TEMPORAL AWARE PRUNING FOR EFFICIENT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generative models have recently enabled high-quality video generation with Diffusion Transformer (DiT) architectures, but remain computationally intensive because generation requires attention computation over long spatiotemporal sequences. Token pruning has proven effective for ViTs and VLMs. However, most prior pruning methods are attention-based and operate per frame, failing to ensure the vital temporal coherence across frames in video generation tasks. In practice, naively adopting attention-only pruning causes noticeable degradation due to worsened background consistency, flickering, and reduced image quality. To address this, we propose TAPE, a training-free Temporal Aware Pruning for Efficient video generation. TAPE (i) applies temporal smoothing to align token importance across adjacent frames and suppress selection jitter; (ii) performs token reselection in selected layers to align token pruning with layers' diverse semantic focus and avoid error accumulation in specific areas; and (iii) adopts a timestep-level budget scheduling that prunes aggressively at early noisy steps and relaxes pruning during fidelity-critical refinement. The experimental results show that TAPE delivers significant speedups while preserving high visual fidelity, outperforming prior token reduction approaches.

1 INTRODUCTION

Recent advances in video generation have enabled increasingly high-quality video synthesis (Seedance et al., 2026; Kong et al., 2024). While early video generative models primarily relied on UNet-based backbones (Blattmann et al., 2023), recent systems increasingly adopt Diffusion Transformers (DiTs) (Peebles & Xie, 2023) due to their strong modeling capacity and scalability. However, video DiTs remain computationally expensive because self-attention operates over long spatiotemporal token sequences. For an image containing $W \times H$ tokens, self-attention has a complexity of $\mathcal{O}((WH)^2)$. Extending the sequence to a video with T frames increases it to $\mathcal{O}((WHT)^2)$. Consequently, the computation grows rapidly with video length. For example, generating a 5-second video (121 frames at 512×768 resolution with 50 denoising steps) using HunyuanVideo (Kong et al., 2024) takes approximately 14 minutes on an NVIDIA A100 GPU. This motivates reducing the computational cost without sacrificing visual quality or temporal coherence.

Recent studies have shown that token pruning can substantially reduce the computation costs of Vision Transformers (ViTs) (Liang et al., 2022; Kong et al., 2022) and Vision-Language Models (VLMs) (Chen et al., 2024; Zhang et al., 2025; Cao et al., 2023). By removing less informative visual tokens while preserving those important to model prediction, these methods reduce computation and memory with little performance degradation. Most existing token-pruning approaches determine token importance using attention scores (Liang et al., 2022; Chen et al., 2024). While such approaches have proven effective in image-generation and video-understanding tasks, video generation imposes an additional requirement: strong frame-to-frame coherence. Pruning decisions that vary inconsistently across adjacent frames can disrupt temporal continuity and lead to visible artifacts such as flickering. This naturally raises the question: *Do existing token-pruning strategies remain effective for video generation?*

We systematically study this question and find that naively applying attention-based pruning substantially degrades video quality. On VBench (Huang et al., 2024), the degradation is concentrated in background consistency, temporal flickering, and imaging quality, indicating that temporally in-

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.

consistent token selection is particularly harmful to video generation. These observations motivate a pruning mechanism that explicitly considers temporal coherence.

To this end, we propose TAPE, a training-free **T**emporal-**A**ware **P**runing framework for **E**fficient video generation. TAPE introduces three complementary mechanisms. First, *temporal smoothing* aligns token-importance scores across adjacent frames to suppress frame-to-frame selection jitter. Second, *token reselection* periodically revisits token importance across layers, allowing pruning decisions to adapt to different semantic focuses and preventing errors from accumulating in fixed regions. Third, *timestep-level budget scheduling* applies stronger pruning during early noisy denoising steps and progressively reduces the pruning rate as generation becomes increasingly sensitive to fine-grained details. We summarize our contribution as follows:

- We systematically investigate attention-only token pruning for video generation and identify temporally inconsistent token selection as a key source of quality degradation.
- We propose TAPE, a training-free pruning framework that combines temporal smoothing, token reselection, and timestep-level budget scheduling to reduce computation while preserving cross-frame coherence.
- Extensive experiments show that TAPE provides significant acceleration with minimal quality degradation. Specifically, TAPE attains a $1.5\times$ speedup while incurring about a 1-point decrease in the VBench total score. Compared with existing token-reduction approaches, TAPE consistently achieves superior performance across all pruning rates thanks to its temporally coherent pruning strategy.

2 Background

2.1 Video Generative Models

Modern video generative models typically synthesize videos through an iterative generative process that transforms noisy latent representations into structured visual content (Ho et al., 2020; Lipman et al., 2023). These models commonly perform generation in a compressed latent space encoded by a variational autoencoder (Rombach et al., 2022). Recent models increasingly adopt Diffusion Transformer (DiT) backbones (Peebles & Xie, 2023; Kong et al., 2024; Wan et al., 2025; Yang et al., 2024), which represent videos as sequences of spatiotemporal tokens and model their interactions through self-attention. Compared with image generation, video generation introduces substantially longer token sequences and additional temporal dependencies across frames, making attention computation a major efficiency bottleneck as video resolution, duration, and frame count increase.

2.2 Token Reduction for Transformer

Despite their strong modeling capacity, Transformers suffer from the quadratic scaling of self-attention with respect to token sequence length (Dosovitskiy et al., 2020), motivating token reduction methods that reduce computation and memory by processing fewer redundant tokens. A widely adopted strategy is token pruning, which estimates token importance, often using attention scores, and removes less informative tokens (Liang et al., 2022; Kong et al., 2022; Yu & Wu, 2023). For example, EViT (Liang et al., 2022) retains tokens with high attention-based importance while aggregating the remaining tokens to preserve global information. Complementary to pruning, token merging methods such as ToMe (Bolya & Hoffman, 2023) reduce redundancy by combining tokens with similar representations, thereby shortening the sequence.

Token reduction has also been extended beyond static vision to temporally structured and multimodal inputs. STA-ViT (Ding et al., 2023) exploits inter-frame redundancy together with semantic importance to prune spatiotemporal tokens. In Vision-Language Models (VLMs), related methods identify redundant visual tokens using signals such as cross-modal attention and visual-token similarity (Chen et al., 2024; Zhang et al., 2025; Cao et al., 2023; Huang et al., 2025). While these approaches demonstrate substantial redundancy in visual token sequences, they are not specifically designed to maintain consistent pruning decisions across frames during video generation, which we investigate in the following section.

Table 1: Results of applying the attention-based token pruning approach to the text-to-video generation. The results are obtained on VBench. The HunyuanVideo model is used as the base model in this experiment.

Method	Speedup ↑	Total Score (%) ↑	Quality Score (%) ↑	Semantic Score (%) ↑
Baseline w/o token Pruning	1.0x	83.24	85.09	75.82
Attn-based token prune (keep rate = 80%)	1.4x	79.70	81.05	74.30
Attn-based token prune (keep rate = 70%)	1.7x	78.25	79.36	73.84

Table 2: Detailed results of the evaluation dimensions of the quality score. "kr" denotes the keep rate. All metrics are higher-the-better, and **dimensions with significant score drops are highlighted in bold.**

Method	Subject Consistency	**Background Consistency**	**Temporal Flickering**	Motion Smoothness	Aesthetic Quality	**Imaging Quality**	Dynamic Degree
Baseline w/o Pruning	97.37	**97.76**	**99.44**	98.99	60.36	**67.56**	70.83
Attn-based prune (kr = 80%)	96.18	**94.06**	**96.20**	97.62	59.10	**63.48**	68.61
Attn-based prune (kr = 70%)	96.24	**93.26**	**95.06**	96.59	58.22	**61.24**	68.06

3 Does Existing Token Pruning Work Well for Text-to-Video Generation?

We first investigate whether conventional attention-based token pruning can be directly applied to text-to-video (T2V) generation. Although attention-based pruning has proven effective in other visual tasks, video generation imposes an additional requirement of maintaining temporal coherence across frames. A token that receives low importance in one frame may still contribute to preserving consistent content in neighboring frames, and independently pruning such tokens can lead to temporally inconsistent computation. To examine this issue, we apply a representative attention-based pruning strategy (Liang et al., 2022) to video generation and analyze its impact on generation quality and temporal coherence.

3.1 Applying Attention-based Token Pruning to T2V Generation

We adopt an attention-based token pruning strategy similar to EViT (Liang et al., 2022). Specifically, the self-attention scores are used to estimate the importance of individual tokens: tokens receiving lower attention scores are regarded as less important, and the lowest-ranked tokens are selected for pruning according to a predefined keep rate. In prior works (Liang et al., 2022; Kong et al., 2023), pruning is applied at selected transformer layers (e.g., the 4th, 7th, 10th blocks of the DeiT model), and once a token is pruned at a given layer, it is discarded and will not appear in subsequent layers. However, unlike conventional discriminative tasks such as object recognition, where less important tokens can be safely discarded, generative tasks like video generation require preserving all tokens for reconstruction. Therefore, importantly, we do not physically discard tokens but skip their computation, maintaining their positional and structural roles so that every token can contribute to the subsequent reconstruction through VAE decoder (Liu et al., 2024).

We use HunyuanVideo (Kong et al., 2024) as the base model and evaluate the generated videos on a representative text-to-video generation benchmark, VBench (Huang et al., 2024). Following prior works (Liang et al., 2022; Kong et al., 2023), token pruning is also applied at selected layers (e.g., the 10th, 30th, and 50th blocks of the HunyuanVideo model, which contains 60 Transformer blocks in total). Once a token is chosen to be skipped, it remains skipped in subsequent layers. We apply the same pruning ratio across temporal frames to avoid frame-wise pruning imbalance. Table 1 reports the quantitative results. The keep rate denotes the fraction of tokens retained for computation at each pruning point. Since pruning is cumulative, tokens skipped at earlier blocks remain inactive in later blocks.

As shown in the table, although this pruning method improves inference speed, it causes a significant degradation in video generation quality. With a keep rate of 80%, the model achieves a 1.4× speedup, but already suffers a noticeable drop in both total and quality scores, decreasing from 83.24 to 79.70 and from 85.09 to 81.05, respectively. When the keep rate is further reduced to 70%, the performance degradation becomes more substantial, where the total score drops to 78.25, and the



Figure 1: An example to show pruning areas in two frames of a video. The token reduction rate is 30% in this example.

quality score falls by nearly 6 points to 79.36 compared to the baseline. Although the semantic score also decreases, its change is relatively small compared to the significant loss in visual and temporal quality. To better understand the cause of this degradation, we further analyze the individual evaluation dimensions that contribute to the quality score.

As shown in Table 2, this attention-based pruning noticeably affects several key evaluation dimensions of the quality score. Among them, temporal flickering, background consistency, and imaging quality, experience the most significant drops as the keep rate decreases. Background consistency declines sharply from 97.76 to 93.26, indicating unstable reconstruction of less-informative background regions across frames. The temporal flickering score drops from 99.44 to 95.06, revealing visible frame-level incoherence and motion discontinuities. Imaging quality also decreases from 67.56 to 61.24, suggesting degraded visual clarity in generated frames.

The observed degradation across multiple quality dimensions can be attributed to how attention-based token pruning (i.e., token skipping) interacts with the spatiotemporal structure of video generation. The large drops in background consistency and temporal flickering stem from incoherent pruning decisions across consecutive frames. Because background tokens generally receive lower attention scores, they are frequently pruned; however, the specific background regions being pruned might vary from frame to frame. Even when these regions are indeed low-attention areas, the lack of temporal coherence in the pruning masks leads to visible inconsistencies and flicker. As shown in Figure 1, although both frames are pruned in less-informative background regions, the pruned areas differ notably. In Frame N, the pruned regions include both the forest and the road, with a larger proportion located in the forest, whereas in Frame M, the pruning primarily occurs along the road. This frame-wise incoherence exposes a key limitation of applying attention-based pruning to video generation without considering temporal coherence.

Moreover, the cumulative pruning mechanism in the current strategy significantly affects the score of imaging quality by allowing errors to propagate through deeper layers. In this mechanism, once a token is skipped, it no longer participates in subsequent computations, causing its representation to become outdated and insufficiently refined. When these degraded features are eventually passed to the VAE decoder for frame reconstruction, the accumulated errors result in a significant decline in overall visual quality.

We use attention-based pruning as an example in this preliminary study. However, this issue is not unique to attention-based token pruning; it also exists in other token reduction approaches that overlook temporal coherence in video generation, such as the popular token merging method ToMe (Bolya & Hoffman, 2023) and the spatial–temporal adaptive transformer STA-ViT (Ding et al., 2023). A more detailed comparison of these approaches is provided in the evaluation section (section 5.2).

4 DESIGN

To address the limitations of prior approaches, we propose TAPE, a training-free Temporal Aware Pruning for Efficient video generation. At its core, TAPE applies ***temporal smoothing*** that aligns token-importance scores across neighboring frames to mitigate temporal incoherence in token selection. Additionally, TAPE performs ***token reselection*** at selected layers, allowing tokens with different semantic roles to be revisited and avoiding overly focused pruning. Moreover, TAPE uses

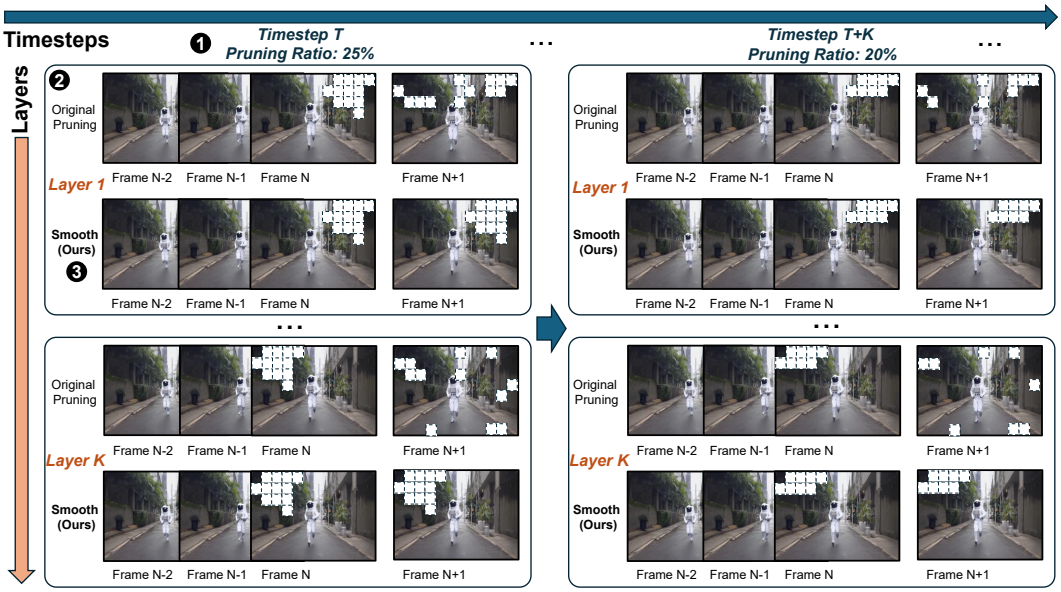


Figure 2: Overview of TAPE. At timestep T, ① timestep-aware scheduling first decides the pruning ratio, which will be reduced at late steps; ② Token reselection is conducted intermittently, aligning pruning decisions with diverse semantic focuses in different layers; upon selection, ③ temporal smoothing blends current and aligned previous scores to enforce temporally coherent pruning.

a ***timestep-level budget scheduling*** that relaxes pruning (i.e., reduces pruning ratio) during fidelity-critical steps, reducing compute without harming temporal coherence or visual quality.

The overview of TAPE is illustrated in Figure 2. Given a timestep T, TAPE first gets the pruning ratio with the *timestep-level budget scheduling*, which gradually reduces the pruning strengths to let more tokens contribute to the computation in late timesteps. Next, *token reselection* revisits token importance at selected layers, ensuring pruning decisions reflect the model's progressively refined representations. The token selection is driven by *temporal smoothing*, which computes smoothed importance by blending each token's current score with its aligned score from the previous frame, ensuring more temporally coherent pruning than attention-score-based selection in prior works. Note that, since generative tasks require all tokens for reconstruction, we skip computation for less important tokens rather than physically discard them, ensuring all tokens contribute to the subsequent VAE-based reconstruction.

4.1 Temporal Smoothing

Motivations. While prior attention-based pruning methods ignore temporal coherence, accurately and efficiently modeling it during video generation is non-trivial. In particular, avoiding incoherent pruning requires balancing semantic change and spatial/temporal stability. On the one hand, the pruning policy must track evolving semantics so that important regions are not mistakenly pruned. On the other hand, the pruned region should remain coherent across adjacent frames so that the affected area does not drift unpredictably, which would otherwise induce flicker and disrupt motion coherence. While Video-analytical solutions, such as optical flow (Cheng et al., 2017; Tu et al., 2022) or dense correspondence estimation (Truong et al., 2021; Zheng et al., 2022), or learned motion prediction (Bao et al., 2019; Kim & Kang, 2020), appear promising for this problem, they fall short for two reasons. First, they require accurate physical inputs to recover motion. In contrast, during the generation process, the input information at most time steps is noisy and often nonphysical, making such signals unreliable or unobservable. Second, these methods introduce substantial overheads, such as additional models, training, or per-step motion estimation. These factors can outweigh the computational savings from pruning. This motivates a lightweight, training-free mechanism that stabilizes token importance scores over time without relying on external motion supervision.

Temporal Smoothing Importance Score. Our design stems from an intuitive observation: *Consecutive frames in videos are strongly correlated.* Specifically, backgrounds, object identities, and coarse geometry change slowly between consecutive frames, and even motion appears as small, continuous changes. Consequently, token-importance scores should remain coherent across neighboring frames within the same region. We therefore propose a *temporal smoothing* scheme that blends each token's current importance with its previous-frame counterpart, preventing drastic inter-frame swings in scores—and hence abrupt changes in token selection—while preserving genuine motion. Specifically, given a token i at frame n, we smooth its importance score as:

$$\tilde{s}_i^n = \alpha \cdot s_i^n + (1 - \alpha) \cdot s_i^{n-1} \tag{1}$$

where s_i^n and s_i^{n-1} are the original importance score (e.g., attention score) at frames n and $n-1$, respectively, and $\tilde{s}_i^n$ denotes the smoothed importance score which blends the token's importance score across frames. The hyperparameter $\alpha \in [0, 1]$ controls smoothing strength: smaller α gives heavier smoothing (more reliance on the previous frame), and larger α trusts the current frame. According to our practice, we set the $\alpha = 0.5$. We have conducted a sensitivity study to analyze the choice of α in supplementary material (Section A).

By doing so, the smoothed importance score $\tilde{s}_i^n$ not only weighs the token's importance in the current frame n but also considers its importance in the previous frame $n-1$. Consequently, as exemplified in Figure 2, tokens selected in the last frame are more likely to be selected again at the current frame, avoiding incoherent pruning across consecutive frames.

4.2 Token Reselection

Motivation: It is neither necessary nor desirable to prune the same spatial token across all layers. First and foremost, different layers emphasize different semantics. As shown in Figure 3, the distribution of attention scores varies across diverse layers. As a result, some layers primarily encode stationary structure (layout, coarse geometry), whereas other layers capture dynamic semantics (fine appearance changes and motion cues). Second, pruning the same set of tokens across diverse layers accumulates errors in that region, leading to noisy outputs such as texture loss and flicker. By contrast, token reselection allows pruned tokens to be distributed across wider spatial ranges, thereby avoiding excessive focus on specific areas.

Token reselection. To this end, we propose to intermittently update token importance by recomputing the *smoothed importance scores* (Eq. 1) for all tokens, re-ranking them, and pruning the least informative ones, which helps prevent pruning from becoming overly biased toward early decisions. While more frequent updates improve adaptivity, they also increase computational overhead, as the "pruned" tokens need to be brought back into the attention computation more often; in contrast, too few reselections may limit their effectiveness. In our practice, we empirically choose to reselect tokens every 10 layers. We provide a sensitivity study on this reselection frequency in the supplementary material (Section A).

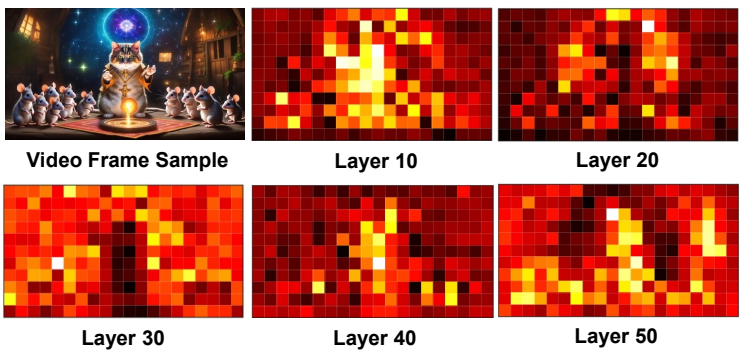


Figure 3: An example of attention distribution across layers. Each block (i.e., token) in the attention map corresponds to the same spatial region of the video frame. Brighter colors indicate tokens that receive higher attention scores from others. The shown attention scores are averaged over all heads.

Table 3: Comparison of different timestep-level pruning schedules under the same attention-based token pruning setting (please refer to Section 3.1 for details of the attention-based pruning).

Method	Total Score (%) ↑	Quality Score (%) ↑	Semantic Score (%) ↑
Constant prune	79.70	81.05	74.30
High→Low prune	81.33	82.85	75.23
Low→High prune	77.53	78.77	72.58

4.3 Timestep-level Budget Scheduling

Motivations. Generation proceeds from highly noisy, non-physical latents to increasingly clean and detail-sensitive frames (Lu et al., 2022; Ho et al., 2020; Lipman et al., 2023). Early in denoising, the signal is coarse and comparatively robust to token removal; late in denoising, fine textures and motion alignment are established, and the model becomes sensitive to drops. To this end, we raise a question: *How should the keep/prune ratio vary along the denoising trajectory to balance efficiency with visual and temporal quality?*

Design Space. To explore the potential effects of changing the pruning budget during video generation, we compare three scheduling policies. (1) *Constant Prune*: We prune 20% tokens at all timesteps. (2) *High→Low Prune*: We allocate 30% tokens to prune at the beginning, then reduce the value by 5% every 10 steps, where the total timesteps is 50 in our experiment. (3) *Low→High Prune*: We allocate 10% tokens to prune at the beginning, then increase the value by 5% every 10 steps. Note that all schedules share the same average pruning ratio, and the resulting FLOPs differences are within 1%, so we regard the computational cost as comparable. In this experiment, we isolate the effect of the pruning schedule for comparison, by following the attention-based token pruning configuration in Section 3.1, without applying temporal smoothing or token reselection.

As shown in Table 3, we find that *High→Low* consistently offers the best trade-off: it harvests substantial compute savings during noisy early steps while preserving textures and motion coherence as frames become detail-sensitive. *Constant* yields moderate savings with moderate impact, whereas *Low→High* tends to degrade late-stage fidelity (texture loss, flicker), indicating that aggressive late pruning harms refinement.

Timestep-level Budget Scheduling. To this end, we adopt a *High→Low Prune* scheduling scheme. Given a target token reduction ratio ρ, we set the pruning rate to start from a slightly higher value $\rho + \Delta$ and end at a slightly lower value $\rho - \Delta$, where Δ controls the deviation (e.g., $\Delta = 10\%$). The pruning rate is updated at M evenly spaced points along the T denoising steps (e.g., every 10 steps when $T = 50$), forming a simple linear decay from $\rho + \Delta$ to $\rho - \Delta$, where the midpoint naturally corresponds to the target ratio ρ. And we set $\Delta = 10\%$ by default in our experiments.

5 Evaluation

5.1 Experimental Setup

In this section, we evaluate the effectiveness of our training-free temporal-aware pruning framework, TAPE, on text-to-video generation. Our evaluation is based on the HunyuanVideo model with 13B parameters and 60 DiT blocks. All experiments are performed directly on the pre-trained model, without any additional training or fine-tuning. The quantitative comparisons reported in this evaluation section follow the standard evaluation protocol of VBench. Unless otherwise specified, videos are generated at a 720p resolution. We have also conducted experiments using another model and metrics, and the results are provided in the supplementary material (Section B and Section C). All experiments are executed on a single A100 (80GB) GPU. We compare TAPE to three representative token reduction methods: i) attention-based token pruning method EViT, ii) similarity-based token merging approach ToMe, and iii) STA-ViT, which prunes tokens using a Semantic-aware Temporal Accumulation score that identifies temporally redundant and semantically weak tokens.

Table 4: Comparison of different token reduction approaches on VBench. The full results on all evaluation dimensions of VBench are in the supplementary material (Section E).

Token Reduction Rate	Method	Speedup ↑	Total Score ↑	Quality Score ↑	Semantic Score ↑
	Baseline	1.0x	83.24	85.09	75.82
20%	EViT	1.3x	81.82	83.62	74.64
	ToMe	1.2x	82.18	83.92	75.22
	STA-ViT	1.2x	80.87	82.77	73.29
	TAPE (Ours)	1.3x	83.16	84.90	76.20
30%	EViT	1.5x	81.15	82.96	73.90
	ToMe	1.4x	80.92	82.74	73.66
	STA-ViT	1.4x	79.30	81.30	71.28
	TAPE (Ours)	1.5x	82.01	83.64	75.50
40%	EViT	1.8x	79.38	81.53	70.77
	ToMe	1.8x	78.26	79.88	71.78
	STA-ViT	1.7x	77.46	79.44	69.53
	TAPE (Ours)	1.8x	81.04	83.20	72.42

5.2 Main Results

Here we evaluate all the token reduction approaches under different token reduction rates on VBench. As shown in Table 4, TAPE delivers significant inference speedups while incurring only minor visual degradation. With a more conservative pruning ratio (e.g., 20%), the VBench score remains nearly unchanged, indicating that our temporal-aware pruning strategy can remove redundant computation without harming generation quality when an appropriate reduction rate is chosen.

When compared with the three representative token-reduction methods, TAPE consistently achieves notably higher scores across all pruning rates. For example, at a 30% reduction rate, TAPE attains a total score of 82.01, whereas EViT, ToMe, and STA-ViT fall to 81.15, 80.92, and 79.30, respectively. This is because these methods do not explicitly account for temporal coherence, which limits their effectiveness in video generation. In contrast, TAPE incorporates temporal smoothing, token reselection, and a timestep-aware pruning schedule, enabling it to maintain cross-frame consistency and preserve visual fidelity while still providing notable acceleration. We provide a visualization of the pruning areas in some samples in the supplementary material (Section D), to illustrate how our pruning method produces more temporally coherent token selection and avoids frame-wise jitter.

5.3 Visualization of Generated Videos

Figure 4 presents qualitative examples comparing the baseline model with our TAPE design at 20% and 40% token reduction rates. Even after pruning a substantial portion of tokens, the generated videos remain highly faithful to the prompts. For instance, in the example of "roaring engine", the model not only preserves the main subject and the details, but also correctly follows the camera instruction (*zoom out*). At a 20% reduction rate, the results are nearly indistinguishable from the baseline. Although a 40% token reduction rate introduces slight softness in some regions, the overall structure, motion, and prompt semantics are still well captured, demonstrating that TAPE maintains strong visual fidelity even under aggressive pruning. We provide additional visualizations of videos generated with our pruning method TAPE in the supplementary material (Figure 6 in Section F), offering a broader view of its visual quality under different scenarios.

5.4 Ablation Study

To evaluate the effectiveness of the components separately, we further compare TAPE's performance with each component enabled. Specifically, we compare the performance of TAPE under four versions: (i) *TAPE-S* enables temporal smoothing only; (ii) *TAPE-R* enables token reselection only; (iii) *TAPE-T* enables timestep-aware budget scheduling only; (iv) *TAPE* enables all. As shown in Table 5, each component of TAPE provides a clear improvement over the naïve attention-based pruning method EViT. While smoothing, reselection, and timestep-aware scheduling each offer noticeable gains on their own, none of them matches the performance achieved when all three are combined.

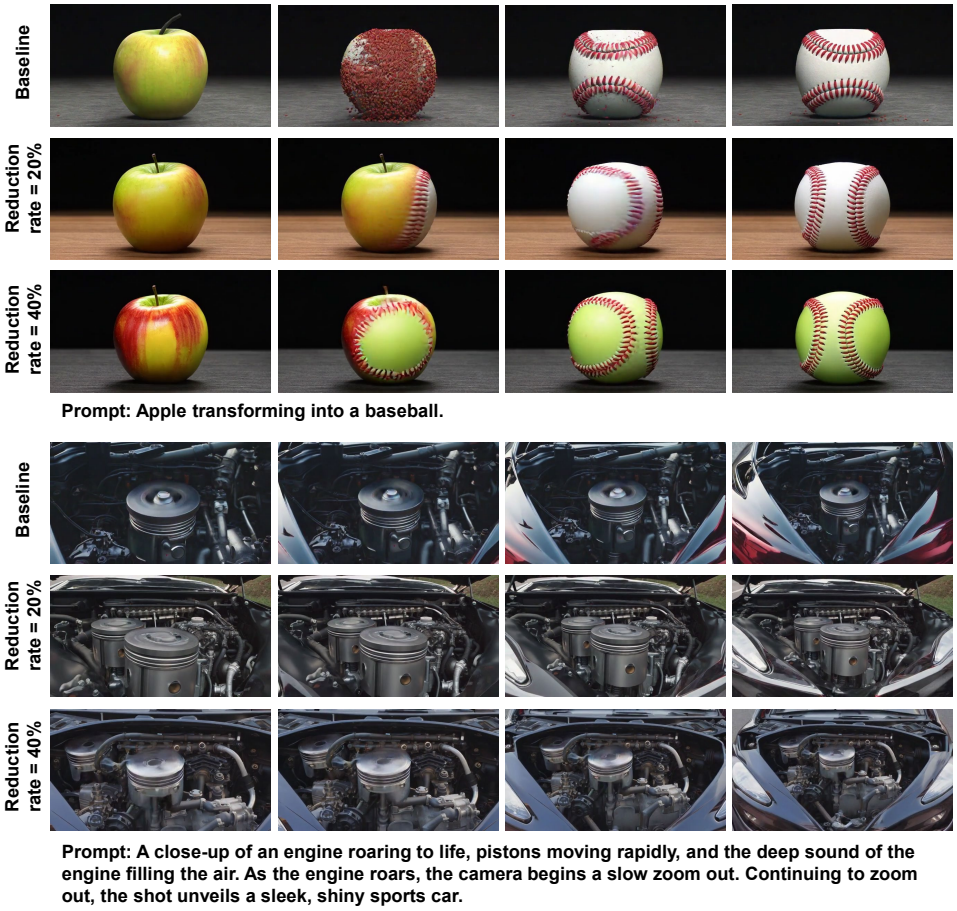


Figure 4: Visualization of the generated video.

Table 5: Ablation results of TAPE components. We compare *Speedup* (higher is better) and total VBench score. The token reduction rate is 30%.

Method	Speedup ↑	Total Score ↑
Baseline (no pruning)	1.0x	83.24
EViT	1.5x	81.15
TAPE-S (smoothing only)	1.5x	81.76
TAPE-R (reselection only)	1.5x	81.52
TAPE-T (timestep schedule only)	1.5x	81.66
TAPE	1.5x	82.01

This demonstrates that the three mechanisms are complementary, and together they enable a significantly stronger speed–quality tradeoff.

6 Conclusion

In this paper, we presented TAPE, a temporal-aware token pruning mechanism that enables efficient inference for video generation. By enforcing temporal consistency in token selection and adapting pruning across layers and timesteps, TAPE effectively reduces redundant computation while preserving semantic and temporal coherence. Comprehensive evaluations validate its advantage over existing token pruning and merging methods, demonstrating its practicality for accelerating modern video generative models without compromising visual fidelity.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research ideation and execution, including brainstorming and refining research directions, methodological choices, and experimental design. We also used generative AI tools for literature discovery and retrieval, as well as for drafting and polishing portions of the manuscript to improve clarity and readability. All AI-assisted research ideas, literature references, technical suggestions, and written content were reviewed and verified by the authors. The authors made all final decisions regarding the research methodology, experiments, analysis, and presentation, and take full responsibility for the final content of this work.

REFERENCES

Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):933–948, 2019.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4599–4603, 2023.

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. PuMer: Pruning and merging tokens for efficient vision language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12890–12903, Toronto, Canada, July 2023. Association for Computational Linguistics.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pp. 686–695, 2017.

Shuangrui Ding, Peisen Zhao, Xiaopeng Zhang, Rui Qian, Hongkai Xiong, and Qi Tian. Prune spatio-temporal tokens by semantic-aware temporal accumulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16945–16956, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nayoung Kim and Je-Won Kang. Dynamic motion estimation and evolution video prediction network. *IEEE Transactions on Multimedia*, 23:3986–3998, 2020.

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 arXiv:2412.03603*, 2024.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, pp. 620–640. Springer, 2022.

Zhenglun Kong, Haoyu Ma, Geng Yuan, Mengshu Sun, Yanyue Xie, Peiyan Dong, Xin Meng, Xuan Shen, Hao Tang, Minghai Qin, et al. Peeling the onion: Hierarchical reduction of data redundancy for efficient vision transformer training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8360–8368, 2023.

Youwei Liang, Chongjian GE, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. EVit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2658–2668, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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.

Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5714–5724, 2021.

Zhigang Tu, Hongyan Li, Wei Xie, Yuanzhong Liu, Shifu Zhang, Baoxin Li, and Junsong Yuan. Optical flow for video super-resolution: A survey. *Artificial Intelligence Review*, 55(8):6505–6546, 2022.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Hao Yu and Jianxin Wu. A unified pruning framework for vision transformers. *Science China Information Sciences*, 66(7):179101, 2023.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025.

Linxin Zheng, Guobao Xiao, Ziwei Shi, Shiping Wang, and Jiayi Ma. Msa-net: Establishing reliable correspondences by multiscale attention network. *IEEE Transactions on Image Processing*, 31: 4598–4608, 2022.

A SENSITIVITY STUDY

A.1 SENSITIVITY STUDY ON SMOOTHING WEIGHT

In this section, we study the sensitivity of our method to the smoothing weight α. We evaluate the total VBench score, speedup, and the three quality-related dimensions where attention-only pruning degrades most, as discussed in Section 3.1 and Table 2 (i.e., *Background Consistency*, *Temporal Flickering*, and *Imaging Quality*) under $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. As shown in Table 6, there is a clear trade-off: larger α gives more weight to the current-frame scores and tends to preserve *Imaging Quality* better but yields weaker gains in *Background Consistency* and *Temporal Flickering* (e.g., at $\alpha = 0.9$, *Imaging Quality* reaches 66.16, while *Background Consistency* and *Temporal Flickering* drop to 95.32 and 97.73). Conversely, smaller α emphasizes the previous-frame scores and improves temporal coherence (e.g., at $\alpha = 0.1$, *Background Consistency* and *Temporal Flickering* rise to 96.33 and 98.68, but *Imaging Quality* decreases to 64.52). Overall, $\alpha = 0.5$ achieves the best balance—with the highest total score and consistently strong performance across dimensions—so we adopt $\alpha = 0.5$ by default.

A.2 SENSITIVITY STUDY ON TOKEN RESELECTION FREQUENCY/TIMES

We further study how often to perform token reselection across layers. Specifically, we vary the reselection times over the 60 attention blocks of Hunyuan-Video while fixing the token reduction rate to 30%, and report the VBench scores and speedups in Table 7. As the reselection becomes more frequent, the model can better adapt token choices to different layers' semantic focus, which generally improves quality. When we increase the reselection times from 2 to 3 and then to 5, the total score rises from 81.27 to 81.51 and 82.01, with both quality (from 83.07 to 83.26 to 83.64) and semantic scores (from 74.08 to 74.53 to 75.50) improving consistently. However, pushing the frequency further to 10 times (every 5 blocks) brings only marginal gains over 5 times (e.g., total score 82.04 vs. 82.01, quality 83.68 vs. 83.64, semantic 75.48 vs. 75.50) while reducing the speedup from 1.5× to 1.4× due to the overhead of repeatedly recomputing pruned tokens (i.e., skipped tokens). These results show that overly infrequent reselection harms quality, whereas overly frequent reselection yields diminishing returns but higher cost; thus, we use 5 reselections (every 10 blocks in the HunyuanVideo model) as our default setting.

Table 6: Sensitivity of TAPE to the smoothing weight α. In this experiment, we set the token reduction rate to 30%. We compare *Speedup* (higher is better), total VBench score, and three key dimensions where attention-only baselines degrade most.

Method	Speedup ↑	Total Score ↑	VBench Dimensions ↑		
			Background Consistency	Temporal Flickering	Imaging Quality
Baseline (no pruning)	1.0x	83.24	97.76	99.44	67.56
TAPE (α=0.1)	1.5x	80.67	96.33	98.68	64.52
TAPE (α=0.3)	1.5x	81.64	96.35	98.43	65.48
TAPE (α=0.5)	1.5x	82.01	96.23	98.36	66.03
TAPE (α=0.7)	1.5x	81.45	95.67	97.90	66.11
TAPE (α=0.9)	1.5x	81.22	95.32	97.73	66.16

Table 7: Sensitivity study on token reselection frequency/times. In this experiment, the Hunyuan-Video model (with 60 attention blocks in total) is used as the base model. The token reduction rate is set to 30%.

Reselection Times (Frequency)	Speedup ↑	Total Score ↑	Quality Score ↑	Semantic Score ↑
10 Times (every 5 blocks)	1.4x	82.04	83.68	75.48
5 Times (every 10 blocks)	1.5x	82.01	83.64	75.50
3 Times (every 15 blocks)	1.5x	81.51	83.26	74.53
2 Times (every 20 blocks)	1.5x	81.27	83.07	74.08

Table 8: Results on VBench. In this experiment, Wan2.1-T2V-1.3B is used as the base model.

Token Reduction Rate	Method	Speedup ↑	Total Score ↑	Quality Score ↑	Semantic Score ↑
	Baseline	1.0x	83.31	85.23	75.65
20%	EViT	1.3x	81.88	84.04	73.22
	ToMe	1.3x	80.93	83.22	71.76
	STA-ViT	1.2x	81.60	83.61	73.54
	TAPE (Ours)	1.3x	83.27	85.20	75.53
30%	EViT	1.5x	80.34	82.28	72.60
	ToMe	1.5x	79.30	81.54	70.34
	STA-ViT	1.4x	79.16	81.05	71.62
	TAPE (Ours)	1.5x	82.52	84.48	74.67
40%	EViT	1.8x	78.44	80.35	70.78
	ToMe	1.7x	77.93	80.03	69.54
	STA-ViT	1.7x	77.90	79.63	70.98
	TAPE (Ours)	1.8x	81.58	83.76	72.85

B Evaluation when using Wan as Base Model

In this section, we use Wan2.1-T2V-1.3B (Wan et al., 2025) as the base model to evaluate the performance of our proposed TAPE pruning framework. As shown in Table 8, under the same token reduction rate, TAPE achieves comparable speedup to other pruning methods but consistently obtains the highest Total, Quality, and Semantic Scores among all pruning methods, indicating the highest video quality. Notably, at the 20% token reduction rate, TAPE's VBench scores are almost unchanged compared to the baseline without pruning, indicating that it largely preserves video quality while still providing acceleration. These results demonstrate the effectiveness of TAPE in delivering efficient token pruning with minimal quality loss, even when applied to a different base model.

Table 9: Comparison of different token reduction methods using the FVD metric. In this experiment, Hunyuan-Video is used as the base model.

Token Reduction Rate	Method	Speedup ↑	FVD ↓
20%	EViT	1.3x	496.2
	ToMe	1.2x	522.9
	STA-ViT	1.2x	559.5
	TAPE (Ours)	1.3x	468.7
30%	EViT	1.5x	543.6
	ToMe	1.4x	559.1
	STA-ViT	1.4x	586.8
	TAPE (Ours)	1.5x	489.3
40%	EViT	1.8x	578.4
	ToMe	1.8x	597.2
	STA-ViT	1.7x	628.9
	TAPE (Ours)	1.8x	516.1

C Evaluation on FVD Metric

In this section, we evaluate our pruning framework TAPE using the metric of Fréchet Video Distance (FVD) (Unterthiner et al., 2018). We use the text prompts from VBench to generate videos. For each prompt, we first generate a reference video with the Hunyuan-Video model without pruning. We then generate videos with all pruning methods using the same prompts and five random seeds (reuse the same set of seeds across all methods). We treat the videos generated by the Hunyuan-Video without pruning as the reference, and calculate FVD for each pruning method with respect to this unpruned baseline. As shown in Table 9, TAPE consistently outperforms the other three token reduction methods by yielding the lowest FVD with similar speedup. These results further confirm

that TAPE better preserves the visual fidelity of the base model compared to other token reduction approaches.

D VISUALIZATION ON PRUNED AREAS

In this section, we use a representative example to visualize the pruned areas (20% token reduction) across multiple frames for both our TAPE method and EViT, where EViT prunes tokens purely based on attention scores. These results are obtained from the first transformer layer at the final denoising step. As shown in Figure 5, both TAPE and EViT mainly remove less-informative background regions. However, our temporal-aware pruning produces much more coherent masks across frames. For example, if we take a closer look at the top-right pruned regions inside the red box, EViT's masks fluctuate noticeably from frame to frame, whereas TAPE maintains a stable and coherent pruning pattern. This helps reduce frame-wise flickering.

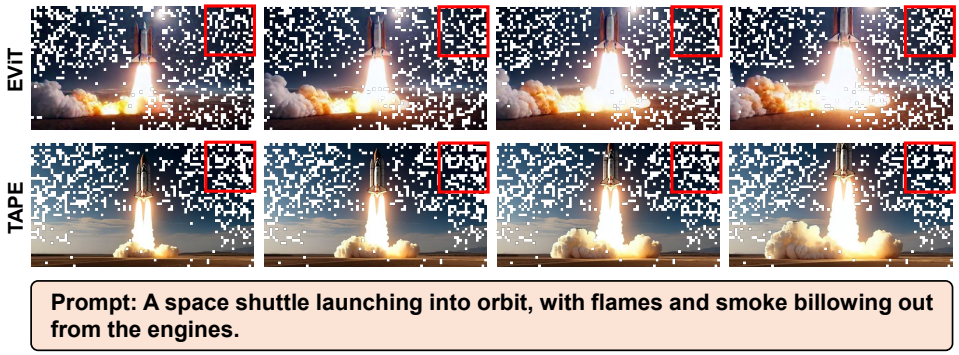


Figure 5: An example visualization of pruned areas across frames for EViT and our proposed TAPE with 20% token reduction.

E DETAILED RESULTS OF ALL VBENCH DIMENSIONS

In this section, we present the detailed VBench results for all evaluation dimensions corresponding to Table 4 in the main paper. Specifically, the scores of all quality-related dimensions are reported in Table 10, while the scores of all semantic-related dimensions are reported in Table 11.

Table 10: VBench quality scores in all evaluation dimensions under different token reduction methods.

Token Reduction Rate	Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Aesthetic Quality	Imaging Quality	Dynamic Degree
	Baseline	97.37	97.76	99.44	98.99	60.36	67.56	70.83
20%	EViT	96.89	96.08	98.22	98.78	59.83	65.64	70.28
	ToMe	97.45	96.06	98.20	98.61	59.98	66.60	71.94
	STA-ViT	96.77	94.78	97.30	98.56	59.52	65.20	70.28
	TAPE (Ours)	97.36	97.55	99.33	98.95	60.13	67.36	70.56
30%	EViT	96.40	95.24	97.66	98.83	59.23	64.98	70.28
	ToMe	96.53	94.95	97.42	98.76	58.94	65.27	69.72
	STA-ViT	96.45	92.80	96.30	98.41	58.68	63.64	69.17
	TAPE (Ours)	96.47	96.23	98.36	98.90	59.10	66.03	70.28
40%	EViT	95.72	93.48	96.82	98.53	58.28	63.38	69.72
	ToMe	95.02	91.07	95.63	98.30	57.30	62.16	68.89
	STA-ViT	95.35	90.77	94.87	98.22	57.63	61.54	68.33
	TAPE (Ours)	95.88	95.75	98.11	98.82	58.77	65.79	70.28

Table 11: VBench semantic scores in all evaluation dimensions under different token reduction methods.

Token Reduction Rate	Method	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
	Baseline	86.10	68.55	94.40	91.60	68.68	53.88	19.80	23.89	26.44
20%	EViT	85.02	67.20	93.20	90.31	66.78	53.52	19.60	23.62	25.75
	ToMe	85.34	67.80	94.20	91.10	68.65	53.38	19.54	23.76	25.98
	STA-ViT	83.51	66.23	91.42	88.07	66.90	52.40	19.20	22.83	25.45
	TAPE (Ours)	85.90	68.44	95.20	92.55	69.87	54.08	19.87	24.03	26.43
30%	EViT	83.31	66.75	91.60	89.53	66.82	52.94	19.39	23.15	25.96
	ToMe	84.23	66.51	91.40	89.38	66.24	52.08	19.04	23.41	25.82
	STA-ViT	81.28	64.90	88.80	86.23	64.00	50.14	18.70	22.42	24.90
	TAPE (Ours)	86.27	68.58	94.20	91.41	68.18	53.12	19.72	23.66	26.31
40%	EViT	80.95	63.58	87.40	85.17	63.79	50.52	18.52	22.47	24.78
	ToMe	81.57	65.60	89.00	86.16	64.57	51.47	18.79	22.49	25.11
	STA-ViT	78.64	63.41	86.80	84.05	62.90	49.77	18.17	21.75	24.08
	TAPE (Ours)	82.71	65.86	90.20	86.85	66.08	51.61	18.94	22.69	25.02

F VISUALIZATION ON GENERATED VIDEOS

Due to the limited space in the main paper, we provide additional visualizations of generated videos in this section, comparing the baseline Hunyuan-Video model without pruning and our TAPE under a 20% token reduction setting. As shown in Figure 6, across diverse prompts, TAPE effectively preserves the semantic content and motion patterns specified by the text, while maintaining clear structures and visually detailed frames. Moreover, the videos produced by TAPE remain visually close to the unpruned baseline in both scene composition and temporal coherence.

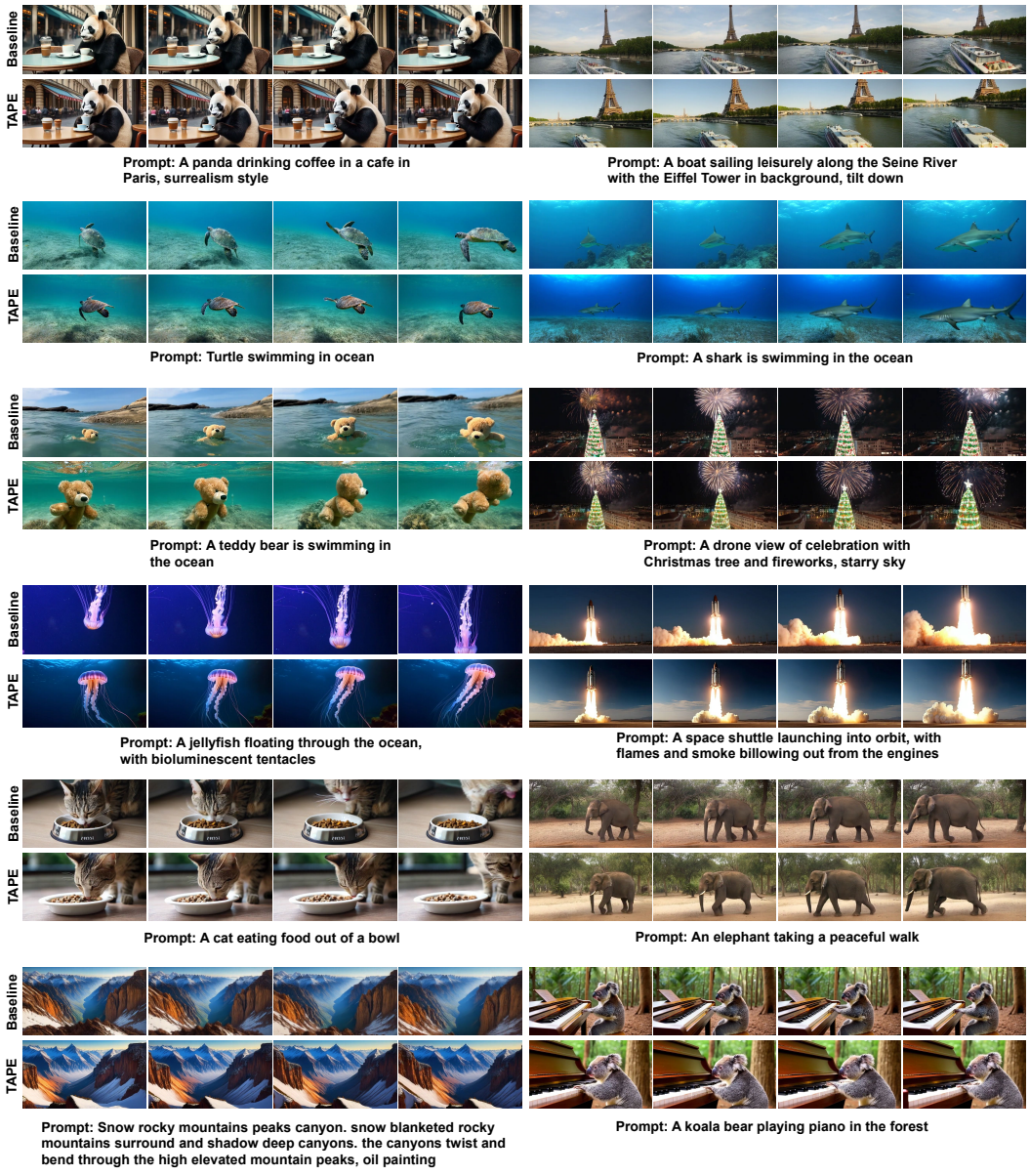


Figure 6: Additional visualizations of the generated videos.