MultiShotMaster: A Controllable Multi-Shot Video Generation Framework (Supplementary Material)

A. More Implementation Details

A.1. Details in Temporal Attention

To clarify the designs in temporal attention, including Multi-Shot Narrative RoPE and Spatiotemporal Position-Aware RoPE, we provide an Algorithm 1. Specifically, the complete in-context latents Z contain multi-shot video latents $z = [z_i]_{i=1}^{N_{shot}}$ and reference latents $z^{ref} = [z^m]_{m=1}^{N_{ref}}$. N_{shot} represents shot count, N_{ref} represents the number of input reference images (subjects and backgrounds). The input bounding box sequences of references $[boxes]_b^{N_{box}}$ contain N_{box} bounding boxes. Each bounding box is represented as $[(m, t, x_1, y_1, x_2, y_2)]$, indicating the bounding box of m-th reference at t-th frame. Note that for background references, the bounding boxes are fixed as $(m, t, 0, 0, H, W)$, where t is the first frame of the corresponding shot.

In temporal attention, the linear projections to_q, to_k, to_v first transform in-context latents Z to $\tilde{Q}$, $\tilde{K}$, $\tilde{V}$. Then, by applying the Multi-Shot Narrative RoPE (i.e., Eq. 2 in the main paper), the query and key of each shot are introduced explicit shot transition signals, while keeping the narrative temporal order. For m-th reference containing N_{box}^m boxes, we copy the query and key of the m-th reference N_{box}^m times. Each copy is then applied with a Spatiotemporal Position-Aware RoPE based on the corresponding box in $[boxes]_b^{N_{box}^m}$. Since RoPE is not applied to value component in attention mechanism, we copy the value for attention computation. After the attention computation with the proposed multi-shot & multi-reference attention mask, we aggregate the N_{box}^m reference copies for m-th reference by taking their mean. Finally, the multi-shot video $\hat{z}$ and the reference latents $\bar{z}^{ref}$ are concatenated along the token dimension and fed into the linear projection to_out. The attention output Z^* maintains the same dimension with the input in-context latents Z.

A.2. Training Paradigm

The three-stage training paradigm consists of: (1) we finetune the temporal attention for spatiotemporal-specified reference injection on 300k single-shot video data with 30 epochs, batch size 8, while keeping other model parameters frozen. (2) we finetune temporal attention, cross attention and FFN on 235k multi-shot & multi-reference data with 3

Algorithm 1 Temporal Attention with Multi-Shot Narrative RoPE and Spatiotemporal Position-Aware RoPE.

Input:

- In-context latents Z containing:
  - Multi-shot video latents $z = [z_i]_{i=1}^{N_{shot}}$
  - Reference latents $z^{ref} = [z^m]_{m=1}^{N_{ref}}$
- Bounding box sequences of references (subjects and backgrounds) $[boxes]_b^{N_{box}} = [(m, t, x_1, y_1, x_2, y_2)]_b^{N_{box}}$

Output: In-context latents Z^* after temporal attention

$\tilde{Q} = to\_q(Z)$, $\tilde{K} = to\_k(Z)$, $\tilde{V} = to\_v(Z)$

// Apply Multi-Shot Narrative RoPE

$Q = [Q_i]_{i=1}^{N_{shot}} = Eq.\ 2([\tilde{Q}_i]_{i=1}^{N_{shot}})$

$K = [K_i]_{i=1}^{N_{shot}} = Eq.\ 2([\tilde{K}_i]_{i=1}^{N_{shot}})$

$V = [\tilde{V}_i]_{i=1}^{N_{shot}}$

// Apply Spatiotemporal Position-Aware RoPE

$Q^{ref} = [Q_b^{ref}]_{b=0}^{N_{box}} = Eq.\ 3(Copy(\tilde{Q}^{ref}), [boxes]_b^{N_{box}})$

$K^{ref} = [K_b^{ref}]_{b=0}^{N_{box}} = Eq.\ 3(Copy(\tilde{K}^{ref}), [boxes]_b^{N_{box}})$

$V^{ref} = [\tilde{V}_b^{ref}]_{b=0}^{N_{box}} = Copy(\tilde{V}^{ref}, [boxes]_b^{N_{box}})$

// Attention Computataion

$\hat{Z}$ = Attention($[Q, Q^{ref}], [K, K^{ref}], [V, V^{ref}]$, Mask)

// Reference Aggregation

$\bar{z}^{ref} = [\bar{z}^m]_{m=1}^{N_{ref}} = [\ mean([\hat{z}^m]_b^{N_{box}^m})\ ]_{m=1}^{N_{ref}}$

$Z^* = to\_out([\hat{z}, \bar{z}^{ref}])$

return Z^*

Table 1. **Ablation study for Multi-Shot RoPE.** We experiment on multi-shot text-to-video generation without reference input.

	Inter-Shot Consistency↑			Transition Deviation↓	Narrative Coherence↑
	Semantic	Subject	Scene		
w/o MS RoPE	**0.702**	0.486	**0.455**	4.68	0.645
Ours (w/o Ref)	0.697	**0.491**	0.447	**1.72**	**0.695**

epochs, batch size 1. (3) following the second stage, we assign $(2\times)$ loss weight to subject regions and $(1\times)$ to backgrounds to train 0.5 epoch. We conduct ablation study for the training paradigm in Sec B.2 and Table 3.

A.3. Labeling Hierarchical Captions

As introduced in Sec 3.5 of the main paper, we employ Gemini-2.5 [1] to label the global caption and per-shot captions. The prompt template is shown in Fig. 1. We begin by proportionally sampling 20 frames from the multi-shot video, ensuring at least one frame is extracted from each shot, and use Gemini-2.5 to produce a comprehensive global caption. Then we employ Gemini-2.5 to reason the per-shot captions based on the global caption and each shot video (with a sampling frame stride of 15). Each subject is denoted by "Subject X, X $\in [1, 2, 3]$". As shown in Fig. 2, the cross-shot consistency of subject annotations is satisfactory due to the powerful Gemini-2.5 and our carefully-designed prompt template.

A.4. Merge Cross-Shot Tracking Annotations

As introduced in Sec 3.5 of the main paper, we conduct the tracking process shot-by-shot to obtain the bounding box sequence of each subject. To merge the cross-shot tracking results, we use Gemini-2.5 [1] to group the subject images by prompting with our carefully-designed prompt template as shown in Fig. 3.

A.5. Narrative Coherence

To comprehensively assess the narrative coherence of generated multi-shot videos, we employ Gemini-2.5 [1] to construct an automated evaluation metric. We begin by proportionally sampling 20 frames from the multi-shot video, ensuring at least one frame is extracted from each shot. Subsequently, we input these frames and the hierarchical captions as a pair into Gemini-2.5. We require Gemini-2.5 to strictly adhere to cinematic narrative logic and scrutinize cross-shot content across four core dimensions: Scene Consistency, Subject Consistency, Action Coherence, and Spatial Consistency, by the constructed elaborate instructions as shown in Fig. 4.

Specifically, Scene Consistency verifies the stability of the background, lighting, and atmosphere during transitions to ensure all shots depict the same setting; Subject Consistency strictly scrutinizes identity features and appearance attributes by comparing core objects across different viewpoints to detect unintended deviations; Action Coherence focuses on evaluating the temporal logic of dynamic behaviors to determine whether actions in subsequent shots constitute reasonable continuations of preceding ones; and Spatial Consistency examines whether the topological structure of relative positional relationships between subjects remains constant in accordance with cinematic language. Functioning as a binary classifier, the model outputs a "True" or "False" verdict for each dimension, thereby quantifying the generative model's capability in handling complex multi-shot spatiotemporal consistency.

Table 2. **Ablation study for reference injection.** We experiment on multi-shot reference-to-video generation.

	Aesthetic Score↑	Narrative Coherence↑	Reference Consistency↑ Subject	Scene	Grounding
w/o Mean	3.84	0.796	0.482	0.452	0.557
w/o Attn Mask	3.72	0.787	0.468	0.414	0.561
w/o STPA RoPE	3.79	0.761	0.425	0.363	✗
Ours (w/ Ref)	**3.86**	**0.825**	**0.493**	**0.456**	**0.594**

B. Ablation Study

B.1. Ablation Study for Network Design

We experiment with different settings to validate the effectiveness of the proposed designs in our framework:

- "w/o MS RoPE": without Multi-Shot Narrative RoPE, the shot transitions rely only on the per-shot captions.
- "w/o Mean": this setting randomly selects one copy from multiple copies of subject tokens after 3D attention, instead of averaging.
- "w/o Attn Mask": without Multi-Shot & Multi-Reference Attention Mask, this setting uses full attention along the temporal dimension.
- "w/o STPA RoPE": without the Spatiotemporal Position-Aware RoPE, this setting directly concatenates the reference tokens along the temporal dimension and applies the RoPE(t=0,h,w) to each reference.
- "Ours (w/o Ref)": this setting is trained using all the proposed designs, and infers multi-shot text-to-video generation without reference input.
- "Ours (w/ Ref)": this setting uses the same trained checkpoint as "Ours (w/o Ref)" and infers multi-shot reference-to-video generation.

Since the spatiotemporal-grounded reference injection might facilitate shot transitions, we do not provide reference input to compare with "w/o MS RoPE" setting. It relies only on the variations between per-shot captions to guide shot transitions, and uses the continuous RoPE to all frames of multi-shot videos in the temporal order. This setting cannot implement precise shot transitions by text prompts only, leading to unsatisfactory transition deviation score, as shown in Table 1. Due to the lack of shot transitions, there is almost no change between shots, resulting in higher semantic and scene consistency scores. With the proposed Multi-Shot Narrative RoPE, we can perform shot transition at user-specified timestamps with superior deviation score.

We further evaluate the designs in spatiotemporal-grounded reference injection. In addition to the mentioned

Table 3. Ablation Study for training paradigm. We experiment on multi-shot reference-to-video generation. The $1^{st}/2^{nd}$ best results of settings are indicated in underline/**bold**.

	Text Align.↑	Inter-Shot Consistency↑			Reference Consistency↑		
		Semantic	Subject	Scene	Subject	Background	Grounding
I: Multi-Shot+Ref. Injection	0.211	0.671	0.464	0.415	0.454	0.426	0.477
I: Multi-Shot II: Multi-Shot+Ref. Injection	0.219	0.695	0.481	0.433	0.472	0.451	0.578
I: Ref. Injection II: Multi-Shot+Ref. Injection	0.222	0.692	0.484	0.437	0.485	0.454	0.583
I: Ref. Injection II: Multi-Shot+Ref. Injection III: Multi-Shot+Subject-Focused Ref. Injection	**0.227**	**0.702**	**0.495**	**0.472**	**0.493**	**0.456**	**0.594**

metrics in the main paper, we further introduce Aesthetic Score [2] to measure the aesthetic quality of the generated multi-shot videos. "w/o Mean" might cause information loss, showing suboptimal results, as shown in Table 2. the excessively long contexts in "w/o Attn Mask" setting have unnecessary interactions between in-context tokens, leading to weak aesthetic score and reference consistency. "w/o STPA RoPE" cannot designate the specific shot or exact spatiotemporal position where the subjects and backgrounds appear, relying only on text prompts for positioning. It shows poor performance on reference consistency. Taking advantage of the effectiveness of the proposed designs, our method shows best performance on all metrics.

B.2. Ablation Study for Training Paradigm

We conduct ablation study for the three-stage training paradigm introduced in Sec 3.4 of the main paper. We first explore the order of multi-shot video generation and reference-to-video generation, then shows the performance of the subject-focused post-training. The first setting involves fintuning the pretrained text-to-video generation model to learn both multi-shot task and reference-to-video task simultaneously. However, because the diffusion loss is computed across all frames to optimize global consistency, this unified training paradigm shows inadequate for effectively learning both tasks. The second setting is first learning multi-shot text-to-video generation, followed by multi-shot reference-to-video generation, both using the curated multi-shot & multi-reference data. This setting achieves slightly lower subject consistency due to insufficient exposure to diverse subjects during training. Since the construction cost of multi-shot & multi-reference data is relatively high, we first train the model to learn spatiotemporal-grounded reference injection task on single-shot data, and then learn both tasks using the curated multi-shot & multi-reference data. It achieves better results on most metrics. Furthermore, we introduce subject-focused post-training that guides the model to prioritize subjects requiring higher consistency, which also promotes the modeling of cross-shot subject variations.

References

[1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 5, 6

[2] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 3

Prompt for Labeling Global Caption

System Instruction: You are an multi-shot video understanding expert that only outputs video captions.

Context Information

User:

Task Overview:
Your task is to analyze the number of subjects that appear in this multi-shot video, and describe the appearance of each subject in the video with one sentence. And describe the video scene roughly.

Task Requirements:
1. People, vehicles, animals, motor vehicles, food, and other independent objects are all subjects that can be described.
2. Subjects who appear only a few times and are not important to the storyline should be omitted.
3. Note that the video might have multiple shots, describe the same person no more than once.
4. Describe no more than four subjects based on importance.
5. Do not describe subjects that are too far away or too small.
6. Use no more than 20 words per description for each subject.

Expected Output Format:
"Subject 1: A young man with blonde hair, wearing a dark jacket over a light-colored shirt; Subject 2: A yellow dog with a big mouth; Subject 3: A brand-new white car features bright headlights and graceful curves. The whole scene takes place in the parking lot."

Input Multi-Shot Video:
These are the frames from the video:{**sampled_frames**}.

Prompt for Labeling Per-Shot Caption

System Instruction: You are an video understanding expert that only outputs video captions based on the input story and video.

Context Information

User:

Task Overview:
Your task is to describe the video content in terms of the subjects' expression and actions, scene background, and camera movement in a single paragraph, based on the subject descriptions in the story setting '{**global_caption**}'. The description should not exceed 80 words.

Task Requirements:
1. Analyze which subjects in the story setting are present in the current video. And use only existing subject numbers in story setting (e.g., 'Subject 1') to denote the visible subject. Some subjects in the story may not appear in the video.
2. Do not describe the subject's appearance. Focus on subjects' expression and actions, scene background, and camera movement.
3. First describe the subject facing the camera, then describe the other subjects. When describing camera position, specify which subjects are facing to the camera.
4. Subjects who appear only a few times and are not important to the storyline should be omitted.
5. Do not describe subjects that are far away or too small or too blurry.
6. Do not create new subject number.
7. Do not repeat the content in the story setting.

Expected Output Format:
"Subject 1 is walking through a dense forest, carrying a large plastic bag. Subject 2 is holding a gun, following behind Subject 1. Subject 3 is running across the ground, with its paws rhythmically hitting the surface. Subject 1 walks slowly and occasionally bends down to pick up items from the ground. The forest is filled with tall, slender trees, and the ground is covered with a mix of grass and fallen leaves. The camera follows them from behind, maintaining a steady and consistent view of their movements. The camera view is a medium shot, capturing the subjects and the surrounding forest. The camera movement is smooth and follows the man's path, maintaining a steady and consistent view of his actions."

Input Each Shot Video:
These are the frames from the video:{**sampled_frames**}.

Figure 1. Prompts of labeling global caption and per-shot captions. We first label the global caption by sampling frames from the input multi-shot video. Then we label the per-shot caption one by one.

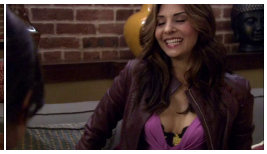
Global Caption:
Subject 1: A woman with long brown hair, wearing a purple top and a dark red leather jacket;
Subject 2: A woman with long blonde hair, wearing a vibrant red dress with a black pattern;
Subject 3: A woman with dark hair in a bun, wearing a black sleeveless dress and black boots.
The whole scene takes place in a modern apartment living room.

Shot1: **Subject 2** is holding a glass of wine, smiling and talking to Subject 1, then drinks from her glass. **Subject 1** is pouring wine into a glass, then holds her glass, looking at **Subject 2** and adjusting her leg. They are seated on a grey couch in a modern apartment living room with a brick wall and a yellow vase in the background. The camera is static, providing a medium shot of the two subjects.

Shot2: Subject 1 is seen in profile, looking to the right, and occasionally sips from her wine glass. **Subject 2** is seated next to **Subject 1**, holding two wine glasses, and looks towards Subject 1 with an engaged expression, appearing to listen or react. The scene takes place in a modern apartment living room with a brick wall visible in the background. The camera remains static, providing a medium shot of the two women.

Shot3: Subject 3 stands holding a wine glass with a serious expression, then bends down before sitting in a chair, gesturing animatedly while speaking. **Subject 1** and **Subject 2** are seated, holding wine glasses, mostly seen from behind. The modern apartment living room features a kitchen area, wall art, and a patterned rug. The camera maintains a medium, mostly static shot, with a slight pan.

Shot4: Subject 2 is smiling while pouring wine from a bottle into a glass. **Subject 1**, holding a wine glass, watches attentively. They are seated on a grey couch in a modern apartment living room with a brick wall background. The camera remains static, capturing a medium shot of the two subjects.



Shot5: Subject 1 is seated on a couch in a modern apartment living room, initially smiling broadly and laughing. She then tilts her head back, looking upwards with an amused expression. The background features a brick wall with a shelf holding a yellow vase and a statue. The camera remains static, providing a medium close-up of **Subject 1**.

Figure 2. Multi-shot video data example. By employing Gemini-2.5 [1] with the carefully-designed prompts as shown in Fig. 1, the labeled subjects could be consistent in global and per-shot captions.

Prompt for Merge Subjects

System Instruction: You are an subject image matching expert that only outputs JSON format.

Context Information

User:

Task Overview:
Your task is to group these '{frame_num}' images, The image names are listed as '{all_shot_all_id_name_list}'. You need to place image name with **the same identity or similar appearance** into the same group. The output group name is named as "new_id_0", "new_id_1"...

Task Requirements:
1. These images might contain diverse categories, including people, vehicles, animals, food, and other subjects.
2. The images from the same subject might be shot from different angles. For example, you may see a person's front view and back view.
3. Each image could only be assigned to one group.
4. If the number of groups exceeds 4, only output the first 4 clearest subjects.
5. Only return the json format and strictly follow the template below.

Expected Output Format:
```
{
"new_id_0": ["shot_0-id_0", "shot_1-id_2"],
"new_id_1": ["shot_0-id_1", "shot_1-id_1", "shot_2-id_3"],
"new_id_2": ["shot_0-id_2", "shot_1-id_0", "shot_2-id_5"],
"new_id_3": ["shot_3-id_0"]
}
```

Input Subject Images from All Shots:
These are the subject images:{**subject_images**}.

Figure 3. By employing Gemini-2.5 [1] to group the subject images, we obtain complete multi-shot tracking results.

Narrative Coherence Metric

System Instruction: You are a professional AI video evaluation assistant.

Context Information

User:

Your task is to strictly evaluate the continuity and consistency of an AI-generated Multi-shot video. This video should depict **the same scene**, just shown from different camera angles (shot changes).
The input will include:
1. Global Caption: An overall description of the entire video scene.
2. Multiple sets of visual and text inputs: The video is divided by shots. For each shot, you will receive:
    * Chronologically ordered sampled frames (at least 1 per shot).
    * A description for that shot (Shot Caption).

Your evaluation includes four core principles. You must evaluate the consistency of the entire video (across all shots) based on the following four criteria:
1. **Scene Consistency**:
    * Objective: Evaluate whether all shots take place in the **same** scene.
    * Checkpoints: Background environment, object placement, lighting conditions, and overall atmosphere.
    * Judgment Criteria:
        * True: The scene, lighting, and atmosphere remain consistent across all shots, clearly indicating the same location and time.
        * False: The background, lighting, or atmosphere changes abruptly and illogically (e.g., suddenly jumping from indoors to outdoors, or from day to night).
2. **Subject Consistency**:
    * Objective: Evaluate whether the core subjects (people, animals, or key objects) in the video remain consistent after shot changes.
    * Checkpoint (Identity): Is it the **same** subject after the cut? (e.g., the same person, the same dog).
    * Checkpoint (Appearance): Does the subject's appearance remain unchanged? (e.g., the same person's clothes, hairstyle, accessories).
    * Judgment Criteria:
        * True: The subject's identity and appearance remain consistent across all shots.
        * False: The subject's identity changes (A becomes B), or the subject's appearance (e.g., clothing color) changes illogically after a cut.
3. **Action Coherence**:
* Objective: (Only applies to dynamic subjects, like people or animals) Evaluate whether the subject's actions are **logically coherent** in time across shot changes.
    * Checkpoint: Is the action in the new shot a reasonable continuation of the action from the previous shot?
    * Judgment Criteria:
* True: The action in the next shot is a reasonable continuation of the action from the previous shot (e.g., Shot 1 shows a hand halfway raised, Shot 2 shows the hand fully raised).
* False: The action is reset (e.g., Shot 1 shows a hand raised, Shot 2 shows the hand back at the starting position; a person who was running is suddenly standing still), or it jumps to a completely unrelated action, breaking the temporal sequence.
4. **Spatial Consistency**:
    * Objective: Evaluate whether the **relative spatial layout** of subjects and their environment remains reasonable after a shot change.
    * Checkpoint (Relative Position): If "A is to the left of B" in Shot 1, does this relative relationship hold in a close-up in Shot 2?
    * Checkpoint (180-Degree Axis): Do the shot changes follow basic cinematic spatial logic (e.g., not "crossing the axis," which could confuse the audience about spatial relationships)?
    * Judgment Criteria:
* True: Spatial relationships remain consistent (e.g., A is on B's left in Shot 1, and this relative position is maintained in the close-up).
* False: Relative positions are disordered (e.g., A suddenly teleports from B's left to B's right), or the shot change causes complete spatial disorientation.

Strict Output Format:
You will receive four evaluation questions. You **must** and **only** answer 'True' or 'False' for each question. Do not add any explanations, justifications, or extra text.
Your final response must strictly adhere to this format:
Scene Consistency: True/False
Subject Consistency: True/False
Action Coherence: True/False
Spatial Consistency: True/False

Input:
Global Caption: {*global_caption*}
--- SHOT 1 ---
{*keyframe_images_from_shot_1*}
Shot 1 Caption:{*shot_1_caption*}
--- SHOT 2 ---
{*keyframe_images_from_shot_2*}
Shot 2 Caption:{*shot_2_caption*}
……

--- Evaluation Questions ---
Based on all provided shots and descriptions, please answer the following questions with only 'True' or 'False'.
Q1: **Scene Consistency**: The background, lighting, and atmosphere remain consistent across all shots, indicating they are in the same location and time.
Q2: **Subject Consistency**: The subjects (people, animals, key objects) in the video maintain the same identity and appearance (e.g., same clothes, features) across all shots.
Q3: **Action Coherence**: The actions of dynamic subjects show logical continuity after shot changes, rather than suddenly resetting or jumping.
Q4: **Spatial Consistency**: The relative spatial relationships between subjects (e.g., A is to the left of B) and the layout with the environment remain reasonable and non-confusing after shot changes (e.g., not violating the 180-degree axis principle).

Figure 4. We require Gemini-2.5 [1] to strictly adhere to cinematic narrative logic and scrutinize cross-shot content across four core dimensions: Scene Consistency, Subject Consistency, Action Coherence, and Spatial Consistency.