

Contents

1	Introduction	2
2	Related Work	3
3	Dataset	4
4	Methodology	5
4.1	Preliminary	5
4.2	Frame correlation in attention module	6
4.3	Mask mechanism	6
4.4	Implementations	7
5	Experiment	7
5.1	Implementation details	8
5.2	Evaluation	8
5.3	Ablation Studies	10
6	Conclusion	10
A	Supplementary Materials Overview	17
B	Details of Dataset Construction	17
B.1	Method of Splitting and Shot Segmentation	17
B.2	Details of Stitching Phase	18
C	Statistic of Cine250K	18
D	Additional Details of Implementation	21
D.1	Visible-First-Frame Attention	21
D.2	Multi-prompt	21
D.3	Transition Point Selection	21
E	Additional Results	22
E.1	The Specific Details of the Attention Probabilities.	22
E.2	Impact of Mask Layers on Results	24
E.3	Qualitative Evaluation for Ablation Studies	26
E.4	User Study	28
E.5	Achieving Smoother Transitions through Soft Masking	28
E.6	Overlapping Ambiguous Shot Boundaries	29
F	Evaluation	30
F.1	Prompt Details	30
F.2	Metric details	30
G	Limitation	33
G.1	Failure Case	33
G.2	Future Work	33
H	Use of Large Language Models	33

A SUPPLEMENTARY MATERIALS OVERVIEW

The supplementary material provides additional experimental results, implementation details, and qualitative analysis. In addition to this PDF, the supplementary package also includes a project page in HTML format (`Cinematic_video.html`), which contains interactive visualizations, video demonstrations, and further results. To view the project page, please open `project_page/Cinematic_video.html` in a web browser. Furthermore, we also provide the source code, which is included in the `code/` directory, along with the training logs stored in `code/log/`.

B DETAILS OF DATASET CONSTRUCTION

Section 3 describes the Cine250K construction pipeline. In this section, we provide a detailed account of different stage.

B.1 METHOD OF SPLITTING AND SHOT SEGEMENTATION

During the Splitting stage, videos are initially segmented using PySceneDetect (Castellano, 2024). In the subsequent gradual-change removal stage, TransNetV2 (Soucek & Lokoc, 2024) is employed to detect and handle gradual transitions, yielding the final shot boundaries and associated shot labels. This pipeline is selected because Pyscenedetect operates exclusively on the CPU, offering greater processing efficiency than Transnetv2; however, for a pre-segmented video, Transnetv2 achieves higher shot segmentation accuracy and can handle gradual transitions more effectively than Pyscenedetect.

For Transnetv2, Figure 10 presents an example of detection. Transnetv2 provides *single-frame-prediction* and *all-frame-prediction*, which are designed to handle hard cuts and gradual transitions, respectively. The *single-frame-prediction* refers to the probability of an individual frame being a transition. In contrast, the *all-frame-prediction* predicts all frames involved in a transition, essentially estimating the probability of a frame being part of a gradual change. We apply threshold-based filtering to exclude frames with a high probability of being transition frames and obtain shot labels. Table 3 presents the threshold settings.



Figure 10: An example of transition detection by Transnetv2. The green line represents the result of *single-frame-prediction*, while the blue line represents the result of *all-frame-prediction*.

Table 3: Threshold settings for dataset construction.

	parameter	value
	Pyscenedetect splitting	27
	α	0.9
	β	0.7
	γ	0.8
Transnetv2	<i>single-frame-threshold</i>	0.45
Transnetv2	<i>all-frame-threshold</i>	0.50

To quantitatively compare accuracy on shot segmentation, we randomly select 200 videos covering multiple categories, each containing multiple shots. We then apply both Pyscenedetect and Transnetv2 for shot segmentation and compare the results manually. A correctly segmented video is

Table 4: Shot Segmentation Accuracy of different methods.

Method	Pyscenedetect	Transnetv2
Shot Segmentation Accuracy	65.50%	87.00%

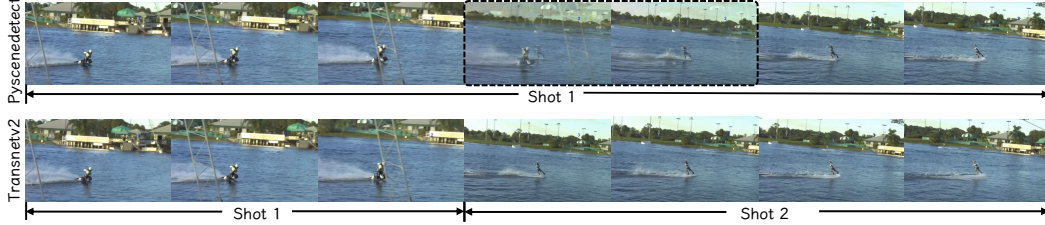


Figure 11: An example of shot segmentation by Pyscenedetect and Transnetv2. In the figure, Transnetv2, after the removing gradual changes step, accurately segments the two shots, while Pyscenedetect fails to identify the gradual changes. The transition frames caused by gradual changes are marked with dashed lines.

assigned 1; otherwise, 0. The shot segmentation accuracy is then calculated accordingly, as demonstrated in Table 4. A specific example is shown in Figure 11.

B.2 DETAILS OF STITCHING PHASE

During the Stitching stage, adjacent segments exhibiting high semantic similarity are merged to form a single video, thereby constructing multi-shot video collections. Specifically, we extract semantic features from the first and last frames of each segment using ImageBind (Girdhar et al., 2023) and quantify their similarity via Euclidean distance. For the i -th segment C^i , we denote the semantic features of its first and last frames as C_{first}^i and C_{end}^i , respectively, with their distance represented as $\text{dis}(C_{\text{first}}^i, C_{\text{end}}^i)$. Based on the distance, video segments are processed sequentially and either stitched or filtered according to the following criteria.

- For a segment C^i , if $\text{dis}(C_{\text{first}}^i, C_{\text{end}}^i) > \alpha$, the segment is filtered.
- For C^i , if C^{i-1} is absent (either non-existent or filtered), C^i is treated as a video’s beginning.
- For C^i , if $\text{dis}(C_{\text{end}}^{i-1}, C_{\text{first}}^i) < \beta$ and $\text{dis}(C_{\text{first}}^{i-1}, C_{\text{end}}^i) < \gamma$, then C^{i-1} and C^i are stitched to form a new segment.

Here, α restricts significant changes within a clip, β enables stitching of semantically similar transitions or originally continuous segments, and γ ensures consistency between the video’s beginning and end. After processing all segments sequentially, each resulting segment is considered a complete video, forming a preliminary dataset of videos with shot transitions.

C STATISTIC OF CINE250K

As we present in Section 3, Cine250K is a carefully curated multi-shot video dataset with detailed captions. This section presents its overall statistics. As shown in Figure 12, the average video duration is 10.75s, and the average caption length is 148.79. Most videos contain 2 to 3 shots. It is important to note that although duration and shot count filtering are applied during data processing, TransnetV2 (Soucek & Lokoc, 2024) is later utilized to remove gradual changes and re-identify shots. As a result, the final shot count and duration do not strictly adhere to the initial filtering criteria. After reidentification, 87.99% of the videos contain 2 to 5 shots. Figure 12 presents the shot distribution for videos with 1 to 10 shots, which represents 99.90% of all videos.

Regarding video categories, following Vimeo’s classification, the dataset is divided into 10 categories. Among them, Travel and Documentary have relatively higher proportions. Overall, the distribution of categories is fairly balanced.

Table 5: Comparison of Cine250K and other video-text datasets.

Dataset	#Videos	Avg video len	Avg text len	Multi-shot	shot label	Aesthetic	Resolution
InternVid (Wang et al., 2023b)	234M	11.7s	17.6	✗	-	High	720P
LVD-2M (Xiong et al., 2025)	2M	20.2s	88.7	✗	-	Medium	Diverse
OpenVid-1M (Nan et al., 2024)	1M	N/A	N/A	✗	-	High	Diverse
Panda70M (Chen et al., 2024b)	70.8M	8.5s	13.2	✓	✗	Medium	720P
LLaVA-Video-178K (Zhang et al., 2024)	178K	~ 40.4s	~ 300	✓	✗	High	Diverse
Shot2Story20K (Han et al., 2023)	20K	16s	201.8	✓	✓	Medium	720P
Shot2Story134K	134K	N/A	N/A	✓	✓	Medium	720P
Cine250K (Ours)	250K	10.75s	148.79	✓	✓	High	720P

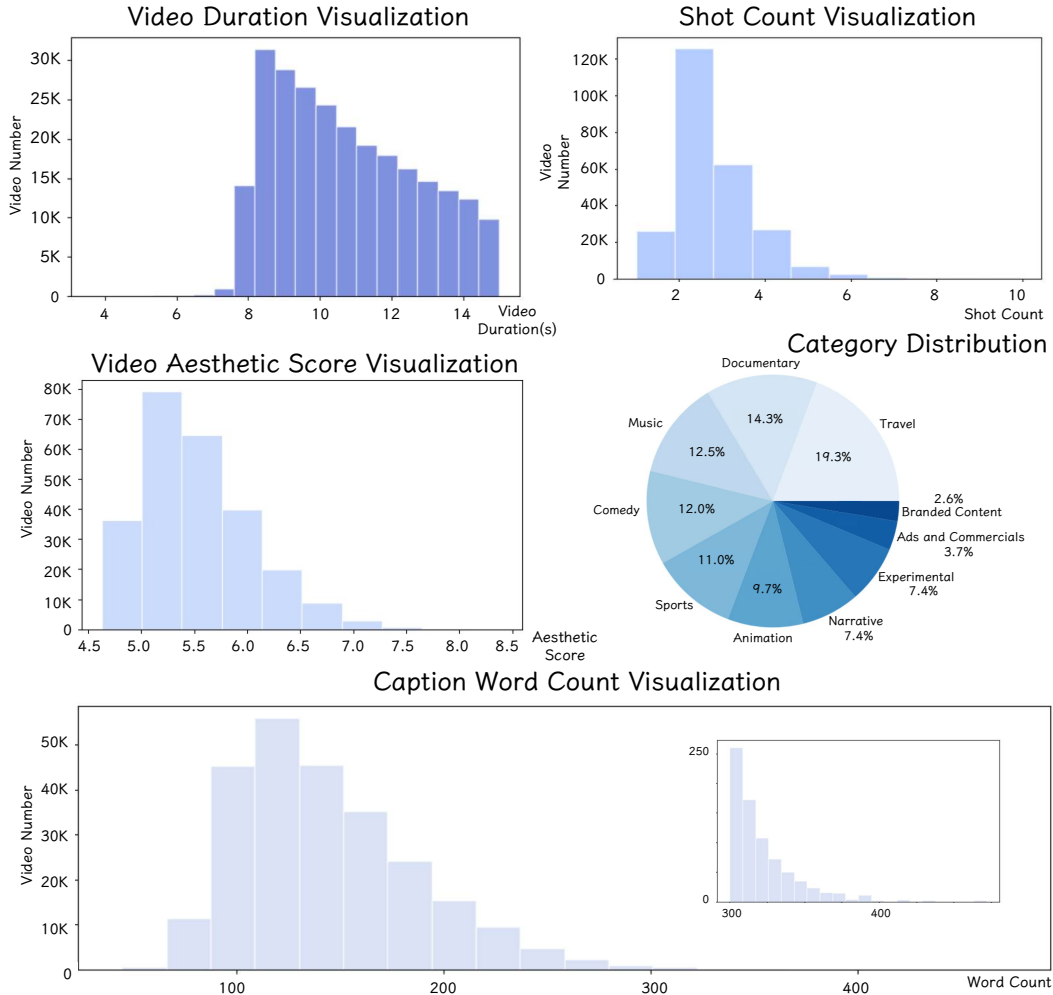


Figure 12: Statistics of Cine250K. To facilitate observation, the figure presents the shot distribution for videos containing 1 to 10 shots. The caption word count distribution only considers data with fewer than 500 words.



Figure 13: Example video-text pairs in Cine250K.

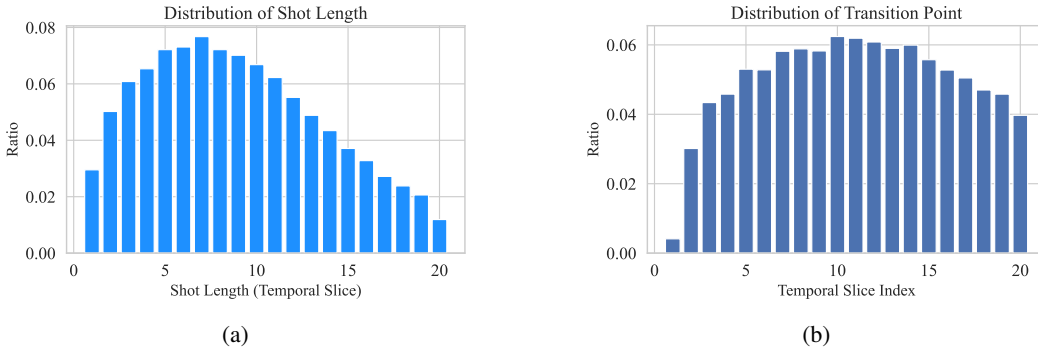


Figure 14: Distribution of shot length and transition points in Cine250K as the training dataset. Given the temporal compression of the VAE encoder, temporal slices are used as the unit.

In Figure 13, we select several example video-text pairs to demonstrate the specific characteristics of the cinematic multi-shot sequences in the dataset and the style of captions. In Table 5, we compare Cine250K with other datasets. As a multi-shot video dataset with detailed shot labels, the high-quality content of Cine250K can significantly facilitate research and exploration in multi-shot video generation.

Additionally, when Cine250K is used as the training dataset for CineTrans, the distribution of transition point positions and shot lengths is relatively uniform (Figure 14), which enables CineTrans to achieve precise frame-level control without timestamp jitter and ensure stability when varying shot lengths are used as conditions.

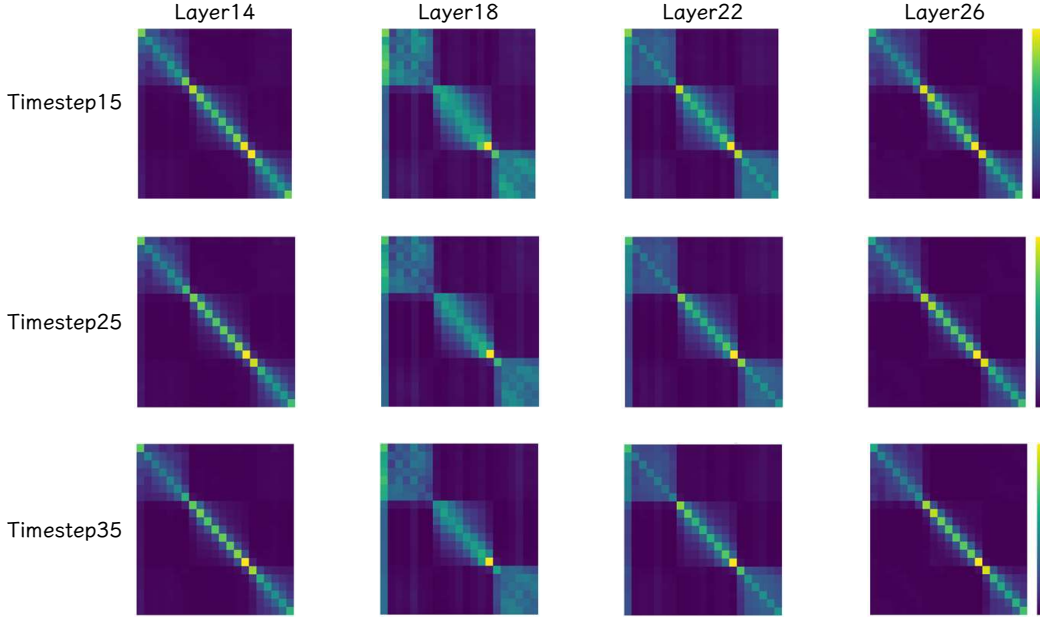


Figure 15: Visualization of the temporal-domain attention maps of Wan2.1 when generating multi-shot videos. Certain layers exhibit a pronounced focus on the first temporal slice, which motivates our Visible-First-Frame Attention mechanism.

D ADDITIONAL DETAILS OF IMPLEMENTATION

D.1 VISIBLE-FIRST-FRAME ATTENTION

In Section 4.4, we introduce the implementation detail dubbed Visible-First-Frame Attention, which serves to further stabilize the mask mechanism, particularly within DiT architectures, as demonstrated in Figure 8a. This mechanism is also motivated by our analysis of DiT’s attention maps during multi-shot video generation. As illustrated in Figure 15, in certain layers all visual tokens assign high attention probabilities to tokens originating from the first frame (owing to temporal compression in the VAE, this in fact corresponds to the first latent temporal slice), suggesting that the initial frame assumes a special function in the diffusion model’s denoising process. In light of this observation, we modify our mask matrix so that its first column is set entirely to zero, thereby effecting the Visible-First-Frame Attention mechanism, which can be formulated as:

$$\mathcal{M}_{ij} = \begin{cases} 0 & \text{if } j = 1 \text{ or } i, j \in \text{same shot} \\ -\infty & \text{if } j \neq 1 \text{ or } i, j \notin \text{same shot} \end{cases} \quad (7)$$

D.2 MULTI-PROMPT

In Section 5.2, we note that some methods employ a multi-prompt strategy at inference, i.e., each shot is prompted by its own text description. CineTrans-DiT also supports this capability. In addition to the primary mask matrix between video tokens, we introduce an additional mask between text and video tokens, following Qi et al. (2025), to enable precise semantic control over each shot. It is worth noting that this mask is applied only to the attention layers governing text–video token interactions, and is distinct from our proposed mask mechanism, which operates on video–video token interactions.

D.3 TRANSITION POINT SELECTION

In Section 4.3, we introduce the Mask Mechanism that achieves temporal control of shot transitions. During evaluation, shot transition points should be predefined in advance to generate multi-shot videos. To address this, we predefine several suitable shot transition points for different shot counts

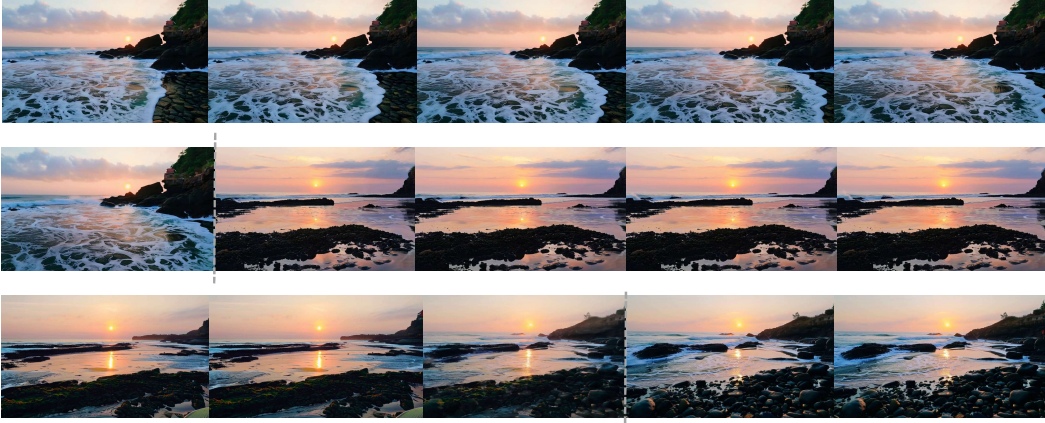


Figure 16: Multi-shot videos with shorter length for last two shots. It can be observed that the visual quality remains intact even when the shot lengths are relatively extreme.

and randomly select them during inference. However, as mentioned in Appendix C, our training set adequately supports shot transitions at various positions, so random sampling of transition points does not lead to issues such as visual collapse (Figure 16). Nevertheless, for optimal visual performance, including action completeness and overall video quality, uniform shot lengths are preferred, as they avoid extremely short shots.

E ADDITIONAL RESULTS

E.1 THE SPECIFIC DETAILS OF THE ATTENTION PROBABILITIES.

In Section 4.2, we explore the frame correlations in the case of cinematic multi-shot video generation in diffusion models, where attention modules in certain layers exhibit a block-diagonal structure, i.e., strong correlations for intra-shot frames and weak correlations for inter-shot frames. In this section, we will discuss the specific details of the attention maps across different architectures, layers, and timesteps.

As for the architecture, we investigate both the temporal-spatial-decoupled framework and the full attention framework. The temporal-spatial-decoupled framework applies the temporal attention module directly to the time sequences, where tokens at different spatial locations do not interact with each other. In contrast, the full attention framework operates on all tokens of the video, allowing correlations across both temporal and spatial dimensions simultaneously. To focus on frame correlations, we group and average the tokens in the full attention framework by the frames before conducting further analysis. In Figure 17 and Figure 18, we use Wang et al. (2024b) and Kong et al. (2024) as representatives of different architectures and present the attention maps across different layers and timesteps in both qualitative and quantitative manners.

As timesteps increase, the discrepancy between intra-shot and inter-shot attention probabilities grows in the temporal-spatial-decoupled framework, whereas it remains stable in the full attention framework. For different layers, the temporal-spatial-decoupled framework exhibits noticeable differences across most layers, whereas the full attention framework shows disparity primarily in the earlier layers. In summary, both frameworks demonstrate significant differences between intra-shot and inter-shot attention probabilities. This finding further substantiates the prevalence of strong intra-shot and weak inter-shot correlations, supporting the proposed approach.

Additionally, we present the inter-shot/intra-shot attention probability ratio for different models and layers, as well as the Pearson correlation between this ratio and the distance from transition points (i.e., whether the ratio increases as partition points approach true transitions), with results recorded in Table 6. It can be observed that the inter-shot/intra-shot ratio tends to increase as partition points approach the transition points across different models, further supporting the general presence of the block-diagonal pattern in diffusion models.

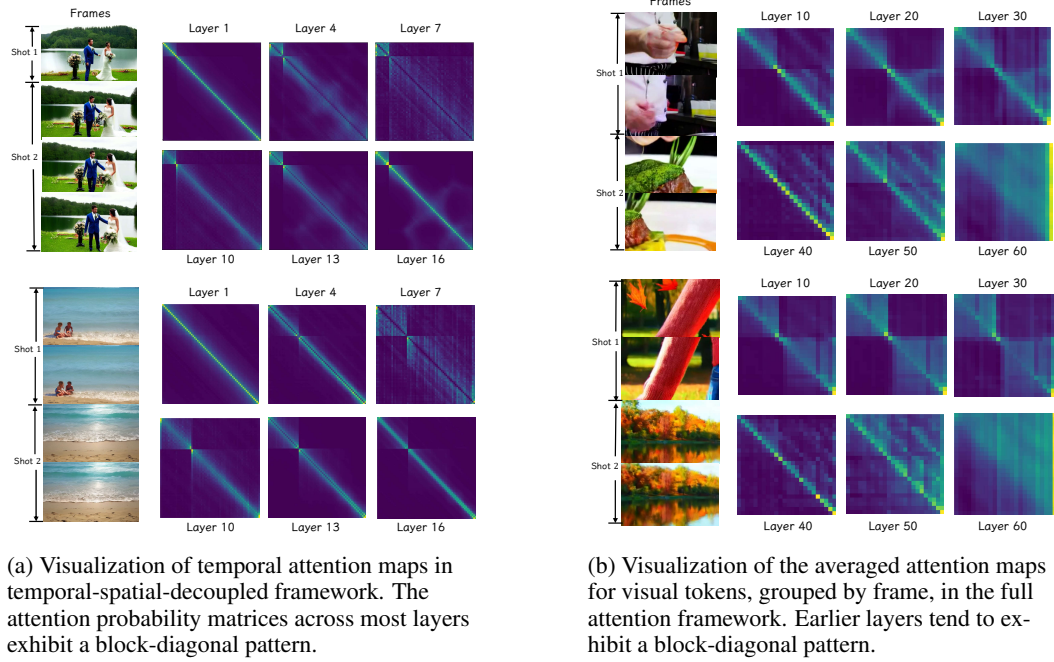


Figure 17: The visualization of attention maps between visual tokens from different frames for individual case of multi-shot video generation. Both in the temporal-spatial-decoupled framework and full attention framework, diffusion models exhibit strong intra-shot attention and weak inter-shot attention in certain layers.

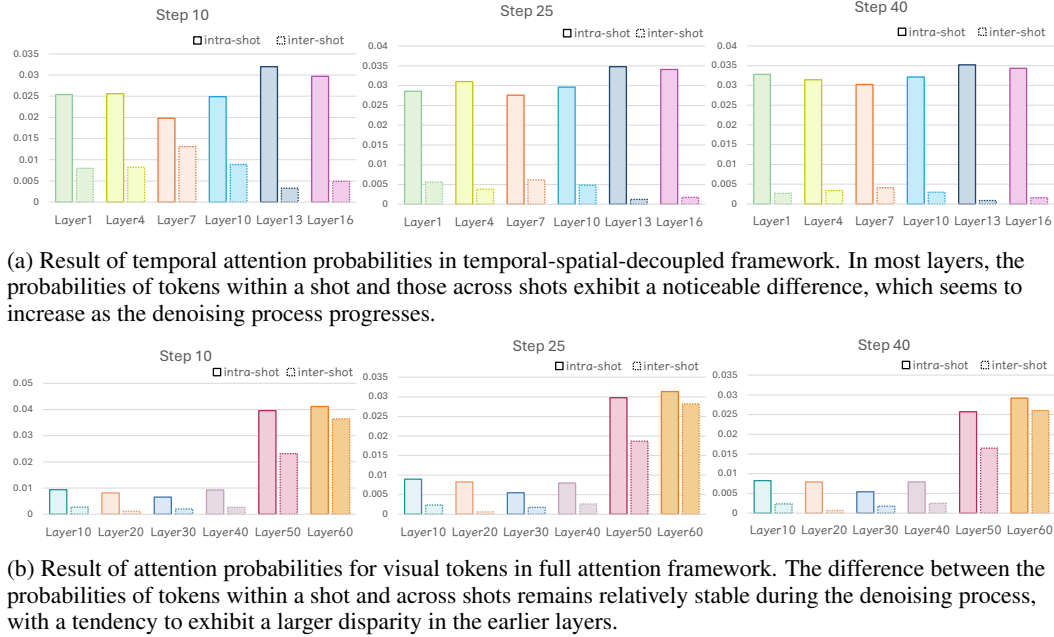


Figure 18: The average attention probability for intra-shot and inter-shot across diffusion layers and denoising steps. In both framework, the average probability within shots is obvious greater than that between shots for certain layers.

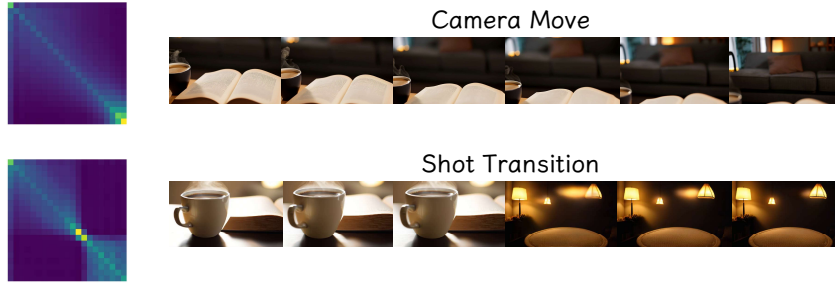


Figure 19: Comparison of attention maps between two cases with identical conditions but different seeds: one exhibiting a shot transition and the other only showing a camera movement.

Table 6: Inter-shot/Intra-shot attention probability ratio and Pearson correlation for different models and layers.

Model	Intra-shot/Inter-shot attention probability ratio	Pearson correlation
Wan2.2 (front layer 33%)	5.6792	0.9267, $p < 0.0001$
Wan2.2 (middle layer 33%)	7.4204	0.9233, $p < 0.0001$
Wan2.2 (late layer 33%)	6.9483	0.8496, $p < 0.0001$
Hunyuan (front layer 33%)	4.8514	0.6571, $p < 0.0001$
Hunyuan (middle layer 33%)	3.4478	0.7507, $p < 0.0001$
Hunyuan (late layer 33%)	1.5194	0.4969, $p < 0.0001$
Wan2.1 (front layer 33%)	8.3925	0.7393, $p < 0.0001$
Wan2.1 (middle layer 33%)	14.1942	0.7319, $p < 0.0001$
Wan2.1 (late layer 33%)	14.6686	0.6159, $p < 0.0001$

Finally, we investigate the potential cause of this pattern by comparing two cases with identical conditions but different seeds: one exhibiting a shot transition and the other only a camera movement (Figure 19). Attention map analysis in Table 7 shows that the case with a shot transition exhibits a stronger block-diagonal pattern, further proving that this attention pattern is strongly correlated with the diffusion model’s understanding of shot transitions.

E.2 IMPACT OF MASK LAYERS ON RESULTS

For CineTrans-Unet and CineTrans-DiT, the mask mechanism employs different masking layers, a design choice motivated by observations of the attention maps when diffusion models generate multi-shot videos. In this section, we examine the effects of applying masks to different layers on the quality of the generated videos, thereby demonstrating the rationale behind our masking strategy.

As shown in Figure 20a, CineTrans-Unet exhibits noticeable visual distortions when the mask is applied to all layers or only to the early layers. In severe cases, some parts of the video become indistinct or heavily degraded. When no mask is applied or when the mask is applied only to the middle layers, cinematic transitions do not emerge clearly. Applying the mask to the later layers

Table 7: Inter-shot/Intra-shot attention probability ratio and Pearson correlation for two cases (with and without transition).

Model	Intra-shot/Inter-shot attention probability ratio	Pearson correlation
Case with shot transition		
Wan2.1 (front layer 33%)	8.3925	0.7393, $p < 0.0001$
Wan2.1 (middle layer 33%)	14.1942	0.7319, $p < 0.0001$
Wan2.1 (late layer 33%)	14.6686	0.6159, $p < 0.0001$
Case without shot transition		
Wan2.1 (front layer 33%)	6.7575	0.3462, $p < 0.001$
Wan2.1 (middle layer 33%)	4.9148	0.3378, $p < 0.001$
Wan2.1 (late layer 33%)	6.7575	0.3462, $p < 0.001$

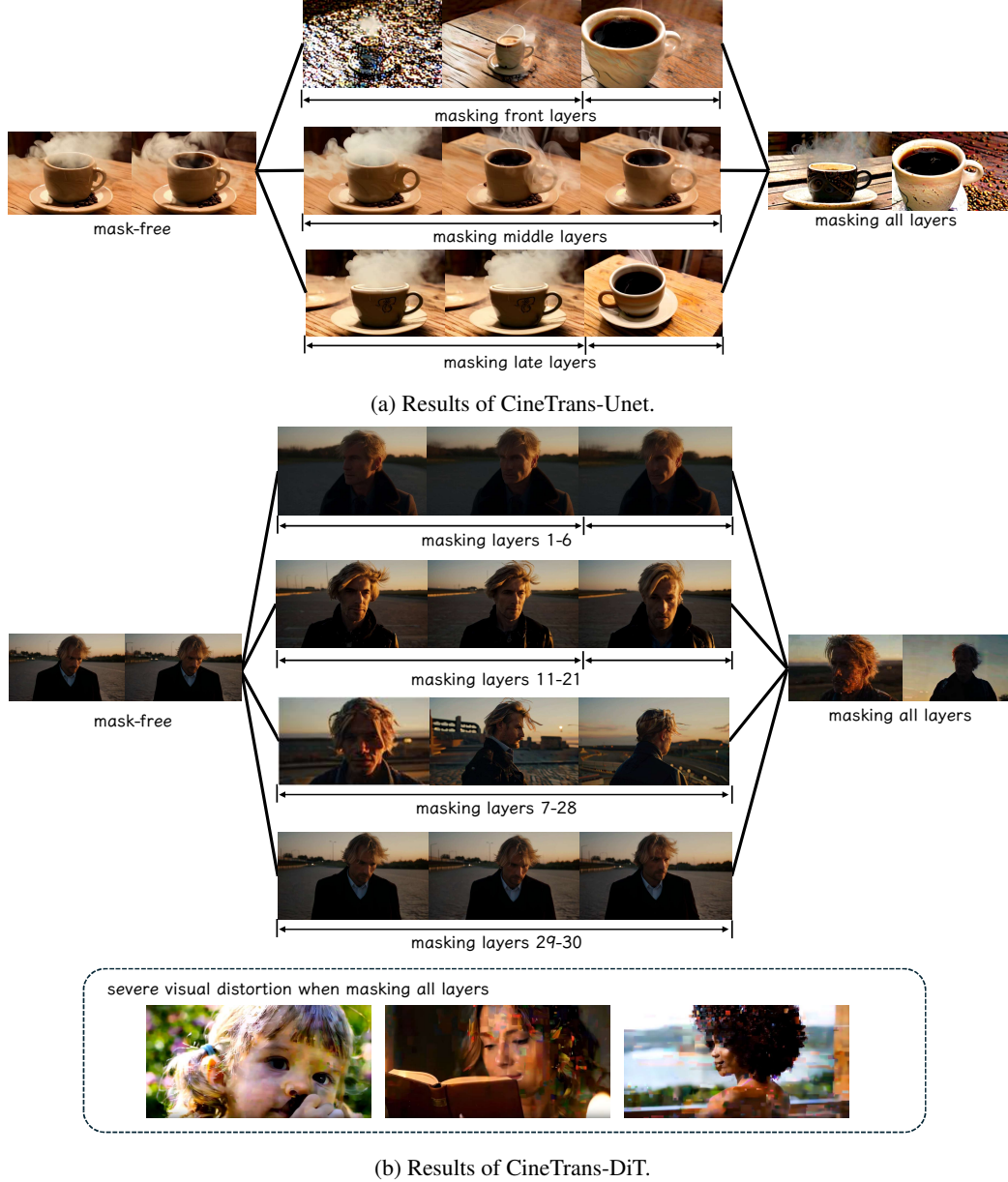


Figure 20: The results of different mask strategies. Applying the mask to the later layers of the spatial-temporal-decoupled architecture (CineTrans-Unet) and the middle layers of the full attention architecture (CineTrans-DiT) is considered more effective. For the full attention architecture, applying the mask to all layers leads to severe visual distortion, as illustrated in the figure.

Table 8: Quantitative results for masking different layers in new different frameworks. The best are in **bold**.

Method	Transition Control Score↑	Inter-shot Consistency				Intra-shot Consistency		Aesthetic Quality↑	Semantic Consistency↑
		Semantic		Visual		Subject↑	Background↑		
		Score↑	Gap↓	Score↑	Gap↓				
Ablation for VideoCrafter2 (Unet)									
Mask front 1/2 layers	0.0889	0.8313	0.3630	0.8338	0.3603	0.9594	0.9723	0.6287	0.2205
Mask late 1/2 layers	0.1889	0.7739	0.1980	0.7766	0.3416	0.9728	0.9693	0.6258	0.2213
Ablation for Wan2.2 (DiT)									
Mask front 1/2 layers	0.3811	0.9425	0.5976	0.9380	0.5205	0.9570	0.9624	0.6298	0.2073
Mask middle 1/2 layers	0.5982	0.6752	0.2084	0.7234	0.1458	0.9474	0.9636	0.6395	0.2140
Mask late 1/2 layers	0.0085	-	-	-	-	0.9466	0.9646	0.6305	0.1943



Figure 21: Results of masking different layers in new different frameworks.

enables effective control over transitions without significantly affecting visual quality, suggesting it as the optimal strategy. This suggests that, in the spatial-temporal-decoupled architecture, correlations in the earlier layers primarily influence visual quality, while the later layers may regulate the consistency between adjacent frames. Therefore, applying the mask mechanism to the last six layers is demonstrated to be effective.

As shown in Figure 20b, the proportion of layers requiring masking in CineTrans-DiT is larger than in CineTrans-Unet, and both the early and late layers need to retain fully visible attention. Masking all layers to this architecture may lead to low inter-shot consistency and severe visual degradation at the transitions. Conversely, masking only the earlier or later layers fails to effectively guide the transitions. Moreover, when fewer layers are masked, the inter-shot differences are reduced, which deviates from the convention of multi-shot video. These observations suggest that the current strategy of masking the middle layers achieves the optimal balance, enabling precise control over transitions while preserving a reasonable degree of consistency.

To validate the generalizability of our heuristic mask-layer selection, we apply the same approach to other architectures: masking the later half of layers in U-Net and the middle half of layers in DiT. We use VideoCrafter2 [Chen et al. \(2024a\)](#) and Wan2.2 [Wan et al. \(2025\)](#) as base models, with results presented in Table 8. Due to VideoCrafter2’s relatively limited duration and representational capacity, the overall Transition Control Scores are relatively low. However, masking later layers still leads to clearer shot-transition behavior. Figure 21 shows that selecting the appropriate masking layer can indeed improve multi-shot video generation performance.

E.3 QUALITATIVE EVALUATION FOR ABLATION STUDIES

The quantitative results of the ablation studies are presented in Section 5.3, and this section provides a supplementary qualitative analysis.

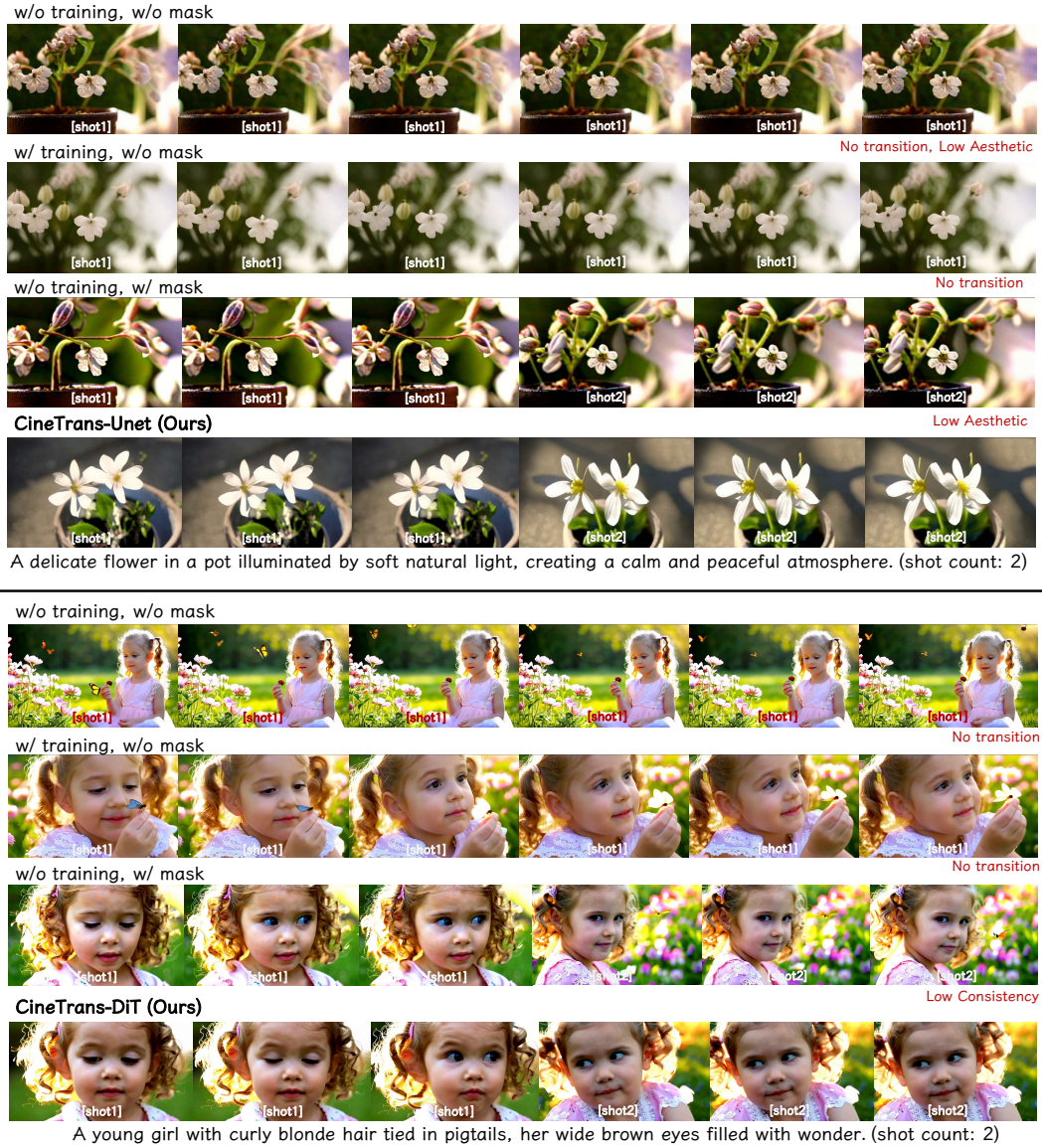


Figure 22: Qualitative results of ablation studies.

As shown in Figure 22, CineTrans can stably control the generation of transitions. In contrast, models without the mask mechanism, whether finetuned or not, are almost incapable of generating cinematic transitions. Direct application of the mask mechanism without further finetuning, i.e., training-free results, does indeed generate transitions. However, further training can help the model produce shot transitions that better align with the film editing style, exhibiting higher aesthetic quality and improved consistency.

E.4 USER STUDY

To complement the experimental results, we report the findings of our user study in this section. Specifically, we select 20 prompts for user evaluation, where participants rate each result on a scale from 1 to 5. We evaluate different methods from the perspective of Transition Control and Overall Consistency, with the MOS results presented in Table 9. It can be observed that CineTrans also demonstrates strong performance in terms of user preference.

Table 9: Results of User Study

Model	Transition Control	Consistency
CineTrans-DiT (ours)	4.60 $\pm$ 0.50	4.15 $\pm$ 0.75
CineTrans-Unet (ours)	4.75 $\pm$ 0.44	4.10 $\pm$ 0.72
StoryDiffusion+CogVideoX12V	-	3.95 $\pm$ 0.76
HunyuanVideo+Cinematron	3.60 $\pm$ 1.95	3.80 $\pm$ 0.68
HunyuanVideo	3.45 $\pm$ 1.88	3.50 $\pm$ 0.76
CogVideoX	2.50 $\pm$ 1.24	3.05 $\pm$ 0.60
Wanx2.1-T2V-turbo	3.25 $\pm$ 1.77	3.45 $\pm$ 0.69

E.5 ACHIEVING SMOOTHER TRANSITIONS THROUGH SOFT MASKING

This section investigates soft masking strategies applied to the hard mask mechanism proposed in 4.3, with the aim of assessing whether soft masking can enable smoother shot transitions or higher consistency. Specifically, we explore strategies including time-dependent penalty and timestep-dependent penalty. The following will present the details of these strategies along with the corresponding experimental results.

Time-Dependent Penalty. In the hard masking scheme, the mask matrix is initially set to be fully invisible, allowing tokens within each shot to interact for multi-shot video generation. The time-dependent penalty allows token interactions near shot boundaries. Specifically, we initialize the mask matrix using a Gaussian decay based on the distance between frames:

$$\mathcal{M}_{ij} = e^{-\frac{(i-j)^2}{2\sigma^2}}. \quad (8)$$

This is then mapped to the range $[0, -\infty)$, with all tokens within each shot being fully visible subsequently, thus forming a soft masking strategy around shot boundaries. The smoothness of the transition is controlled by the parameter σ . The evaluation results are presented in Table 10 and Figure 23, where L denotes the sequence length. As the soft mask approaches full visibility, the transition control effect weakens until no shot transition occurs. With the appropriate hyperparameter settings (e.g., $\sigma = L/12$), the visual break at the transition boundary becomes less obvious. However, excessively large σ values lead to minimal difference across shot compositions, reducing the content diversity in the multi-shot video and resulting in a more uniform appearance, which explains the decline in semantic consistency.

Timestep-Dependent Penalty. In contrast, the Diffusion-Timestep-Dependent Penalty focuses on the timestep of the denoising process, making the invisible region of the mask partially visible during the early denoising steps and gradually approaching full invisibility as the process progresses. Compared to the Time-Dependent Penalty, which focuses on shot boundaries, this soft masking enhances interactions between all inter-shot tokens. As shown in Figure 24, the increased token correlations result in a significant reduction in compositional differences between shots, leaving only a residual boundary transition effect. This comes at the cost of multi-shot content diversity, as reflected by the substantial drop in the Transition Control Score in Table 11. Therefore, the Diffusion-Timestep-Dependent Penalty tends to strengthen inter-shot consistency, while its effect on smoother transitions is less pronounced than that of the time-dependent penalty.

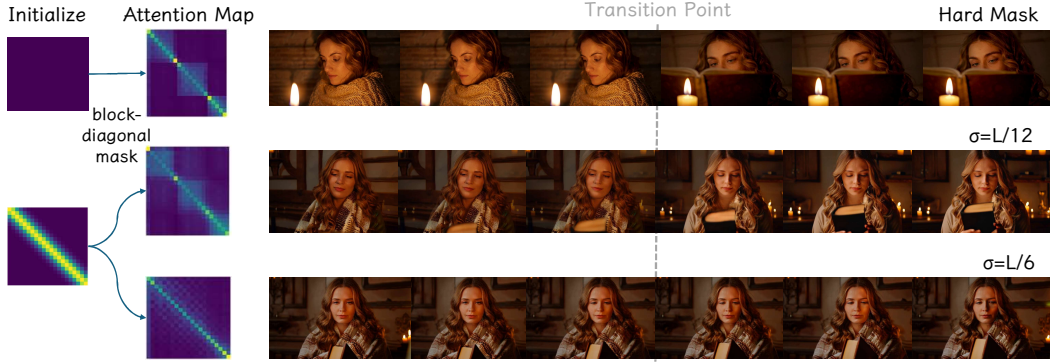


Figure 23: An illustration of the Time-Dependent Penalty.

Table 10: Quantitative results for Time-Dependent Penalty. The best are in **bold**. Smaller σ values approach the hard mask.

Method	Transition Control Score $\uparrow$	Inter-shot Consistency				Intra-shot Consistency		Aesthetic Quality $\uparrow$	Semantic Consistency $\uparrow$
		Semantic		Visual		Subject $\uparrow$	Background $\uparrow$		
		Score $\uparrow$	Gap $\downarrow$	Score $\uparrow$	Gap $\downarrow$				
Ablation for CineTrans-DiT (training-free)									
$\sigma = L/4$	0.0220	0.8685	0.4261	0.8231	0.2499	0.9629	0.9750	0.6513	0.2025
$\sigma = L/6$	0.0700	0.8238	0.2411	0.8248	0.2431	0.9445	0.9606	0.6514	0.2073
$\sigma = L/12$	0.4036	0.7987	0.2005	0.8186	0.2139	0.9400	0.9544	0.6562	0.2086
Hard Mask	0.6564	0.7838	0.1772	0.7844	0.1943	0.9618	0.9746	0.6556	0.2093
Ablation for CineTrans-DiT (trained)									
$\sigma = L/4$	0	-	-	-	-	0.9541	0.9660	0.6453	0.2004
$\sigma = L/6$	0.0455	0.7944	0.1885	0.8377	0.3074	0.9424	0.9598	0.6455	0.2105
$\sigma = L/12$	0.5193	0.7902	0.1796	0.8145	0.2575	0.9519	0.9659	0.6489	0.2095
Hard Mask	0.7003	0.7858	0.1552	0.7874	0.1901	0.9673	0.9775	0.6508	0.2109

E.6 OVERLAPPING AMBIGUOUS SHOT BOUNDARIES

The results in Table 1 demonstrate that our model maintains precise control over shot transitions at specified transition points using the mask mechanism. In this section, we examine the model’s performance under overlapping and ambiguous shot boundaries. Specifically, we introduce fluctuations of $\pm t$ frames (temporal slices) around the fixed transition points during inference, which generates overlapping and ambiguous boundaries in the denoising process. Quantitative results are

Table 11: Quantitative results for Timestep-Dependent Penalty. The best are in **bold**. Smaller t values approach the hard mask.

Method	Transition Control Score↑	Inter-shot Consistency			
		Semantic		Visual	
		Score↑	Gap↓	Score↑	Gap↓
Ablation for CineTrans-DiT (training-free)					
$t = 1.0$	0	-	-	-	-
$t = 0.5$	0	-	-	-	-
Hard Mask	0.6564	0.7838	0.1772	0.7844	0.1943
Ablation for CineTrans-DiT (trained)					
$t = 1.0$	0	-	-	-	-
$t = 0.5$	0.0741	0.7864	0.2211	0.8029	0.3655
Hard Mask	0.7003	0.7858	0.1552	0.7874	0.1901

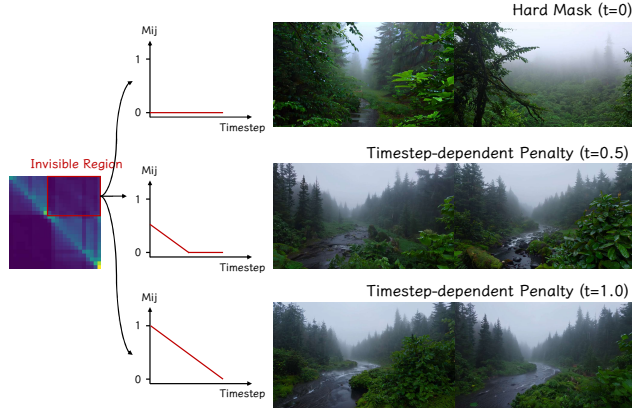


Figure 24: An illustration of the Timestep-Dependent Penalty.

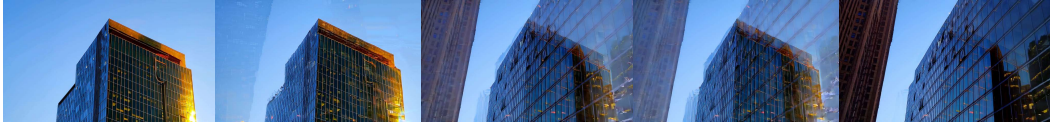


Figure 25: A Result of Overlapping Ambiguous Shot Boundary.

shown in Table 12 and Figure 25. It can be observed that, even under overlapping and ambiguous shot boundaries, the model retains a reasonable Transition Control Score and produces effects resembling fade-in/fade-out transitions to some extent. These results confirm the robustness of the mask mechanism, demonstrating that the model retains effective control even in the presence of ambiguous or overlapping shot boundaries.

F EVALUATION

F.1 PROMPT DETAILS

In Section 5.2, we present a set of prompts generated by GPT-4o (Achiam et al., 2023) to evaluate the performance of multi-shot video generation. Figure 26a provides the prompt used to guide GPT.

Given the long inference time of video generation models, we design 100 prompts for evaluation, covering multiple categories. For these prompts, the transition guidance can be divided into two types. One type explicitly specifies a significant semantic change, such as prompts that include phrases like *The video transition to*, clearly indicating shot changes. The other type implicitly guides the shot transition, where the prompt does not explicitly differentiate content changes between shots, but rather provides an overall description of the video. This type of prompt design accounts for the fact that some cinematic multi-shot videos only involve switching camera angles without significant semantic change. The multi-shot guidance in such cases primarily relies on the prompt’s opening phrase: *The multi-shot video consists of {num_shot} shots*. Figure 26b presents examples of these two types of prompts, and Figure 26c shows the word cloud distribution for this set of prompts. Table 13 outlines the categories of the prompts. It is worth noting that, for methods requiring multi-prompt inference, we employ GPT-4o to expand the general prompt into shot-specific captions according to the designated shot count.

F.2 METRIC DETAILS

In Section 5.2, we establish evaluation metrics from three aspects: transition control, temporal consistency, and overall video quality. This section will provide a detailed explanation of the specific definitions of these metrics.

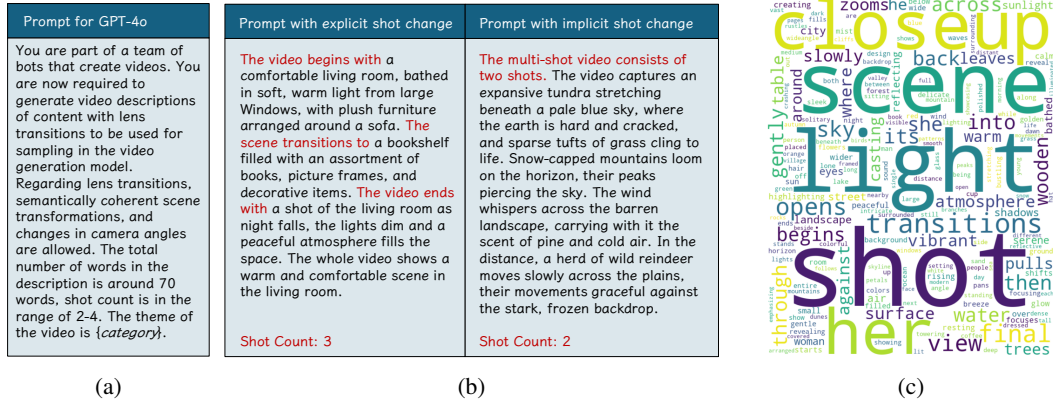


Figure 26: The details of the prompts used for evaluation. (a) Prompt for GPT-4o. The designated video theme vary. (b) Prompt examples for evaluation. (c) Word cloud of prompts for evaluation.

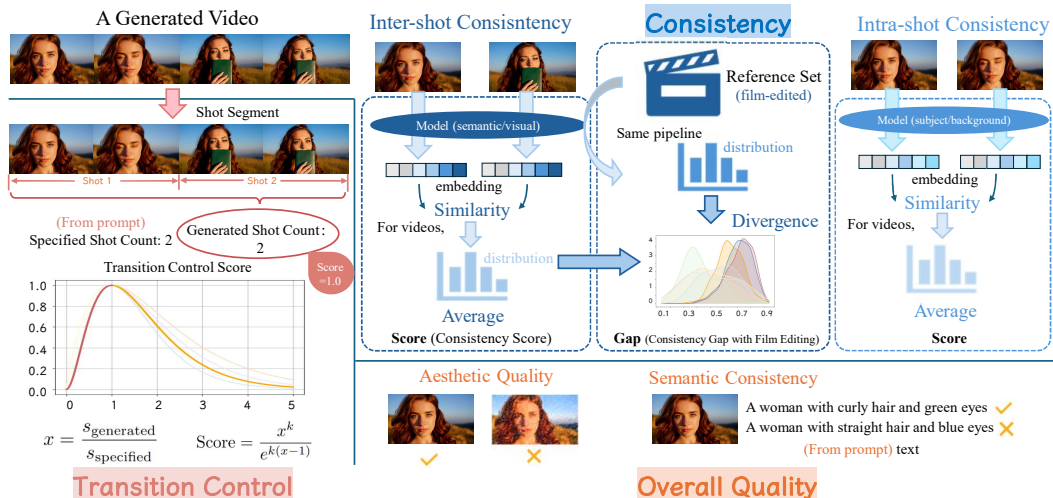


Figure 27: Overview of the metric design. We devise evaluation measures along three complementary dimensions: transition control, temporal consistency, and overall video quality.

Table 12: Quantitative results for Overlapping Ambiguous Shot Boundaries. The best are in **bold**. Smaller t values approach the fixed transition points.

Method	Transition Control Score $\uparrow$	Inter-shot Consistency			
		Semantic		Visual	
		Score $\uparrow$	Gap $\downarrow$	Score $\uparrow$	Gap $\downarrow$
Ablation for CineTrans-DiT (training-free)					
No Mask	0.2051	0.5924	0.3421	0.5274	0.3574
$t = 2$	0.2698	0.7227	0.1918	0.8109	0.2612
$t = 1$	0.4123	0.7344	0.1846	0.8044	0.2299
Fixed Transition Points	0.6564	0.7838	0.1772	0.7844	0.1943
Ablation for CineTrans-DiT (trained)					
No Mask	0.2112	0.6532	0.3422	0.6087	0.3312
$t = 2$	0.4240	0.7526	0.1976	0.8311	0.3082
$t = 1$	0.5461	0.7302	0.1837	0.8245	0.2764
Fixed Transition Points	0.7003	0.7858	0.1552	0.7874	0.1901

Table 13: Details of prompt categories for evaluation.

category	count
scenario	34
architecture	10
human	25
object	31

Transition Control. For transition control, we define the Transition Control Score to measure whether the shot count in the generated video aligns with the specified count, as formulated in Equation 5.

Figure 27 visualizes the calculation method for the Transition Control subpanel. In practice, when $x < 1$, k is set to 2, and when $x \geq 1$, k is set to 1.6. Given that the prompts used for evaluation all specify multiple transitions, the score is set to 0 when the generated video consists of a single shot. When the shot count in the generated video matches the specified value, the score is recorded as 1. More generally, the score is determined based on the absolute difference from the specified value.

Temporal consistency. In terms of temporal consistency, we consider both intra-shot consistency and inter-shot consistency. Intra-shot consistency treats each shot as a separate video and calculates the metric between adjacent frames using a method similar to that in VBench (Huang et al., 2024). The focus of this section is on inter-shot consistency. As for frame extraction, we use the middle frame of each shot for calculation. However, if the video does not generate multiple shots, inter-shot consistency cannot be evaluated. As shown in Table 1, the original LaVie (Wang et al., 2024b) lacks the ability to generate transitions, and therefore its inter-shot consistency metric does not have a corresponding value.

Regarding the metric definition, inter-shot consistency cannot directly serve as the final evaluation metric. In multi-shot video generation, the goal is to ensure that the generated video aligns with the editing style of film-edited videos. High consistency would imply pixel-level similarity, which may contradict the multi-shot nature of real video editing. To address this, we extract 1000 film-edited videos as a validation dataset and compute their inter-shot consistency as a reference set. The final metric is then determined by the Jensen-Shannon Distance (JSD) between the inter-shot consistency of the generated video and that of film-edited videos, as shown in Equation 6. A lower JSD indicates a closer alignment with the reference distribution.

Videos whose inter-shot consistency distribution aligns with film editing styles are considered to exhibit higher inter-shot consistency performance, while those deviating from film editing practices are regarded as having lower performance. This defines the evaluation metric based on inter-shot consistency.



Figure 28: Failure case with similar composition consistency, which probably results from insufficient training.

G LIMITATION

G.1 FAILURE CASE

The role of the mask mechanism in controlling the occurrence of transitions remains relatively stable. Nevertheless, in occasional cases (particularly under the training-free setting), the compositions of different shots may become overly similar, as illustrated in Figure 28. Although such content changes are perceptible to the human eye, the shot segmentation model does not recognize them as multi-shot videos due to the high compositional similarity, and the resulting visual experience also deviates from the film editing style. These cases can be attributed to insufficient training or the absence of such data types in the training set, which prevents the model from generating videos of the same scene with varied compositions. Potential improvements may come from expanding the dataset or further training.

G.2 FUTURE WORK

Although CineTrans has achieved promising results in cinematic multi-shot video generation using the mask mechanism, there are still limitations to be addressed, along with several promising directions for future research.

- While the mask mechanism enables control over the occurrence of shots, the specification of camera viewpoints could be made more controllable, for example, enabling changes in shooting perspective within a scene that are more consistent with the film production pipeline. Therefore, achieving higher content consistency and more precise control over camera viewpoints will be a key future direction, potentially requiring the incorporation of 3D information as prior knowledge.
- Multi-shot videos could also be extended to greater lengths. Incorporating auto-regressive generation into the video generation pipeline represents a promising approach for producing longer multi-shot videos.
- Fine-grained consistency presents an area for further improvement. While our method demonstrates strong overall consistency, achieving fine-grained alignment across shots remains challenging due to the intricate spatial understanding required. This limitation is primarily due to the lack of detailed background and scene context annotations in the training data. Future work will focus on incorporating more granular annotations and leveraging advanced model designs to address this challenge.

H USE OF LARGE LANGUAGE MODELS

We clarify the involvement of large language models (LLMs) in the preparation of this work. LLMs are not used for research design, methodological decisions, or experimental analysis. Their use is limited to two aspects: (i) improving the clarity and readability of the manuscript through language editing, and (ii) generating evaluation prompts during the assessment phase, as explicitly noted in Section 5.2. All substantive research contributions, including the conception of the problem, model design, implementation, and analysis, are conducted entirely by the authors.