

# ShotDirector: Directorially Controllable Multi-Shot Video Generation with Cinematographic Transitions

## Supplementary Material

### A. Details of Dataset Construction

In Sec. 3.1, we introduce ShotWeaver40K, a high-quality dataset with detailed shot transition annotations. This section provides additional details of the dataset construction pipeline to further support clarity and reproducibility.

**Raw Video Source.** To capture rich cinematography language and maintain narrative flow, we collect 16K full-length films as our raw video source. This large corpus provides shot transitions with strong semantic coherence.

**Segmentation and Stitching.** To obtain initial multi-shot clips that exhibit both semantic and visual similarity, we perform segmentation and similar-segment stitching on the raw videos. Specifically, we first segment each video using TransNetV2 [38], and extract image features for adjacent shots using ImageBind [12]. A clip is discarded if the similarity between its first and last frames falls below a predefined threshold. The thresholds used during this stage are summarized in Tab. 4.

Table 4. Threshold settings for the segmentation and stitching.

Threshold Type	Value
Segmentation threshold	0.45
First/last frame similarity threshold	0.90
Stitching threshold	0.65

**Coarse Filtering.** After acquiring the initial multi-shot clips, we apply a coarse filtering stage based on fundamental video attributes, including frame rate, spatial resolution, temporal duration (5-12 seconds), and overall aesthetic quality. To further ensure clear transition boundaries, the aesthetic score filtering is applied with particular attention to frames near the shot boundary, which helps prevent ambiguous or visually unclear transitions. As our study focuses on shot transition types, which are defined between compositions of shot pair, we retain only clips containing two shots. This stage yields a pool of approximately 500K candidate videos.

**Fine-Grained Transition Filtering.** Inter-shot consistency is crucial for selecting meaningful shot transitions. Extremely low similarity leads to abrupt or implausible visual break, while excessively high similarity reduces diversity and often corresponds to flash-like effects rather than genuine transitions. Although the segmentation-and-stitching stage provides coarse control, a more refined selection step is required to ensure spatial or causal coherence between shots and avoid potential confusion during model training.

To address excessive similarity, we compute CLIP [34] feature similarity and remove pairs with similarity greater than 0.95, effectively filtering out cases of incorrect segmentation or no-transition scenes (e.g., light flickering). For low similarity, we use a VLM-based [43] method, as image-feature-based metrics tend to focus on vibe, style, and tone rather than spatial or causal relationships. The prompts used for VLM filtering are shown in Fig. 7. This fine-grained filtering yields a final set of 40K videos.

**Caption Generation.** We employ GPT-5-mini to generate hierarchical captions for each curated video. As shown in Fig. 8, each video is annotated with a general description that captures the main subjects across the two shots, together with more fine-grained shot-specific captions that articulate the relevant cinematographic characteristics. This hierarchical annotation scheme endows the dataset with structured and professionally oriented cinematic transition knowledge.

### B. Statistic of Dataset

This section presents detailed statistical characteristics of ShotWeaver40K, and visualizes the distributions of aesthetic score, shot duration, and inter-shot CLIP feature similarity in Fig. 9. Across the 40K curated two-shot clips, the dataset exhibits an average duration of 8.72 seconds, an average aesthetic score of 6.21, and a mean CLIP feature similarity of 0.7817 between adjacent-shot frame pairs. These statistics indicate that the dataset maintains high aesthetic quality and strictly controlled inter-shot consistency, making it well suited for training frameworks aimed at the preliminary exploration of shot transition modeling.

### C. Complete Results of Qualitative Results

In Sec. 4.2, the transition-type-aware multi-shot video generation performance of ShotDirector is compared against several baseline models, and representative qualitative results are presented in Fig. 4. This section provides the full set of qualitative visualizations, offering a more comprehensive and convincing analysis that further supports the effectiveness of ShotDirector. We present the results of different models in video form on the project page provided in the supplementary material, and the corresponding full prompts are included in Fig. 10. These extensive comparisons enable a more thorough examination of the strengths and limitations of each method.

For multi-shot video generation methods, shot-by-shot

### Prompt for filtering low-similarity or low-quality videos

Here are two consecutive frames from a video. Please complete the following tasks:

1. Determine whether the images depict a clear, high-quality realistic scene (not animation, not 3D rendering, not blurry, not vulgar, no text symbols in the image).
2. Determine whether the two images show the same scene. To be considered the same scene, the images must share certain overlapping visual content. Merely having a similar atmosphere, lighting, or style without overlapping objects or regions should be classified as different scenes.

Finally, return the results in the following format, and only return the results without any additional explanation: High-quality / Low-quality, Same Scene / Different Scene (e.g., High-quality, Different Scene)

Figure 7. Prompt for filtering low-similarity or low-quality videos using Qwen.

stitching approaches exhibit low consistency across shots. StoryDiffusion [53] produces frames with substantial disparities between shots, preventing coherent multi-shot composition. Phantom [28], which incorporates reference images, improves subject consistency but struggles with background continuity, failing to maintain a unified scene and thus breaking the narrative flow.

For end-to-end multi-shot generation models, CineTrans [46] is capable of producing shot transitions but lacks an understanding of transition-type semantics, resulting in outputs without professional cinematic characteristics. Mask2DiT [33] tends to generate animation-like visuals, leading to relatively low perceptual quality.

Regarding large-scale pretrained models, HunyuanVideo [26] and Wan2.2 [41] demonstrate certain transition effects owing to their strong semantic understanding, yet the outcomes remain unstable, and neither consistency nor transition type can be reliably preserved.

For camera-controlled methods, SynCamMaster [5] achieves reasonable camera control but at the expense of visual quality, likely due to the reliance on synthetic training data, which causes notable deviations from real-world video appearance. ReCamMaster [4] tends to perform camera motions instead of hard shot transitions and may suffer from visual artifacts during switching.

Across all these comparisons, the proposed framework demonstrates the ability to produce stable transitions while maintaining an understanding of professional cinematographic language, thereby enabling more controllable and semantically grounded shot transitions.

## D. Details of Evaluation Metrics

In Sec. 4.2, a comprehensive evaluation protocol is introduced to quantitatively assess the performance of the proposed framework. This section provides additional details regarding the computation of several key metrics, further supporting the reproducibility and validity of the evaluation.

**Transition Confidence Score.** To quantify the clarity and

reliability of shot transitions in generated videos, the Transition Confidence Score is computed using TransNetV2 [38]. Given an input video sequence, TransNetV2 produces a frame-wise transition likelihood feature  $d \in \mathbb{R}^{F \times 1}$ , where each element corresponds to the transition probability of a specific frame after applying a sigmoid activation. The score for the entire video is defined as the maximum confidence over all frames:

$$\text{Transition Confidence Score} = \max(\sigma(d)), \quad (7)$$

where  $\sigma(\cdot)$  denotes the sigmoid function. This metric captures whether a transition occurs within the sequence and how sharply it is presented (e.g., distinguishing hard cuts from gradual transitions), providing an intuitive measure of the model’s ability to generate well-structured shot transitions.

**Transition Type Accuracy.** Beyond detecting transitions, correctly modeling transition types is crucial for evaluating transition-aware multi-shot video generation. To this end, Transition Type Accuracy is introduced to assess a model’s ability to adapt to different categories of transitions. A vision-language model (VLM) [43] is employed to classify the transition type of each generated video, and accuracy is computed against the ground-truth prompts. Tab. 5 reports the distribution of transition types within the evaluation set, and Figure 11 illustrates the prompt used for VLM-based recognition.

Table 5. Distribution of transition types in the evaluation set.

Transition Type	Count
Cut-in	24
Cut-out	26
Shot/Reverse Shot	25
Multi-Angle	15

Prompt for hierarchical captioning
<p>You are a Visual-Language Model. You will be given two images (they are consecutive frames intended to represent a shot transition). Your task is to perform a check and caption process based on these two images according to the following guidelines.</p> <p>General Caption (across both images):</p> <p>Subject: Identify the main subject(s) (if present), assign a name(s) with &lt;mark&gt; token, e.g., &lt;Man 1&gt; and describe the appearance. The name(s) should be used consistently throughout the captions. If no subject is present, this item can be omitted.</p> <p>Description: Using the assigned subject name(s), provide an overall description of the two images together, covering the subject, environment, scene, and ongoing action.</p> <p>Shot Caption (per image):</p> <p>Shot n: Using the assigned subject name(s), first describe the visual content of the frame, then add attributes from a cinematography perspective (Framing, Focus, Lighting, Color, Camera Angle, etc.).</p> <p>Transition Caption (across two shots):</p> <p>First assign a Transition Label from the following four categories: (Shot/Reverse Shot, Cut-in, Cut-out, Multi-Angle).</p> <p>Shot/Reverse Shot: Alternating shots between two subjects (commonly used in dialogues), switching perspective back and forth.</p> <p>Cut-in: Transition from a wider shot to a closer detail of the same subject or scene.</p> <p>Cut-out: Transition from a close-up or detail shot to a wider shot of the same subject or scene.</p> <p>Multi-Angle: Transition between two different camera angles of the same subject within the same scene.</p> <p>Then provide an explanation of how the first image transitions into the second.</p> <p>Final Output Template:</p> <p>General Caption:</p> <p>Subject: &lt;Girl 1&gt; a girl with blonde hair. &lt;Boy 1&gt; a boy in red T-shirt.</p> <p>Description: &lt;Girl 1&gt; is sitting at the piano in a softly lit room, with some flowers decorating. &lt;Boy 1&gt; happily claps hands for her.</p> <p>Shot Caption:</p> <p>Shot 1: &lt;Girl 1&gt; is seated at the piano, and &lt;Boy 1&gt; is clapping hands near her. Medium shot, eye-level, shallow focus, soft daylight, muted tones.</p> <p>Shot 2: &lt;Girl 1&gt;'s hands press the keys. Close-up, high angle, sharp focus on hands, warm lighting.</p> <p>Transition Caption: Cut-in. Cut from a medium shot of &lt;Girl 1&gt; at the piano to a close-up of her hands, emphasizing detail and continuity.</p> <p>Do not return anything beyond what is specified in the template. Each description should be around 25 words.</p>

Figure 8. Prompt for hierarchical captioning using GPT-5-mini.

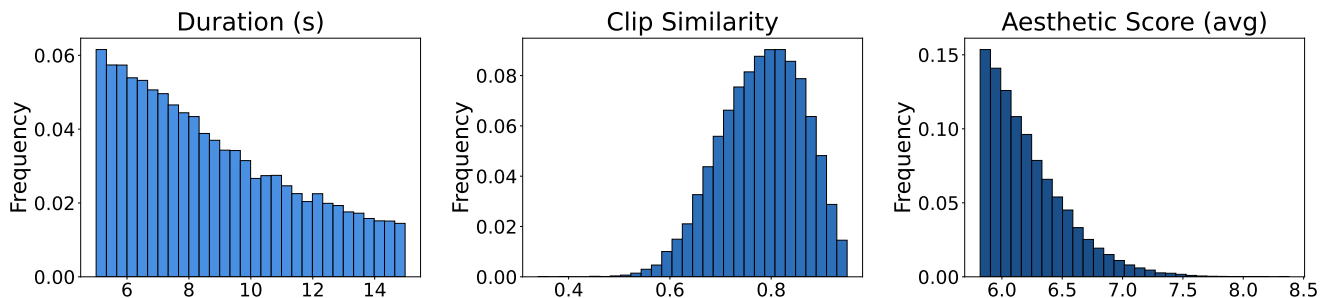


Figure 9. Distribution of key video attributes in ShotWeaver40K.

Cut-in
<p><b>[Subject]</b> &lt;Man 1&gt; an elderly Asian man in an ornate peach and black robe. &lt;Man 2&gt; a bespectacled man in a gray suit. &lt;Man 3&gt; a large bearded man in a blue vest.</p> <p><b>[General]</b> &lt;Man 1&gt; speaks with &lt;Man 2&gt; and &lt;Man 3&gt; in a lush outdoor garden; the second frame cuts closer to &lt;Man 1&gt; revealing hooded attendants standing behind him.</p> <p><b>[Transition]</b> Cut-in. Cut from a medium-wide group shot including &lt;Man 2&gt; and &lt;Man 3&gt; to a closer, intimate portrait of &lt;Man 1&gt;, highlighting his expression and attendants.</p> <p><b>[Shot 1]</b> &lt;Man 1&gt; holds a book facing &lt;Man 2&gt; and &lt;Man 3&gt; outdoors. Medium-wide shot, eye-level, natural daylight, moderate depth-of-field, vibrant costume colors, balanced composition.</p> <p><b>[Shot 2]</b> &lt;Man 1&gt; in a tighter close-up with hooded attendants blurred behind. Close-up, shallow focus on face, soft natural lighting, warm tones, intimate framing emphasizing expression.</p>
Shot/Reverse Shot
<p><b>[Subject]</b> &lt;Woman 1&gt; a blonde woman wearing a bright pink blouse. &lt;Man 1&gt; an older balding man in a beige collared shirt.</p> <p><b>[General]</b> &lt;Woman 1&gt; and &lt;Man 1&gt; stand outdoors in a leafy residential setting, exchanging dialogue; alternating camera perspectives capture her listening and him responding during daylight.</p> <p><b>[Transition]</b> Shot/Reverse Shot. Alternating perspective between &lt;Woman 1&gt; and &lt;Man 1&gt; as they converse; cut switches from her mid-close to his medium close-up, maintaining spatial and conversational continuity.</p> <p><b>[Shot 1]</b> &lt;Woman 1&gt; faces the camera in a bright pink blouse, mid-close framing; shallow depth of field, soft natural light, vibrant color, eye-level angle, attentive expression.</p> <p><b>[Shot 2]</b> &lt;Man 1&gt; looks toward &lt;Woman 1&gt;, medium close-up with her shoulder blurred in foreground; sharp facial focus, soft daylight, neutral warm palette, eye-level perspective.</p>
Cut-out
<p><b>[Subject]</b> &lt;Man 1&gt; a middle-aged man with short curly hair and a mustache, wearing a dark blue bathrobe, cigarette in mouth, holding a mug and a lighter.</p> <p><b>[General]</b> &lt;Man 1&gt; stands in a brick stoop doorway across two frames, smoking while holding a mug and inspecting a lighter, suggesting a quiet morning routine on the steps.</p> <p><b>[Transition]</b> Cut-out. The sequence pulls back from a medium close-up of &lt;Man 1&gt; inspecting lighter and smoking to a wider shot that reveals his full figure and the surrounding stoop.</p> <p><b>[Shot 1]</b> &lt;Man 1&gt; framed chest-up in doorjamb, cigarette in mouth, lighter and mug visible; medium close-up, shallow focus, low-key lighting, cool muted tones, slight high-angle.</p> <p><b>[Shot 2]</b> &lt;Man 1&gt; seen full-body on stoop holding mug and lighter, wearing robe and sandals; wide shot, eye-level, deeper focus, naturalistic lighting, revealing environment and stairs.</p>
Multi-Angle
<p><b>[Subject]</b> &lt;Man 1&gt; an older man in a dark suit and fedora wearing a Nazi lapel pin.</p> <p><b>[General]</b> &lt;Man 1&gt; inspects and holds a large framed coat-of-arms in a dim, wood-paneled foyer, moving from a side view to a frontal, backlit view.</p> <p><b>[Transition]</b> Multi-Angle. Cut between two camera angles of &lt;Man 1&gt;: a side biased medium-close emphasizing hands and object, then a frontal medium revealing silhouette and surroundings.</p> <p><b>[Shot 1]</b> &lt;Man 1&gt; holds a framed coat-of-arms at chest level, looking down. Medium-close, slight side angle, shallow focus, cool low-key lighting, muted teal tones.</p> <p><b>[Shot 2]</b> &lt;Man 1&gt; viewed from front with hat shadowing his face as he opens the package. Medium shot, centered, strong backlight from stained glass, high-contrast rim lighting.</p>

Figure 10. Prompt for qualitative evaluation using different models. (The results in video form are on the project page.)

## E. Additional Results

### E.1. User Study

To complement the experimental results, we report the results of our user study in this section. Specifically, the par-

### Prompt for recognition of shot transition type

You are a Visual-Language Model. You will be given two consecutive frames from a video, representing a shot transition. Your task is to determine which type of transition the two frames illustrate.

There are four possible transition types, each defined as follows:

1. Shot/Reverse Shot: Used primarily in dialogue or interaction scenes between two subjects. The camera alternates between two characters or perspectives, typically one facing left and the other facing right, while maintaining spatial continuity.
2. Cut-in: A transition from a wider shot to a closer detail within the same scene or subject. The second shot focuses on a smaller, more specific region or action visible in the first shot.
3. Cut-out: The opposite of a cut-in. It moves from a closer shot (e.g., a detail or close-up) to a wider framing of the same subject or environment, providing more spatial context.
4. Multi-Angle: A transition between two shots showing the same subject or scene from different camera angles, while keeping framing scale roughly similar (e.g., from front view to side view, or from 45° to 90° rotation).
5. No-Transition: No explicit shot transition between the two images.

Your goal is to analyze the two given images carefully and classify the transition into exactly one of the above four categories. Do not explain your reasoning or describe the images. Only return one of the following labels as the final answer:

Shot/Reverse Shot, Cut-in, Cut-out, Multi-Angle.

Figure 11. Prompt for recognition of shot transition type using Qwen.

ticipants rate each result on a scale from 1 to 5. We evaluate different methods from the perspective of Transition Control, Visual Quality and Consistency, with 5-point Likert-scale evaluation presented in Tab. 6. It can be observed that ShotDirector also demonstrates strong performance in terms of user preference.

Table 6. Ablation results for camera information. The best is **bold**.

Method	Transition Control	Visual Quality	Consistency
StoryDiffusion	2.72±0.74	3.13±0.52	2.06±1.05
Wan2.2	1.73±0.98	3.16±0.88	3.27±0.79
SynCamMaster	3.53±0.69	1.38±0.59	<b>3.98±0.65</b>
CineTrans	3.22±0.89	3.35±0.81	3.26±0.94
<b>Ours</b>	<b>3.96±0.84</b>	<b>3.38±0.97</b>	3.96±0.86

## E.2. Results of More Shots

In Sec. 4, we primarily focus on the modeling of transition itself, which occurs between two consecutive shots by definition. Therefore, the core experimental results are presented in the 2-shot setting, as it provides the most direct and controlled scenario for analyzing transition quality and controllability. Nevertheless, the proposed framework is not limited to two shots. Our model is fully capable of generating multi-shot videos with more segments, as shown in Fig. 12, where multiple consecutive transitions are handled in a coherent and stable manner. These results further demonstrate that the transition modeling mechanism generally works beyond the 2-shot case.

## E.3. Additional Ablation and Robustness Analysis

In Sec. 4.3, we performed an ablation study to verify the effectiveness of the key components of ShotDirector. In this



Figure 12. Result of 4-shot video.

section, we further extend the analysis with additional ablations focusing on the hierarchical prompting design and the generalization ability of the model.

To isolate the contribution of hierarchical prompting design, we compare the full model against three variants: (i) w/o local prompts, (ii) w/o global prompts, and (iii) w/o hierarchy design, with evaluation metrics consistent with Tab. 1 in the main paper. The results shown in Tab. 7 indicate that global prompts play a crucial role in guiding transition type control and overall text adherence. In contrast, local prompts, together with the hierarchical structure, substantially improve cross-shot semantic consistency. The best shot-transition quality is achieved only when both global and local prompts are integrated in a hierarchical manner. These findings verify that the hierarchical design is necessary and that a simple concatenation of prompts is insufficient to achieve comparable performance.

Additionally, to examine whether the camera condition design merely overfits synthetic trajectories, we evaluate the model on out-of-domain multi-view camera control signals from a separate synthetic dataset. Compared to in-domain signals, the performance degradation is minimal (RotErr: 0.6156 vs. 0.5907; TransErr: 0.5419 vs. 0.5393). The small increase in error indicates that the model generalizes well to unseen camera trajectories, demonstrating that the architectural design contributes to robustness.



Table 7. Results of ablation study on hierarchical prompting design.

Method	Transition Control		Overall Quality				Cross-shot Consistency	
	Confidence $\uparrow$	Type Acc $\uparrow$	Aesthetic $\uparrow$	Imaging $\uparrow$	Overall Consistency $\uparrow$	FVD $\downarrow$	Semantic $\uparrow$	Visual $\uparrow$
w/o local	0.8813	0.6333	<u>0.6271</u>	0.6812	<u>0.2367</u>	78.73	0.6572	0.7551
w/o global	<u>0.8881</u>	0.3778	0.6108	0.6839	0.1364	76.74	<u>0.7318</u>	<u>0.8131</u>
w/o hierarchy	0.8865	<u>0.6556</u>	0.6249	<b>0.7024</b>	0.2341	<u>74.07</u>	0.6044	0.7343
<b>Ours</b>	<b>0.8956</b>	<b>0.6744</b>	<b>0.6374</b>	<u>0.6984</u>	<b>0.2394</b>	<b>68.45</b>	<b>0.7918</b>	<b>0.8251</b>

## F. Limitation

### F.1. Failure Case

Fig. 13 presents failure cases observed during inference. We find that in certain samples the visual characteristics of different subjects become mixed, suggesting that the model lacks a clear one-to-one correspondence when multiple subjects are present. This issue may indicate insufficient understanding of multi-subject scenarios within the model. A possible improvement is to provide more detailed bounding-box-level annotations in the dataset, which could enhance the model’s ability to better understand and model multi-subject situations.



Figure 13. A representative failure case in which the visual characteristics of multiple subjects become unintentionally blended during generation.

### F.2. Future Work

Although ShotDirector demonstrates strong performance in Sec. 4.2, it still exhibits several limitations and opens up promising avenues for future exploration.

- **Integrating camera-control and semantic cues more cohesively.** Our approach employs two separate modules to govern shot transitions: one conditioned on high-level semantic information and the other on parameter-level camera control signals. A compelling future direction is to investigate how to unify these two forms of conditioning more seamlessly, enabling a more coherent and expressive transition modeling process.
- **Toward longer videos with more shot transitions.** Extending our framework to generate longer videos that contain a richer set of shot transitions represents another valuable research direction. We believe that scaling to longer temporal horizons is feasible with additional data. Given that the effectiveness of our method has been verified on ShotWeaver40K, further fine-tuning on extended datasets

may enable the model to generalize to videos with more shots and significantly longer durations.