188	first+boundary	5	max	94.68	<u>95.81</u>	63.39	67.73	<u>82.13</u>
$K = 50$	1188	first+boundary	50	max	94.62	95.73	63.31	**68.13**	<u>82.13</u>
How (Amplification Rule)									
Scaling	1188	first+boundary	20	scaling	94.51	95.71	**63.47**	67.56	82.06
Ours(STAS)	1188	first+boundary	20	max	<u>94.74</u>	**95.84**	**63.47**	<u>68.05</u>	**82.24**

Effect of steering boundary tokens $\mathcal{S}$. Table 5 compares different steering token positions. Boundary steering outperforms first-frame steering, while replacing boundary tokens with interior tokens lowers overall quality. Combining first-frame and boundary tokens performs best.

Effect of timestep threshold K. We further ablate the timestep threshold K for applying STAS. As shown in Table 5, all tested settings improve over the vanilla baseline, while K=20 yields the best overall performance. In comparison, a smaller K (e.g., K=5) provides weaker gains, and a larger K (e.g., K=50) becomes less balanced across metrics. This matches our observation in Figure 3 that MA magnitudes decay over denoising, suggesting that steering is most beneficial during the early high-noise steps when global structure is formed.

Effect of amplification rule g. We compare max-referenced calibration with fixed scaling. Scaling preserves the heterogeneous magnitudes of structural MAs, leaving weak activations relatively small. By instead calibrating selected tokens to the current peak activation magnitude while preserving their signs, our max-based rule achieves the best overall quality in Table 5.

8 Limitations

STAS strengthens the model's native suppression of residual writing at selected structural positions, and attention propagates this to other tokens, which may slightly attenuate subtle motion. Samples near the decision boundary can then fall below threshold-based metrics such as Dynamic Degree and be counted as static; the same trade-off is reported for other training-free methods (Shaulov et al., 2025; Jang et al., 2026; Tursynbek et al., 2026) and even for post-training ones (Bu et al., 2026). STAS also depends on the base model: its configuration may require minor adaptation across models, and it cannot fully recover severely degraded generations. We discuss more in Appendix I.

9 Conclusion

In this paper, we investigate Massive Activations in video DiTs and uncover an emergent structural bias aligned with video-VAE encoding asymmetries: MA magnitudes peak at first-frame tokens and recur at latent-frame boundaries, where the encoded content load is reduced. Bidirectional interventions further reveal that MAs act as implicit token-level rescalers, with larger magnitudes suppressing self-attention and feed-forward residual writes. These findings suggest that video DiTs have learned to regulate residual computation according to the structural reliability of their latent tokens. Building on this native behavior, we propose STAS, a training-free method that explicitly strengthens it by calibrating MAs at the identified structural positions during early denoising. Across diverse backbones, STAS consistently improves video quality and temporal consistency without additional forward passes, while remaining compatible with existing training-free methods. Our results establish structured MAs as both an interpretable internal mechanism and a practical signal for improving video generation.

AI USE STATEMENT

In this work, we used generative AI tools for writing aid for grammar, wording, and formatting suggestions. We have not used generative AI tools for autonomous generation of experimental evidence or figure drawing, and tasks involving AI-generated or simulated study participants are not applicable to this work. Additionally, we used generative AI tools for drafting and debugging analysis and visualization code. We have reviewed all AI-assisted work. AI-assisted code was manually inspected, executed, and verified against the underlying data by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The full algorithm, backbone configurations, hyperparameters, activation profiling protocol, and experimental settings are provided in the Appendix. All experiments use publicly available models and benchmarks. And our code will be released.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*, 2023.

Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Bytheway: Boost your text-to-video generation model to higher quality in a training-free way. In *CVPR*, 2025.

Jiazi Bu, Pengyang Ling, Yujie Zhou, Yibin Wang, Yuhang Zang, Xuanlang Dai, Shengyuan Ding, Tianyi Wei, Xiaohang Zhan, Jiaqi Wang, et al. HPSD: Hybrid-policy self-distillation for text-image-to-video diffusion models. *arXiv preprint*, 2026.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *ICML*, 2025.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *ICLR*, 2026.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. VPO: Aligning text-to-video generation models with prompt optimization. In *ICCV*, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. CFG-Zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint*, 2025.

Gongfan Fang, Kunjun Li, Xinyin Ma, and Xinchao Wang. TinyFusion: Diffusion transformers learned shallow. In *CVPR*, 2025.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *NeurIPS*, 2025.

Chaofan Gan, Zicheng Zhao, Yuanpeng Tu, Xi Chen, Ziran Qin, Tieyuan Chen, Mehrtash Harandi, and Weiyao Lin. Massive activations are the key to local detail synthesis in diffusion transformers. In *ICLR*, 2026.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Liefeng Bo, and Di Huang. I4VGen: Image as free stepping stone for text-to-video generation. *arXiv preprint*, 2024a.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024b.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint*, 2022.

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 *CVPR*, 2024.

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *CVPR*, 2025.

Diljeet Jagpal, Xi Chen, and Vinay P Namboodiri. Eidt-v: Exploiting intersections in diffusion trajectories for model-agnostic, zero-shot, training-free text-to-video generation. In *CVPR*, 2025.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. In *ICML*, 2026.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. In *ICML*, 2025a.

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. Video-lavit: unified video-language pre-training with decoupled visual-motional tokenization. In *ICML*, 2024.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025b.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, 2024.

Kwanyoung Kim and Sanghyun Kim. Model already knows the best noise: Bayesian active noise selection via attention in video diffusion model. In *ICLR*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint*, 2024.

Jialu Li, Shoubin Yu, Han Lin, Jaemin Cho, Jaehong Yoon, and Mohit Bansal. Training-free guidance in text-to-video generation via multimodal planning and structured noise initialization. *arXiv preprint*, 2025.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. In *ICML Workshop*, 2024.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *NeurIPS*, 2025.

Wenxuan Liu and Sai Qian Zhang. HQ-DiT: Efficient diffusion transformer with fp4 hybrid quantization. *arXiv preprint*, 2024.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint*, 2025.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint*, 2025.

Louis Owen, Nilabhra Roy Chowdhury, Abhay Kumar, and Fabian Güra. A refined analysis of massive activations in llms. *arXiv preprint*, 2025.

Zihan Qiu, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, et al. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training. *arXiv preprint*, 2026.

Yihua Shao, Deyang Lin, Minxi Yan, Siyu Chen, Fanhu Zeng, Minwen Liao, Ao Ma, Ziyang Yan, Haozhe Wang, Yan Wang, Zhi Chen, Xiaofeng Cao, Haotong Qin, Hao Tang, and Jingcai Guo. TR-DQ: Time-rotation diffusion quantization. In *AAAI*, 2026.

Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. Flowmo: Variance-based flow guidance for coherent motion in video generation. In *NeurIPS*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-Compbench: A comprehensive benchmark for compositional text-to-video generation. In *CVPR*, 2025.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *COLM*, 2024.

Evelyn Turri, Davide Bucciarelli, Sara Sarto, Lorenzo Baraldi, and Marcella Cornia. Few channels draw the whole picture: Revealing massive activations in diffusion transformers. *arXiv preprint*, 2026.

Nurislam Tursynbek, Zhiqiang Lao, Heather Yu, Gedas Bertasius, and Marc Niethammer. TeDiO: Temporal diagonal optimization for training-free coherent video diffusion. In *CVPR Workshop*, 2026.

Wan Team, 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*, 2025.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging initialization gap in video diffusion models. In *ECCV*, 2024.

Tian Xia, Xuweiyi Chen, and Sihan Xu. UniCtrl: Improving the spatiotemporal consistency of text-to-video diffusion models via training-free unified attention control. *TMLR*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *ECCV*, 2024.

Xiaomeng Yang, Zhiyu Tan, and Hao Li. IPO: Iterative preference optimization for text-to-video generation. *arXiv preprint*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025b.

Yunlong Yuan, Yuanfan Guo, Chunwei Wang, Wei Zhang, Hang Xu, and Li Zhang. FreqPrior: Improving video diffusion models with frequency filtering gaussian noise. In *ICLR*, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Rui Wan, Widyadewi Soedarmadji, Enshu Liu, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. ViDiT-Q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *ICLR*, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint*, 2024.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient video generation with latent diffusion models. *arXiv preprint*, 2022.

APPENDIX

STEERING VIDEO DIFFUSION TRANSFORMERS WITH MASSIVE ACTIVATIONS

A MASSIVE ACTIVATION ANALYSIS

A.1 MEASUREMENT PROTOCOL

To identify the MA dimensions of different video DiT models, we compute the average absolute activations over 100 prompts using a streaming pass (running sums), at a fixed set of transformer blocks (and, when needed, a fixed subset of denoising steps). For each block/step, we first take the maximum absolute activation over all token positions for each feature dimension, and then take the mean of these per-dimension maxima as a reference scale. We then report MAs using a peak-to-mean rule: a dimension is marked as MA dimension only if its peak exceeds the mean reference by more than $50\times$. For a new model, this identification requires only a one-pass activation profiling procedure, which means that activations are collected during ordinary forward evaluations and selected directly using the fixed criterion, without backpropagation, parameter search, or task-specific tuning. The profiling cost is therefore small and incurred only once per model.

A.2 STATISTICS ACROSS MODELS

We summarize representative MA statistics at block 15 and timestep 5 across four models in Table 6. For each model, we take the maximum activation magnitude over token positions for every feature dimension, and treat dimensions whose peaks exceed $\mu + 3\sigma$ (computed over dimensions) as potential MA outliers. Among these candidates, dimensions with a peak-to-mean ratio larger than $50\times$ are marked as MA dimensions, while the remaining candidates are reported as weak-MA, indicating that they are elevated but less significant than the MA dims. As shown in Table 6, different models have different MA/weak-MA dimensions.

Table 6: **Model-specific massive activation dimensions across video DiTs.**

Model	Block	k	Type	Dim	Max Value	Peak-to-Mean	Peak-to-Median
Wan2.1–1.3B	15	5	MA	**1188**	**45.4400**	**58.3**	**59.6**
Wan2.1–1.3B	15	5	weak-MA	71	4.8645	6.2	7.5
CogVideoX–5B	15	5	MA	**1982**	**412.5200**	**100.1**	**127.3**
Wan2.2–5B	15	5	MA	**1938**	**63.3419**	**83.8**	**97.1**
Wan2.2–5B	15	5	weak-MA	1389	5.0704	6.7	7.8
Wan2.2–5B	15	5	weak-MA	2357	4.4530	5.9	6.8
LTX2(base)	15	5	MA	**3999**	**321.0118**	**110.2**	**201.2**
LTX2(base)	15	5	weak-MA	1076	50.6361	17.4	31.7

B STAS IMPLEMENTATION AND REPRODUCIBILITY DETAILS

B.1 ALGORITHM

Algorithm 1 summarizes the complete inference procedure of STAS. Working with CFG, STAS calibrates the MA dimensions at selected structural tokens, DiT blocks, and early denoising steps using the max-referenced target. The operation is implemented as an in-place masked assignment over a small subset of hidden-state entries and requires neither additional denoiser evaluations nor auxiliary branches.

B.2 MODEL-SPECIFIC CONFIGURATIONS

We summarize the backbone configurations and the corresponding STAS settings in Table 7. For each model, we report the number of transformer blocks, hidden size, the identified MA dimensions, and the layer index where STAS is applied. These configurations are used consistently across all experiments unless otherwise specified.

Algorithm 1 Structured Activation Steering (STAS)

Require: Video DiT D_θ, VAE decoder $\mathcal{D}$
Require: Sampling schedule $\{t_k\}_{k=0}^{N-1}$ and CFG scale λ
Require: MA dimensions $\mathcal{M}$ and intervention blocks $\mathcal{L}$
Require: Steering strength α, boundary ratio p, and step window K
Sample $z_0 \sim \mathcal{N}(0, I)$
$\mathcal{S} \leftarrow \mathcal{F}_0 \cup \mathcal{B}(p)$
function STEER(H^m, k)
  if $k < K$ **and** $m \in \mathcal{L}$ **then**
    for each $d \in \mathcal{M}$ **do**
      $r^m_{k,d} \leftarrow \max_{j \in \mathcal{V}} \left| H^m_{k,j,d} \right|$
      for each $i \in \mathcal{S}$ **do**
        $H^m_{k,i,d} \leftarrow \alpha \cdot r^m_{k,d} \cdot \operatorname{sign}\Big(H^m_{k,i,d} \Big)$
      end for
    end for
  end if
  return H^m
end function
for $k = 0, 1, \ldots, N-1$ **do**
  $v_k^{(+)} \leftarrow D_\theta\big(z_k, t_k, c; \text{STEER}(\cdot, k)\big)$ ▷ Conditional branch
  $v_k^{(-)} \leftarrow D_\theta\big(z_k, t_k, c_{\text{neg}}; \text{STEER}(\cdot, k)\big)$ ▷ Unconditional branch
  $v_k^{\text{cfg}} \leftarrow v_k^{(-)} + \lambda\Big(v_k^{(+)} - v_k^{(-)}\Big)$
  $z_{k+1} \leftarrow \text{UPDATE}(z_k, v_k^{\text{cfg}}, t_k)$
end for
return $\mathcal{D}(z_N)$

Table 7: **Model configurations used for implementing STAS across backbones.**

Model	Blocks	Hidden size	MA dim	Applied layer
Wan2.1-1.3B	30	1536	1188	9
Wan2.2-5B	30	3072	1938	9
CogVideoX-5B	42	3072	1982	8
LTX2 (base)	48	4096	3999	7

Table 8: **Hyperparameter setup across models.**

Model	λ (CFG scale)	α (STAS scale)	p (boundary %)	K (timestep cut-off)
Wan2.1-1.3B	5.0	2.5	8	20
Wan2.2-5B	5.0	2.0	12	20
CogVideoX-5B	6.0	1.2	8	20
LTX2 (base)	3.0	1.2	8	30

Since STAS is applied on top of CFG (Ho & Salimans, 2021), we use the default guidance scale λ for each model. The STAS scaling factor α, latent-boundary percentage p, and denoising timestep threshold K are summarized in Table 8.

B.3 Experimental Protocol for Main Figures

In this subsection, we detail the experimental settings and implementation specifics used to generate Figures 1, 2 and 4.

Figure 1. For all the 3D bar charts in Figure 1, the z-axis values (activation magnitudes of hidden states) are collected from block 15 at an early denoising timestep ($k = 5$, counting from the start of the denoising process). The corresponding generated video resolutions are $81 \times 480 \times 832$, $81 \times 704 \times 1280$, $49 \times 480 \times 720$, and $81 \times 512 \times 768$, respectively.

Figure 2. All VAE experiments in Figure 2 use the Wan2.1 video VAE on videos of resolution $T \times H_x \times W_x = 81 \times 480 \times 832$, producing 21 latent frames. We disable tiled encoding to avoid additional tile-boundary effects.

For the temporal analysis in Figure 2a (the purple curve), we quantify the content carried by each latent frame through a decoder-ablation probe. Given video $V^{(q)}$ and its VAE latent $z^{(q)} = \mathcal{E}(V^{(q)})$, we ablate latent frame f by replacing its spatially varying features with their per-channel spatial means:

$$\widetilde{z}^{(q,f)}_{c,\ell,h,w} = \begin{cases} \dfrac{1}{H_z W_z} \displaystyle\sum_{h',w'} z^{(q)}_{c,f,h',w'}, & \ell = f, \\ z^{(q)}_{c,\ell,h,w}, & \ell \neq f. \end{cases} \tag{5}$$

Here, $f \in [0, 20]$ denotes the ablated latent frame, whereas ℓ indexes an arbitrary latent frame. This operation preserves the channel-wise mean while removing the frame's spatial structure. We decode the original and ablated latents and measure the resulting change over the physical frames associated with f:

$$\mathcal{L}^{(q)}_f = \frac{1}{3H_x W_x} \sum_{\tau \in \mathcal{G}_f} \sum_{c,h,w} \left| \mathcal{D}(z^{(q)})_{c,\tau,h,w} - \mathcal{D}(\widetilde{z}^{(q,f)})_{c,\tau,h,w} \right|, \tag{6}$$

where $\mathcal{G}_0$ contains the first decoded physical frame, while $\mathcal{G}_f$ contains the four physical frames represented by each later latent frame. The purple curve plots, at each latent-frame index f, the median across all tested videos of $100\,\mathcal{L}^{(q)}_f / \operatorname{median}_{r=1,\dots,20} \mathcal{L}^{(q)}_r$. Hence, the value of approximately 26% at $f = 0$ means that the first latent frame carries about one quarter of the decoded content of a typical subsequent latent frame under the same token budget, in line with the VAE's $4\times$ temporal compression.

For comparison in Figure 2a (the orange curve), we collect the absolute activation of MA dimension 1188 from block 15 at sampling step $k = 5$ over 100 prompt-conditioned generations. For each generation and latent-frame index f, we first average the absolute MA magnitude over the frame's interior tokens, excluding a two-token margin from every spatial boundary. The orange curve plots the mean across generations of this frame-wise magnitude divided by its median over all latent frames after the first and multiplied by 100. Thus, a value of 119% at $f = 0$ means that the first latent frame has an average MA magnitude approximately 19% larger than a typical subsequent latent frame.

For the spatial analysis in Figure 2b, we construct eight control videos whose frames are spatially uniform. Their intensities span $[-0.9, 0.9]$, with a small temporally varying but spatially constant perturbation. Because the inputs contain no spatial structure, systematic spatial variation in their latents reflects the VAE encoder rather than image content. We average the absolute latent values over Q videos, C_z channels, and all latent frames after the first, producing a 60×104 spatial map:

$$M^{\mathrm{VAE}}_{h,w} = \frac{1}{Q\,C_z\,(F_z - 1)} \sum_{q=1}^{Q} \sum_{c=1}^{C_z} \sum_{f=1}^{F_z - 1} \left| z^{(q)}_{c,f,h,w} \right|. \tag{7}$$

Excluding the first latent frame separates the spatial-padding effect from the distinct temporal-padding effect.

The corresponding DiT map is computed from MA dimension 1188 at block 15 and sampling step $k = 5$ over $Q_{\mathrm{MA}} = 100$ generations. We reshape the flattened video tokens into a $21 \times 30 \times 52$ grid and average their absolute MA magnitudes over generations and latent frames $f \geq 2$:

$$M^{\mathrm{MA}}_{h,w} = \frac{1}{Q_{\mathrm{MA}}(F_z - 2)} \sum_{q=1}^{Q_{\mathrm{MA}}} \sum_{f=2}^{F_z - 1} \left| H^{15,(q)}_{k,(f,h,w),1188} \right|. \tag{8}$$

The earliest two latent frames are excluded here to suppress the dominant temporal transient and isolate the recurring spatial pattern. After 2×2 spatial patchification, the 60×104 VAE latent grid becomes a 30×52 DiT token grid for each latent frame.

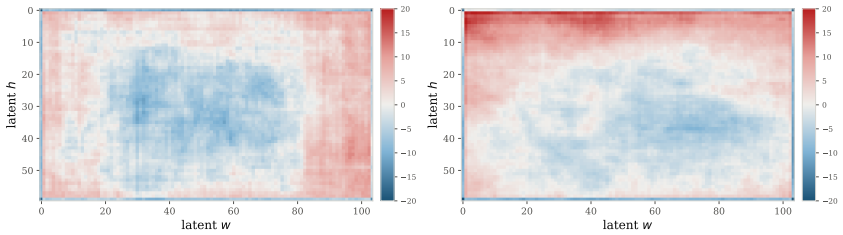
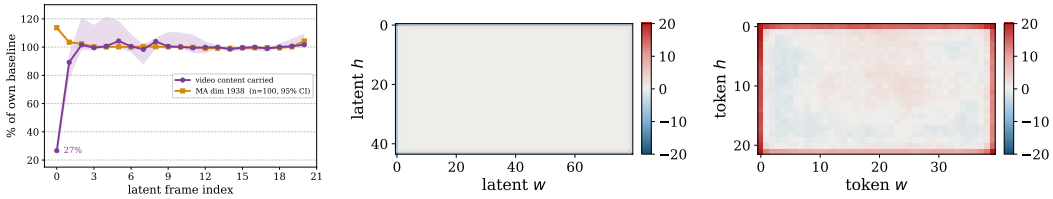
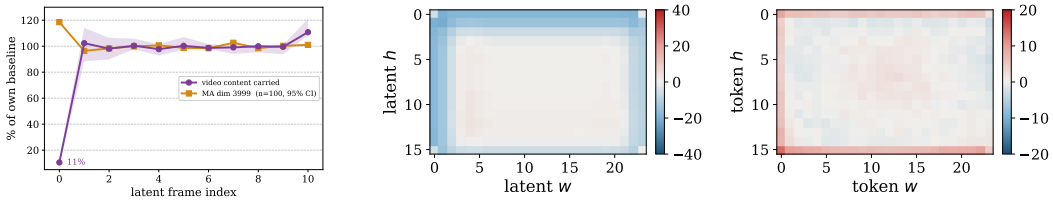


Figure 8: **VAE boundary responses on nonuniform inputs.** Natural videos are shown on the left and model generated videos on the right.



(a) Per-frame content load vs. MA (b) Latent boundaries of the VAE (left) and the DiT MA map (right)

Figure 9: **The MA hierarchy aligns with video-VAE encoding asymmetries on Wan2.2 VAE.**



(a) Per-frame content load vs. MA (b) Latent boundaries of the VAE (left) and the DiT MA map (right)

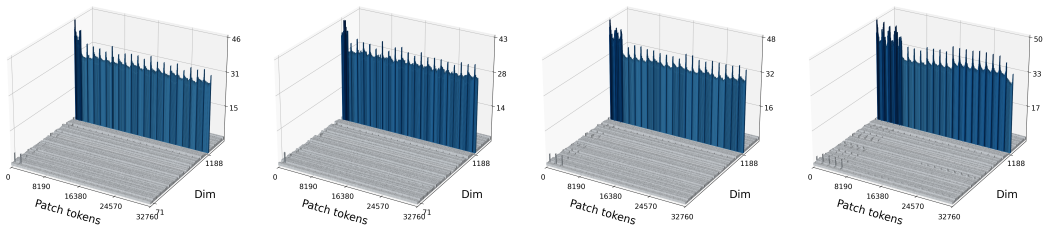
Figure 10: **The MA hierarchy aligns with video-VAE encoding asymmetries on LTX-2 VAE.**

For visualization, each map is independently normalized by the median of its interior positions, excluding the outermost spatial boundaries, and shown as the percentage deviation from this reference. Accordingly, 0% denotes the interior median, while positive and negative values indicate larger and smaller magnitudes, respectively.

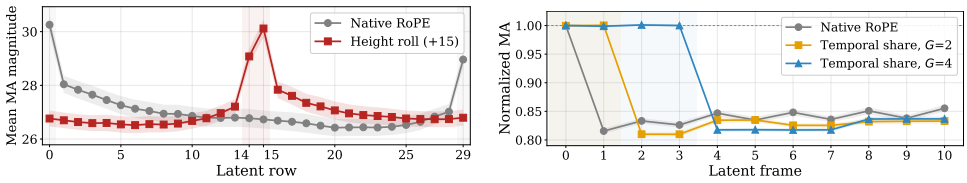
Figures 4 & 5. For the Figure 4, over 100 paired Wan2.1-1.3B generations, we steer MA dimension 1188 at the selected structural tokens in block 9 and measure the response in block 10 during the first ten denoising steps. We report post-normalization content and self-attention/FFN residual-write magnitudes, normalized by their paired native values. Dashed curves show the corresponding responses at non-target interior tokens. We use the same videos for the experiments in Figure 5.

C Additional Analysis of the VAE–MA Connection

VAE Boundary Responses on Natural and Generated Videos. The spatial boundary response is not restricted to the uniform controls used in our main analysis. Figure 8 shows the average VAE latent magnitude for natural videos and generated videos (averaged over 100 videos each). Both exhibit content-dependent, low-frequency spatial variation, yet retain a sharp one-cell attenuation at the outermost latent boundary. Thus, the boundary signature introduced by the VAE remains visible under realistic content and in the model's own outputs. Its alignment with elevated MAs at the same positions, further supporting the VAE–MA connection.



(a) Native RoPE (b) Height roll (+15) (c) Temporal share, G=2 (d) Temporal share, G=4



(e) Spatial MA profiles (f) Temporal MA profiles

Figure 11: **MA peaks follow the RoPE coordinates.** (a) Native RoPE. (b) Rolling the height RoPE indices by $+15$ moves the boundary coordinates to the center rows, and the MA peaks move with them. (c–d) Sharing each temporal RoPE coordinate across groups of G=2 or G=4 consecutive latent frames makes the first G frames share the first-frame coordinate and exhibit a common elevated MA plateau. (e) The aggregated spatial profiles show the MA peak moving from the boundary rows to the center rows under the height roll. (f) The normalized temporal profiles show elevated plateaus spanning one, two, and four latent frames under native RoPE, G=2, and G=4, respectively.

Generalization across different VAEs. We further extend the VAE–MA correspondence experiments on Wan2.2 and LTX2 video VAEs, which use different spatial and temporal compresion ratios shown in Figures 9 and 10. Temporally, the first latent frame carries substantially less video content than later frames, while exhibiting a consistently higher MA magnitude. Spatially, both models exhibit attenuated VAE latent magnitude at boundary positions alongside elevated MAs at the corresponding tokens. These results show that the inverse relationship between VAE content load and MA magnitude generalizes across different VAEs.

D Analysis of the Role of RoPE in Localizing MA Patterns

Asymmetric spatial padding and causal temporal processing in the video VAE produce a systematic information asymmetry between boundary and interior latent tokens. Because the same VAE transformation is applied throughout training, this pattern is expected to persist in the latent training distribution. We hypothesize a three-stage account: the VAE creates the asymmetry, the DiT learns its recurring structure, and RoPE supplies the coordinates that localize the learned response in the flattened token sequence. We test the localization stage by modifying only the RoPE lookup and no latent token is explicitly moved or overwritten.

Specifically, we use Wan2.1-T2V-1.3B and follow the MA identification protocol described in Appendix A.1. All conditions use the same 100 prompts, random seeds, model weights, and sampling settings. We record activations at block 15 and timestep 5. An 81-frame video corresponds to 21 latent frames of 30×52 tokens, leading to 1,560 consecutive positions in the flattened token sequence.

Spatial RoPE roll. We first shift every height coordinate by half the grid in RoPE:

$$\pi_h(r) = (r + 15) \bmod 30. \tag{9}$$

The boundary rows 0 and 29 thus receive central indices and the central rows 14 and 15 receive boundary indices, with the native order $0, 1, \ldots, 29$ becoming $15, 16, \ldots, 29, 0, 1, \ldots, 14$. The shift is a bijection on the index set, and the token order and all other coordinates are unchanged.

With 52 tokens per row, shifting by 15 rows predicts a displacement of $15 \times 52 = 780$ tokens, exactly half the 1,560-token frame period. This is what we observe in Fig. 11b: in every latent frame, including the first, the peaks move off the physical boundary rows and toward the frame center. The aggregated row profiles in Fig. 11e further make this relocation explicit.

Temporal RoPE grouping. Native indexing assigns latent frame i the temporal coordinate $\pi_t(i) = i$. We next use

$$\pi_t^{(G)}(i) = \left\lfloor \frac{i}{G} \right\rfloor, \qquad G \in \{2, 4\}, \tag{10}$$

so that G consecutive latent frames share one temporal coordinate, while the height and width coordinates remain unchanged. For $G = 2$ the native sequence $0, 1, 2, \ldots, 20$ becomes $0, 0, 1, 1, \ldots, 9, 9, 10$, so the tokens of the first two frames all carry the temporal coordinate of the first frame. The first latent frame MA elevation should then cover both.

Figures 11c and 11d show this behavior, in which the elevated region expands to two latent frames for $G = 2$ and four latent frames for $G = 4$. Consistently, the averaged profiles in Fig. 11f contain elevated plateaus of 1, 2, and 4 latent frames under native indexing, $G = 2$, and $G = 4$, respectively. Although repeated temporal indices constitute an out-of-distribution coordinate assignment, the systematic widening follows the prediction of a coordinate-driven mechanism.

Interpretation. The two interventions probe complementary axes of the factorized RoPE. Spatial half-period shift changes its phase within each frame, whereas temporal grouping changes the extent of the MA pattern across latent frames. Together, these results establish the assigned RoPE coordinates as the localization mechanism for MA. RoPE transports the VAE-induced asymmetry to the corresponding token positions, and the VAE generates the underlying asymmetry.

E Additional Analysis of MA-Induced Residual Rescaling

We give a normalization-based analysis of why MA magnitude rescales residual writing. Both LayerNorm and RMSNorm fix the overall scale of each token, so a single unusually large hidden dimension inflates the normalization denominator and shrinks the relative contribution of all remaining dimensions. We derive this effect for the normalization layers used by the studied backbones, omitting the stabilization constant ϵ, which is negligible at the observed magnitudes.

Fix a block m, denoising step k, and token i, and abbreviate $h_d = H^m_{k,i,d}$, where C is the number of channels. Let $d^* \in \mathcal{M}$ denote the dominant MA dimension, with signed activation $a = h_{d^*}$ and magnitude $A = |a|$.

LayerNorm. Wan2.1 and CogVideoX use LayerNorm before self-attention and feed-forward branch, which subtracts the token mean μ from every dimension and divides by the token standard deviation $\sqrt{v}$. To see how the MA affects this normalization, let $\bar{c}$ be the mean of the remaining dimensions and let $E_{\mathrm{LN}} = \sum_{d' \neq d^*} (h_{d'} - \bar{c})^2$ denote their centered energy. Separating the MA dimension gives

$$v = \frac{1}{C} \sum_{d'=1}^{C} (h_{d'} - \mu)^2 = \underbrace{\frac{E_{\mathrm{LN}}}{C}}_{\text{remaining dimensions}} + \underbrace{\frac{C-1}{C^2}(a - \bar{c})^2}_{\text{MA contribution}}. \tag{11}$$

The first term is the variance already present among the remaining dimensions, while the second is the additional variance introduced by the MA. As $|a - \bar{c}|$ grows, every dimension is divided by a larger standard deviation. Mean subtraction also shifts all non-MA dimensions by the same amount, but this shift cancels in their pairwise differences. Their contrast is therefore attenuated by the enlarged denominator. Since $A \gg |\bar{c}|$ and C is large, this attenuation is approximately $(1 + A^2/E_{\mathrm{LN}})^{-1/2}$, so that erasing the MA releases the remaining dimensions, whereas amplifying it suppresses them.

RMSNorm. LTX-2 divides each coordinate by the root-mean-square magnitude of the token. Writing $E_{\mathrm{RMS}} = \sum_{d' \neq d^*} h_{d'}^2$ for the energy of the remaining dimensions, every $d \neq d^*$ is mapped

to

$$\mathrm{RMSNorm}(h)_d = \frac{h_d}{\sqrt{(E_{\mathrm{RMS}} + A^2)/C}} = \rho(A)\frac{h_d}{\sqrt{E_{\mathrm{RMS}}/C}}, \qquad \rho(A) = \frac{1}{\sqrt{1 + A^2/E_{\mathrm{RMS}}}}. \tag{12}$$

RMSNorm therefore yields the same attenuation form directly. The factor $\rho(A)$ equals one when zeroing out MA and decreases monotonically as A grows.

LayerNorm and RMSNorm thus share the same mechanism: a dominant MA increases the token's normalization denominator, leaving less normalized scale for its remaining dimensions. LayerNorm measures their centered energy, whereas RMSNorm uses their uncentered energy. When several MA dimensions are active, their joint contribution enters the denominator and produces the same effect.

Implications for residual writing. As normalization is applied to each token independently and a token is divided by its own RMS magnitude, a larger MA inflates that token's denominator and leaves less normalized scale for its remaining dimensions. The attenuation therefore depends on the MA magnitude relative to the energy E_{RMS} of those dimensions. Scaling the MA dimension alone does not change E_{RMS}, so ρ decreases strictly monotonically in A at that token. The normalized token then feeds both the attention-value and feed-forward projections. Amplifying a token's MA therefore weakens its residual writes, whereas erasing it strengthens them.

Interestingly, although the feed-forward branch processes each token independently, self-attention mixes tokens and can propagate the intervention to non-target positions. After projection, queries and keys are normalized again by QK normalization before the attention scores are computed. This equalizes their vector lengths but does not restore their original coordinate values. Consequently, the query–key dot products and attention weights can still change. Figure 4 supports this explanation. At selected tokens, MA erasure strengthens both residual branches, while amplification progressively suppresses them. At non-target tokens, the main response occurs in the self-attention write, with feed-forward writes remaining close to their native values. Affine and AdaLN parameters are shared across video tokens at a given block and denoising step, so they can change absolute magnitudes but do not eliminate this token-dependent attenuation.

F Additional Analysis of STAS Design Choices

Despite structural token positions exhibiting larger MA magnitudes on average, as discussed in Sec. 4, this elevation does not hold for every token: an individual structural token may still have an MA magnitude no greater than that of a typical interior token. To quantify this overlap, we measure the fraction of structural tokens whose native MA falls below the interior-token median, across 100 generated videos. This fraction would be 50% if structural and interior tokens followed the same distribution, whereas 0% would mean that every structural token has a larger MA than the interior median. Figure 12 shows values of 18% for first-frame tokens and 34% for boundary-band tokens, indicating that structural MAs are generally stronger but remain heterogeneous.

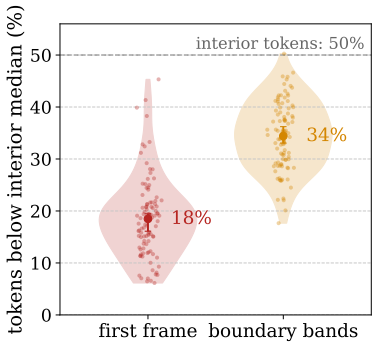


Figure 12: **Structural MAs are not uniformly elevated.**

Recall that our steering target in Eq. 3 determines the magnitude of each selected activation using the maximum response of the corresponding dimension while preserving its original polarity. Since this target is independent of the token's native magnitude, it brings both weak and strong structural responses to a consistent adaptive scale. Simple multiplicative scaling instead preserves their relative differences, leaving weak structural MAs comparatively weak. This explains why max-referenced calibration provides a better performance than scaling as shown in Figure 5 and Table 5.

Table 9: **Unified formulations of output guidance and STAS activation steering.**

Type	Formulation	Applied to	Reference / Target
CFG	$\widehat{D} = D + w(D - D_{\hat{c}})$	Denoiser output	Unconditional $D_{\hat{c}}$
Autoguidance	$\widehat{D} = D + w(D - D_{\theta^*})$	Denoiser output	Under-capacity D_{θ^*}
DG	$\widehat{D} = D + w(D - D_{\hat{z}})$	Denoiser output	MA-disrupted $D_{\hat{z}}$
STAS	$\widehat{H} = H + \chi(g - H)$	Hidden activation	MA-related adaptive target g

G Analysis of STAS Through the Lens of Guidance

We analyze STAS through the lens of guidance by comparing it with CFG (Ho & Salimans, 2021), Autoguidance (Karras et al., 2024), and DG (Gan et al., 2026). Each method can be written as an original quantity plus a correction. Specifically, CFG, Autoguidance, and DG correct the denoiser prediction, while STAS corrects internal hidden activations. To make this comparison explicit, let $D = D_\theta(z_t, t, c)$ denote the original conditional denoiser prediction and define

$$D_{\hat{c}} = D_\theta(z_t, t, \hat{c}), \qquad D_{\theta^*} = D_{\theta^*}(z_t, t, c), \qquad D_{\hat{z}} = D_\theta(\hat{z}_t, t, c).$$

Here, $\hat{c}$ denotes an unconditional or negative condition, θ^* denotes the parameters of an under-capacity model, and $\hat{z}_t$ denotes the MA-disrupted state used to construct the DG reference prediction.

STAS does not combine two denoiser predictions. Instead, it directly replaces selected hidden activations with the calibrated target g. To express this replacement as an additive update, let $\chi^m_{k,i,d}$ indicate whether an activation is selected: it equals 1 when $k < K$, $m \in \mathcal{L}$, $i \in \mathcal{S}$, and $d \in \mathcal{M}$, and equals 0 otherwise. Equation 4 can then be rewritten as

$$\widehat{H}^m_{k,i,d} = H^m_{k,i,d} + \chi^m_{k,i,d} \left(g^m_{k,i,d} - H^m_{k,i,d} \right). \tag{13}$$

For a selected activation, $\chi^m_{k,i,d} = 1$, so the added term changes its value exactly from $H^m_{k,i,d}$ to $g^m_{k,i,d}$. For every other activation, $\chi^m_{k,i,d} = 0$, so no correction is applied.

Table 9 highlights both the shared formulation and the key distinction. CFG, Autoguidance, and DG construct an auxiliary reference prediction and extrapolate away from it in denoiser-output space. CFG obtains its reference by removing the text condition, Autoguidance by replacing the denoiser with an under-capacity model, and DG by disrupting MAs in an auxiliary branch. Their guidance weight w controls the correction magnitude along the difference between the original and reference predictions.

STAS instead operates directly in activation space. Rather than constructing an auxiliary prediction to extrapolate from, it rescales selected MA activations toward the max-referenced target g, whose magnitude is controlled by α in Equation 3. MAs have previously been used as something to disrupt: DG perturbs them to obtain a degraded reference and then corrects the denoiser output. STAS instead treats them as a quantity to calibrate, adjusting them in place so that the corrected activations propagate through the remaining blocks. Consequently, unlike Autoguidance and DG, STAS requires neither an auxiliary model nor an additional degraded forward pass. It applies a single element-wise operation at a few structural token positions, adding no parameters and negligible computational overhead.

H Additional Results

H.1 Additional Quantitative Results

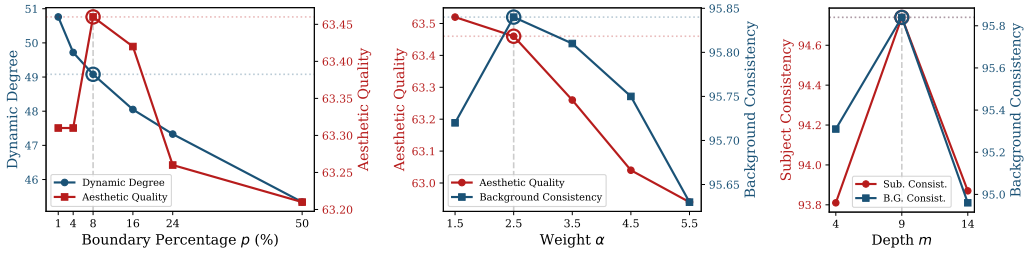
Table 10 presents the complete VBench breakdown over all 16 evaluation dimensions, together with the aggregated quality, semantic, and final scores across the four backbone models, showing consistent improvements across most dimensions.

H.2 Additional Ablations

We here provide additional ablation studies to further analyze more design choices in STAS: the boundary-percentage p, the scaling factor α and the applying layer depth m. Due to limited compu-

Table 10: **Quantitative comparison on VBench.**

Dimension	Wan2.1-1.3B		CogVideoX-5B		Wan2.2-5B		LTX-2 (base)	
	Vanilla	**+Ours**	Vanilla	**+Ours**	Vanilla	**+Ours**	Vanilla	**+Ours**
Subject Consistency	94.63	**95.00**	93.40	**93.80**	95.13	**95.37**	92.03	**92.11**
BG. Consistency	95.81	**95.93**	95.29	**95.47**	96.63	**96.70**	94.73	**94.77**
Temporal Flickering	97.93	**98.09**	95.97	**96.21**	97.81	**97.87**	**97.80**	**97.80**
Motion Smoothness	98.78	**98.85**	98.18	**98.28**	**98.94**	**98.94**	**98.90**	**98.90**
Dynamic Degree	**45.15**	43.64	**48.71**	45.85	**39.32**	38.24	**51.74**	51.63
Aesthetic Quality	61.91	**62.03**	59.98	**60.31**	61.67	**61.72**	**60.34**	**60.34**
Imaging Quality	68.14	**68.95**	64.62	**65.12**	69.02	**69.39**	**66.36**	66.35
Object Class	90.38	**91.64**	86.88	**88.07**	**92.34**	92.23	**88.97**	88.94
Multiple Objects	77.47	**78.61**	67.47	**68.78**	78.60	**82.86**	64.59	**66.14**
Human Action	96.40	**97.20**	98.60	**98.80**	**98.00**	97.80	98.20	**98.40**
Color	**86.93**	86.72	82.04	**84.12**	87.18	**87.54**	75.91	**77.47**
Spatial Relationship	73.63	**78.38**	64.17	**64.73**	86.68	**87.31**	**77.23**	77.07
Scene	54.87	**55.41**	**54.62**	54.41	54.18	**55.47**	54.74	**55.71**
Appearance Style	22.90	**22.93**	24.83	**24.92**	**22.84**	22.77	22.67	**22.75**
Temporal Style	25.97	**26.02**	25.49	**25.59**	**25.81**	25.80	**24.62**	24.59
Overall Consistency	**27.03**	27.02	27.59	**27.65**	**27.50**	27.44	**26.57**	26.54
Quality Score	81.81	**82.03**	79.78	**79.95**	81.75	**81.82**	81.11	**81.13**
Semantic Score	79.70	**80.66**	77.59	**78.24**	81.68	**82.35**	76.82	**77.32**
Final Score	81.39	**81.76**	79.34	**79.61**	81.74	**81.93**	80.26	**80.37**



(a) Effect of boundary percentages. (b) Effect of STAS weights. (c) Effect of depths.

Figure 13: **More ablations.** Impact on boundary percentage p, STAS weight α and applied layer depth m. Moderate values achieve the best trade-off across metrics.

tational resources, all ablations are evaluated on the VBench quality-dimension prompts only (1,255 videos per variant).

Effect of boundary percentage p. We present the quantitative results across different boundary percentage p (head $p\%$ + tail $p\%$) of latent chunks. As shown in Figure 13a, increasing p from very small values initially improves aesthetic quality, reaching a peak at moderate percentages (around 8%), while the dynamic degree gradually decreases as p becomes larger. Note that overly large p, which involves more interior tokens for calibration, would lead to a degradation for both metrics. For instance, when $p = 50$, which all the visual tokens are selected for amplification, the performance is poor. These results suggest that a moderate boundary coverage achieves the best trade-off.

Effect of scaling weight α. We next show the effect of varying scaling weight α, which controls the magnitude of amplification. As illustrated in Figure 13b, both metrics improve when increasing α from small values, peaking at a moderate weight (around $\alpha = 2.5$). However, further increasing α leads to a consistent decline in both metrics.

Effect of applying layer depth m. We further ablate the transformer layer depth m at which STAS is applied. As shown in Figure 13c, steering at an intermediate layer (m=9) achieves the best performance on both metrics. In contrast, applying STAS at a shallower layer (m=4) or a deeper

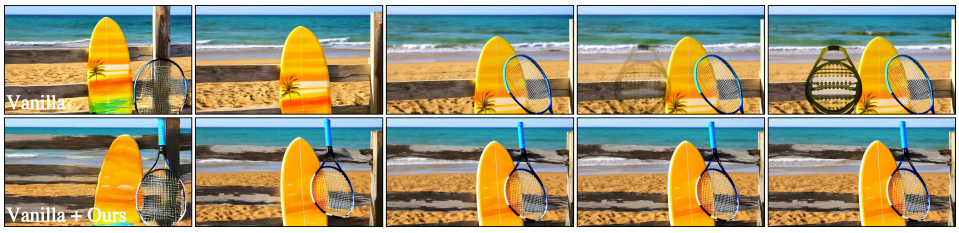


Figure 14: **A failure case of Dynamic Degree evaluation.** Matched frames for "a surfboard and a tennis racket". In the baseline (top), the tennis racket disappears in the early frames and later reappears together with an additional green racket. Dynamic Degree counts these identity changes as motion and predicts `True`. With STAS (bottom), one racket persists across all frames and the prediction changes to `False`.

layer (m=14) leads to weaker results. This suggests that intermediate layers provide a more suitable representation for our method, while steering too early or too late in the network is less effective.

H.3 Additional Qualitative Results

We provide additional qualitative comparisons across different base models in Figures 15, 16, 17, 18, 19, 20, 21 and 22. The example videos are also provided in the supplementary materials.

I Additional Discussion of Limitations

Threshold-based metrics like Dynamic Degree in VBench can misinterpret temporal artifacts as genuine motion. Figure 14 shows matched frames from the baseline and STAS for the prompt "a surfboard and a tennis racket". The baseline exhibits both failure modes that inflate frame-to-frame variation: the tennis racket first disappears over the early frames, then returns later in the sequence together with an additional green racket. Neither change represents meaningful motion and both arise from inconsistent object identity. Dynamic Degree nevertheless counts this variation as dynamics and predicts `True`. With STAS, a single racket persists across the entire sequence, the artifact-driven variation disappears, and the sample falls below the metric's decision threshold, changing the prediction to `False`. This transition therefore reflects the removal of artifact-driven variation rather than of genuine motion.

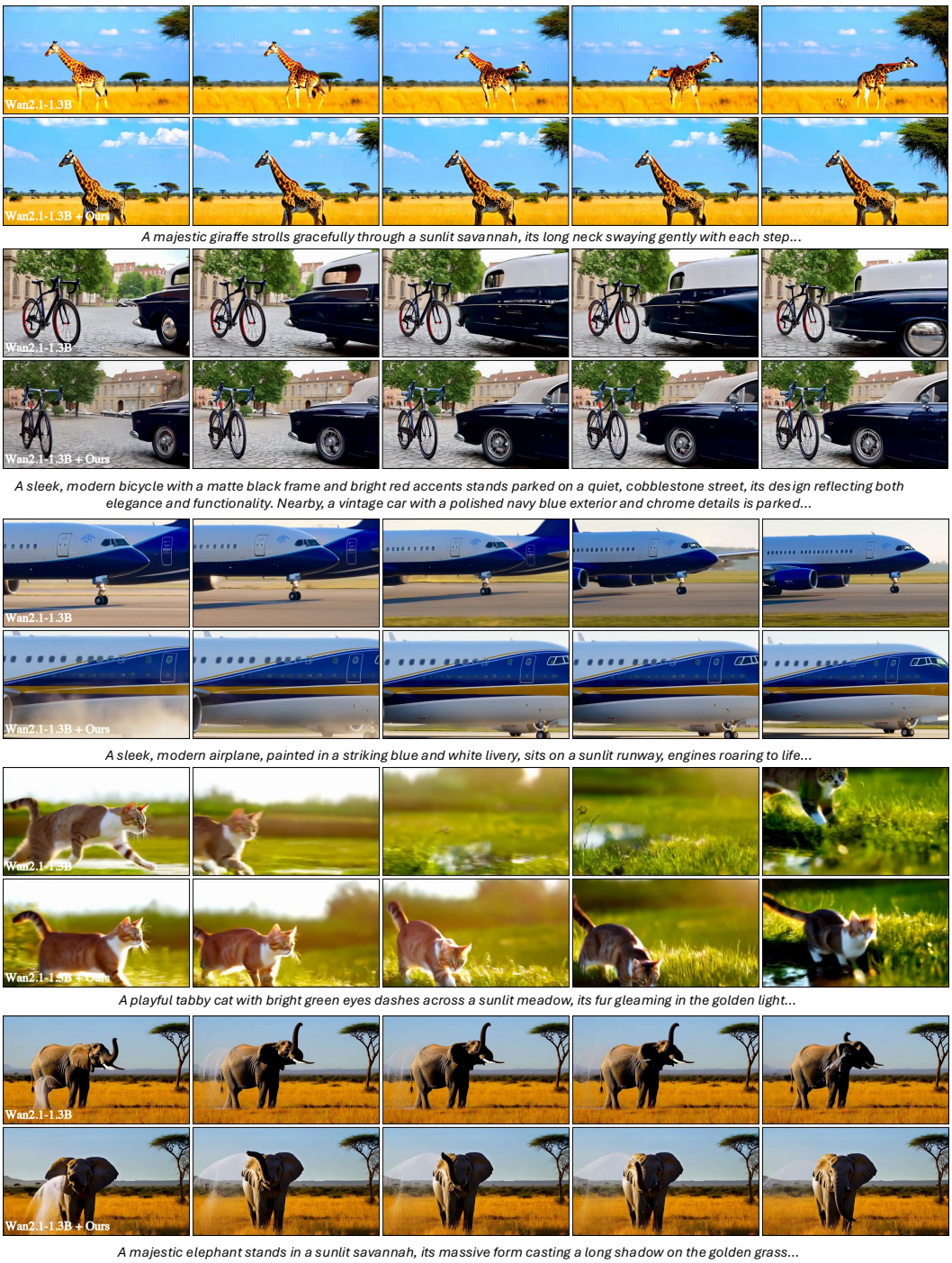
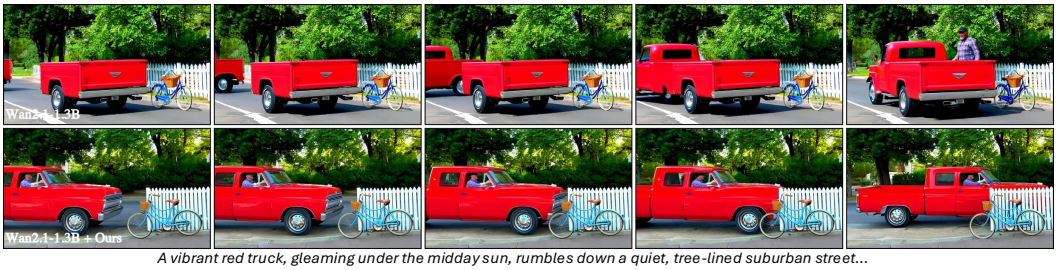


Figure 15: **More qualitative comparisons with and without STAS on Wan2.1-1.3B.**



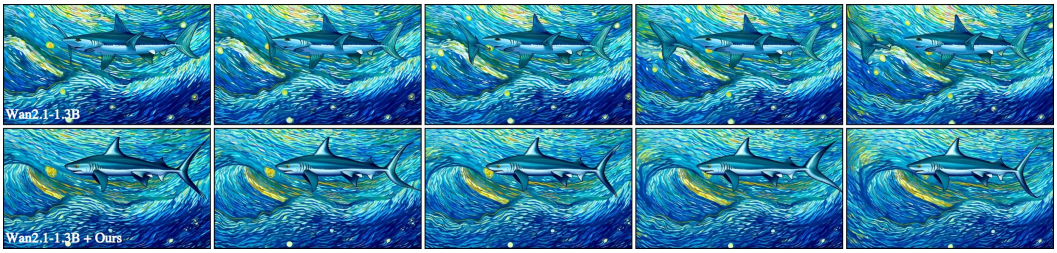
A vibrant red truck, gleaming under the midday sun, rumbles down a quiet, tree-lined suburban street...



A person with short, curly hair and wearing a cozy, oversized sweater stands in a warmly lit room, their eyes closed in a moment of deep connection. They embrace another individual, whose face is partially visible, showing a gentle smile...



A vibrant purple umbrella opens against a backdrop of a bustling city street, its rich hue standing out amidst the gray, rainy day...



A majestic shark glides through the swirling, vibrant waters of the ocean, depicted in the iconic Van Gogh style. The scene is alive with dynamic, swirling brushstrokes of deep blues, teals, and hints of golden yellows, capturing the movement of the water and the shark's sleek form...



A refined couple, dressed in elegant evening attire, navigates a bustling street under a heavy downpour. The man, in a tailored black tuxedo, and the woman, in a flowing crimson gown, both hold delicate paper umbrellas adorned with intricate patterns...

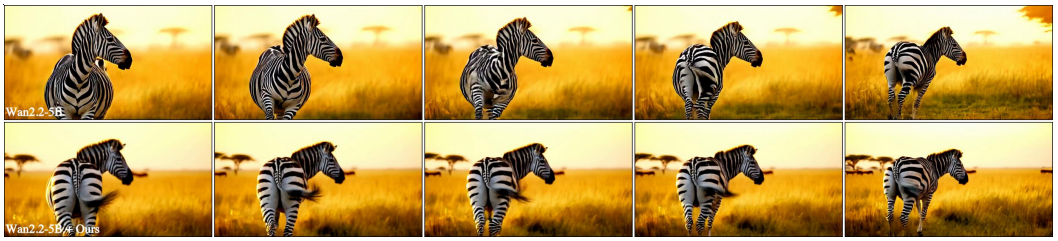
Figure 16: **More qualitative comparisons with and without STAS on Wan2.1-1.3B.**



A lone bicycle, with its sleek frame and black tires, glides effortlessly through a vast, snow-covered field under a pale winter sky. The rider, bundled in a red parka, black gloves, and a woolen hat, pedals steadily, leaving a delicate trail in the pristine snow...



A couple, elegantly dressed in formal evening wear, navigates a bustling city street under a heavy downpour. The man, in a tailored black tuxedo, and the woman, in a flowing emerald gown, hold large, ornate umbrellas that barely shield them from the relentless rain...



A majestic zebra stands in the golden savannah, its black and white stripes contrasting vividly against the tall, sunlit grasses. The camera captures a close-up of its face, highlighting the intricate patterns around its eyes and muzzle...



In a rustic barn bathed in the soft morning light, a person in a plaid shirt, denim overalls, and sturdy boots kneels beside a gentle, brown-and-white cow. The person carefully places a metal pail beneath the cow's udder, their hands moving with practiced ease...



A lone rider, clad in a sleek black leather jacket, matching helmet, and dark jeans, navigates a winding mountain road on a powerful motorcycle. The sun sets behind the peaks, casting a golden glow on the rugged landscape...

Figure 17: **More qualitative comparisons with and without STAS on Wan2.2-5B.**



A sleek, vibrant orange sports car glides effortlessly along a winding coastal road, its glossy finish reflecting the golden hues of the setting sun. The car's aerodynamic design and polished chrome accents catch the light, creating a dazzling display of color and motion...



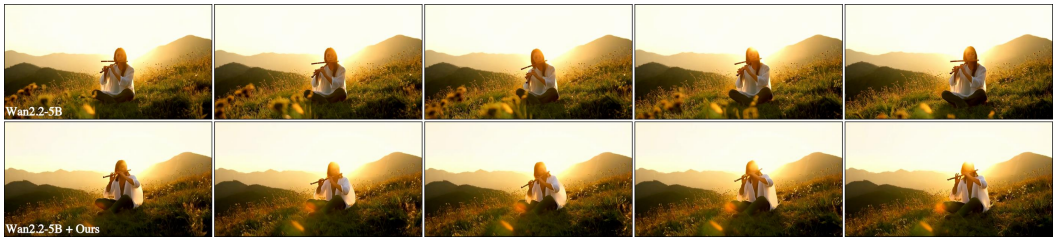
A serene individual sits comfortably in a cozy, softly lit room, wearing a plush white robe. They gently massage their legs, starting from the calves and moving upwards with slow, deliberate motions...



A thrill-seeker, clad in a bright red jumpsuit and a secure harness, leaps off a towering cliff, the vast canyon below stretching out in breathtaking detail. The camera captures the moment of freefall, the wind rushing past their exhilarated face, eyes wide with a mix of fear and excitement...



A vintage steam locomotive chugs along a winding railway through a picturesque countryside, its billowing smoke blending with the early morning mist. The train, with its polished brass and deep green carriages, glides past fields of golden wheat and vibrant wildflowers...



A serene individual, dressed in a flowing white shirt and dark trousers, sits cross-legged on a grassy hilltop at sunset, playing a wooden flute...

Figure 18: **More qualitative comparisons with and without STAS on Wan2.2-5B.**

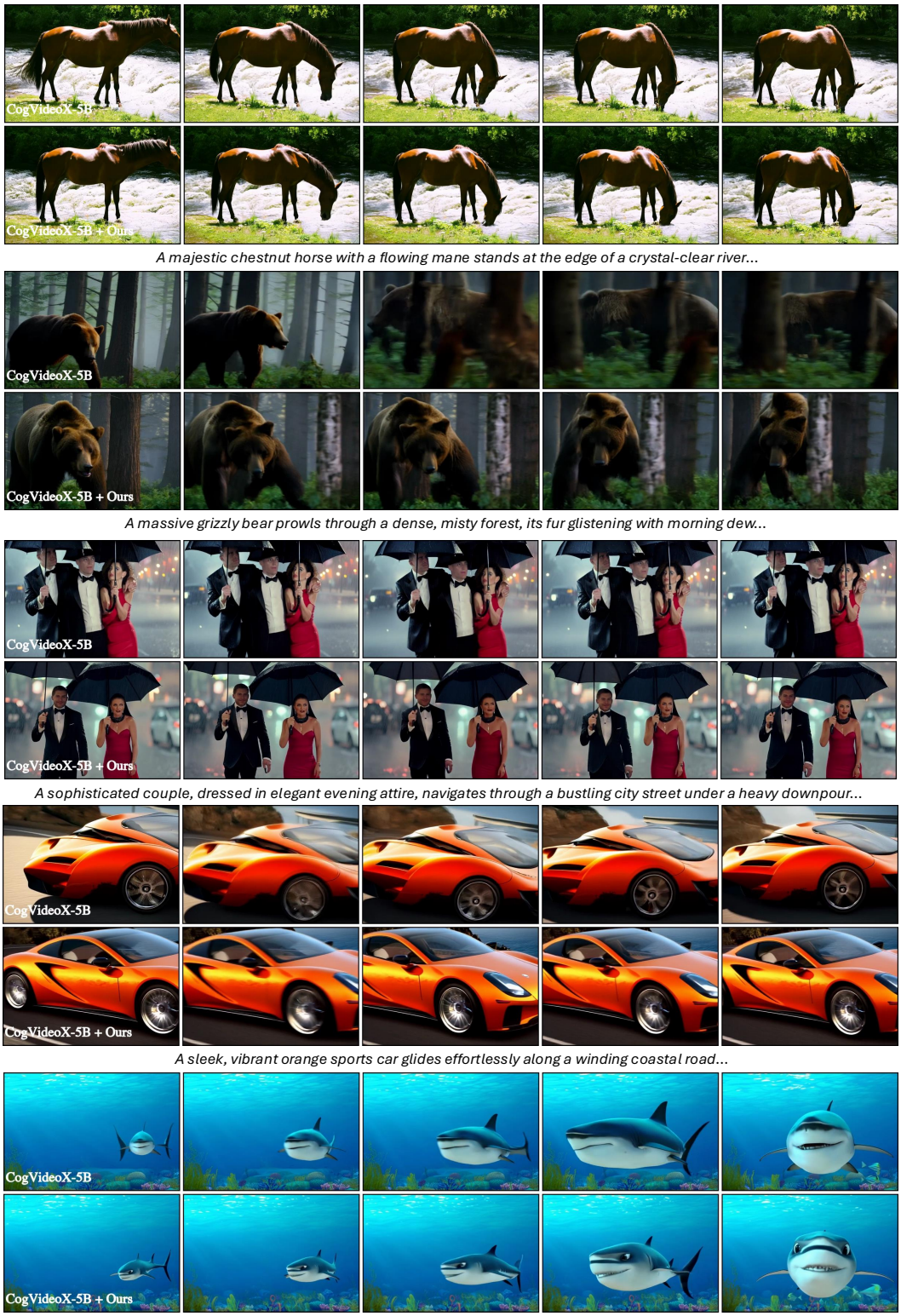


Figure 19: **More qualitative comparisons with and without STAS on CogvideoX-5B.**



Figure 20: **More qualitative comparisons with and without STAS on CogvideoX-5B.**



A golden retriever with a shiny coat strolls leisurely through a sun-dappled forest path, the morning light filtering through the trees casting a warm glow. The dog's tail wags gently as it sniffs the air, ears perked up, taking in the serene surroundings. The camera...



A young woman with long, flowing hair stands in a small, dimly lit bathroom, wearing a casual white t-shirt and jeans. She gazes thoughtfully at an old-fashioned porcelain toilet with a wooden seat, the room's vintage tiles adding a nostalgic touch. The scene...



A sleek, black motorcycle with chrome accents stands proudly on a winding mountain road, its polished surface gleaming under the midday sun. The camera zooms in to capture the intricate details of the engine, the leather seat, and the handlebars, showcasing...



A vibrant pink chair sits elegantly in the center of a sunlit room, its plush velvet upholstery catching the light. The chair's sleek, modern design features gently curved armrests and polished wooden legs, adding a touch of sophistication. Surrounding the...



A pixel art scene depicts a couple in elegant evening attire, caught in a sudden downpour. The man, in a sharp black tuxedo, holds a black umbrella, while the woman, in a flowing red gown, clutches a white umbrella. Raindrops cascade around them, creating a...

Figure 21: **More qualitative comparisons with and without STAS on LTX2 (base).**



In a sunlit meadow, a majestic chestnut horse with a flowing mane grazes peacefully beside a fluffy white sheep. The horse, with its sleek coat glistening in the sunlight, occasionally lifts its head to survey the serene landscape. The sheep, with its woolly fleece,...



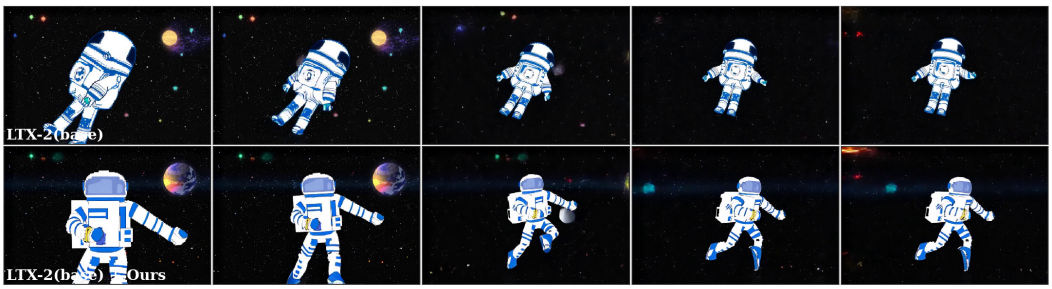
A joyful individual, bundled in a red winter coat, knitted hat, and gloves, stands in a snow-covered park, rolling a large snowball to form the base of a snowman. The scene is set against a backdrop of snow-laden trees and a serene, overcast sky. Next, they...



A delicate porcelain teacup, adorned with intricate floral patterns in soft pastels, sits on a rustic wooden table. Sunlight streams through a nearby window, casting a warm glow and gentle shadows on the cup's surface. The camera zooms in to reveal the fine...



A stylish couple, dressed in sleek, futuristic evening wear, navigate a neon-lit cityscape under a heavy downpour. The man, in a sharp, metallic silver suit, and the woman, in a shimmering, holographic gown, hold transparent umbrellas that reflect the...



A pixel art astronaut, clad in a white spacesuit with blue accents and a reflective helmet, floats gracefully through the vast expanse of space. Stars twinkle in the dark, pixelated sky, while distant planets and colorful nebulas add depth to the cosmic scene. The...

Figure 22: **More qualitative comparisons with and without STAS on LTX2 (base).**