MANICAST: MULTI-SHOT ANIMATION CREATION WITH STRUCTURED CHARACTER-BOUND SCRIPTS

Anonymous authors
Paper under double-blind review

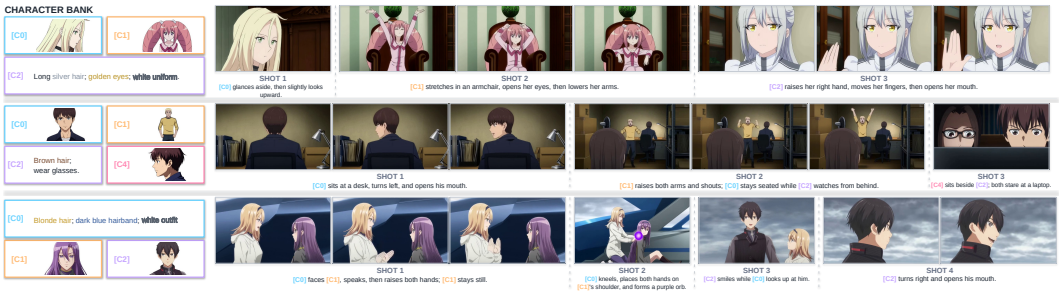

Figure 1: MAniCast, a framework for controllable multi-shot animation creation with Structured Character-Bound Scripts. Given characters defined by text or reference images together with story- and shot-level scripts containing explicit character placeholders, MAniCast generates multi-shot animations with the specified characters and actions.

ABSTRACT

Creating controllable multi-shot animations remains difficult for non-expert users because existing approaches typically require dense frame-level sketches, poses, or keyframes produced with professional drawing and animation skills. A more intuitive alternative is to create animations using only text descriptions and reference images. However, integrating characters defined by text descriptions and reference images with textual narratives remains challenging, particularly in multi-character, multi-shot animation creation. To address this problem, we present *MAniCast*, the first framework to combine modality-unified character casting with explicit identity-to-role binding for multi-shot animation generation. Structured Character-Bound Scripts combine story- and shot-level instructions with a shared character bank in which each identity is defined by either a detailed text description or a reference image, while explicit placeholders assign selected identities to roles and actions across shots. To realize these assignments at the network level, we introduce Character-Bound Attention-over-Attention (AoA), which derives identity features from textual or visual character definitions and injects them into the corresponding placeholders in the narrative. We further construct *MAniScript*, a multi-shot animation dataset annotated with Structured Character-Bound Scripts, each of which contains a character bank of text- or image-defined identities. Experiments demonstrate that MAniCast outperforms prior text-controlled and reference-based baselines in casting text- and image-defined characters, executing identity-bound actions, and generating complex interactions across shots. Our model and dataset will be publicly released upon acceptance.

1 INTRODUCTION

Creating controllable multi-shot animations remains difficult for non-expert users. Existing animation generation methods commonly obtain fine-grained controls from line-art sequences, reference keyframes, or storyboard drawings (Meng et al., 2025; Li et al., 2026; Xu et al., 2026). Although effective, these inputs require drawing expertise and substantial manual effort, making them difficult

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.

for non-professional users to produce. Therefore, a more intuitive approach is to let users describe the intended story and shots in natural language and define characters through text descriptions and reference images. Existing reference-conditioned approaches associate visual references with subject mentions through global conditioning (Jiang et al., 2025; Zhou et al., 2026a), cross-modal alignment (Liu et al., 2025; Chen et al., 2025; Fei et al., 2025), subject identifiers (Zhou et al., 2026b; Tencent Hunyuan Team, 2026), or specialized feature routing (Huang et al., 2025a; Zhang et al., 2025b). Recent structured representations further separate persistent entity definitions from shot narratives using reference banks and identifiers (Zhou et al., 2026b; Tencent Hunyuan Team, 2026). However, integrating characters defined by either text descriptions or reference images with textual narratives remains challenging: existing methods do not jointly treat both types of identities as cast members and explicitly bind them to narrative roles and actions through a unified mechanism. Such modality-unified explicit binding is critical for intuitive and accurate character casting in multi-character, multi-shot animation creation.

To address this problem, we introduce *MAniCast*, the first framework to combine modality-unified character casting with explicit identity-to-role binding for multi-shot animation generation. Rather than requiring frame-level line art, drawings, or motion annotations, MAniCast organizes user inputs through Structured Character-Bound Scripts, each comprising a story-level description, shot-level instructions, and a shared character bank in which each identity is defined by either a detailed text description or a reference image. As shown in Figure 1, explicit character placeholders bind selected identities in the character bank to their narrative roles and actions (e.g., "[C0] shakes hands with [C1]"). This design allows users to intuitively define characters once, conveniently reuse them across shots, select who appears in each shot, and specify the actions they perform.

Although Structured Character-Bound Scripts establish identity-to-role binding at the input level, realizing the same assignments at the network level remains challenging. Textual features of character definitions and narrative scripts are encoded separately, while visual references and narrative text are processed through separate self-attention and cross-attention, respectively. The model must therefore recover identity-to-role correspondence across different representations and attention paths, but standard DiTs lack a unified mechanism for routing textual or visual identity features to the correct roles and actions.

To enable this unified routing, we propose Character-Bound Attention-over-Attention (AoA). For image-defined characters, masked self-attention ensures that video tokens observe only the visual references assigned to the current shot. In the subsequent cross-attention layers, AoA derives identity-specific features from either text descriptions or reference images and injects them into the corresponding character placeholders in the narrative condition. By connecting the multi-modal character bank to the narrative script, AoA enables accurate casting of both text- and image-defined characters.

In summary, our main contributions are as follows:

- We introduce *MAniCast*, the first framework to combine modality-unified character casting with explicit identity-to-role binding through Structured Character-Bound Scripts, which organize text- and image-defined identities in a shared character bank and assign them to narrative roles and actions across shots.
- We propose Character-Bound Attention-over-Attention, which realizes these bindings at the network level by routing identity-specific textual or visual features from the character bank to the corresponding placeholders in the narrative text.
- We construct and contribute *MAniScript*, a multi-shot animation dataset annotated with Structured Character-Bound Scripts, each of which contains a character bank of text- or image-defined identities, to support research on controllable, narrative-driven animation creation.
- We conduct extensive experiments to demonstrate that MAniCast outperforms prior text-controlled and reference-based baselines in casting text- and image-defined characters, executing identity-bound actions, and generating interactions across shots.

2 Related Work

Controllable Animation Creation. Professional animation workflows rely on keyframes and line-art sequences. LVCD (Huang et al., 2024a) and AniDoc (Meng et al., 2025) colorize line-art sequences using character references. ToonComposer (Li et al., 2026) generates in-between frames and colors from sparse sketches and a colored reference frame. DrawVideo (Xu et al., 2026) uses storyboard keyframes with appearance and motion prompts to control individual shots. Although these methods reduce manual work, preparing line art, keyframes, or storyboards still demands professional drawing or animation skills. Text-controlled prompting and reference-based conditioning (Jiang et al., 2025; Zhou et al., 2026a; Liu et al., 2025) offer a more intuitive way to specify stories and character appearances. MAniCast follows this direction for animation creation without requiring frame-level visual guidance.

Structured Narrative and Character Control. Story-level and multi-shot methods organize narratives as sequences of prompts or shot descriptions (Zhou et al., 2024; Kara et al., 2025; Wang et al., 2026; Guo et al., 2025; Qiu et al., 2025; Zheng et al., 2025; Atzmon et al., 2025). Free-form shot prompts often mix characters, actions, and scenes (Zhou et al., 2024; Kara et al., 2025; Wang et al., 2026; Zheng et al., 2025). Subject identifiers and hierarchical formats provide more structure, but do not necessarily separate reusable character definitions from the narrative (Guo et al., 2025; Qiu et al., 2025). Aura (Zhou et al., 2026b) separates cast and setting definitions from shot narratives. Script-a-Video (Tencent Hunyuan Team, 2026) stores persistent entities in a Reference Stream linked to Shot and Event streams. Both demonstrate the value of reusable entity definitions, while leaving open how text- and image-defined characters can be cast through one unified mechanism.

Reference Grounding & Modality-Unified Character Casting. *Implicit reference grounding.* Some methods associate reference images with ordinary subject words or appearance descriptions rather than explicit identity markers in the narrative. Phantom (Liu et al., 2025) aligns visual references with text. Video Alchemist (Chen et al., 2025) and ConceptMaster (Huang et al., 2025b) fuse paired image-text subject conditions, while MAGREF (Deng et al., 2026) aligns subject labels with visual regions. For multi-shot generation, AnimeShooterGen (Qiu et al., 2025) combines references with current and previous shot context, MultiShotMaster (Wang et al., 2026) assigns reference tokens to shots or regions, and StoryMem (Zhang et al., 2025a) uses a memory of earlier keyframes. Their actions are still described using ordinary subject words, leaving identity-to-role assignment implicit.

Explicit reference binding. Other methods connect narrative markers directly to visual references. Movie Weaver (Liang et al., 2025) anchors references after subject phrases. HunyuanCustom (Hu et al., 2025a) and PolyVivid (Hu et al., 2025b) insert image tokens into textual conditions. PoCo (Huang et al., 2026) routes `@character` IDs, while EM-Vid (Vandersanden et al., 2026) uses character IDs to retrieve visual memory. Aura (Zhou et al., 2026b) binds structured cast and shot definitions to visual references, while Script-a-Video (Tencent Hunyuan Team, 2026) connects persistent entities to shots through identifiers. However, they do not route text- and image-defined identities to narrative roles through a common binding mechanism to support unified character casting. MAniCast instead stores both identity types in a shared character bank and uses Character-Bound Attention-over-Attention to inject their features at the designated narrative placeholders.

3 Preliminary

MAniCast builds on Wan 2.2 (Wan et al., 2025), a Diffusion Transformer (DiT) (Peebles & Xie, 2023) operating on video latents. Its self-attention models spatiotemporal interactions among video tokens, while cross-attention conditions them on textual tokens $\boldsymbol{T}$. We train the model with rectified flow (Lipman et al., 2023; Liu et al., 2023). Let $\mathbf{X}_0$ be a clean video latent, $\mathbf{X}_1$ Gaussian noise, and $\tau \in [0, 1]$ the flow time. The linear path is $\mathbf{X}_\tau = (1-\tau)\mathbf{X}_0 + \tau\mathbf{X}_1$, and the model predicts its velocity field by minimizing:

$$\mathcal{L} = \mathbb{E}_{\tau,\mathbf{X}_0,\mathbf{X}_1}[\|\mathbf{U}_{\boldsymbol{\theta}}(\mathbf{X}_\tau, \tau, \boldsymbol{T}) - (\mathbf{X}_1 - \mathbf{X}_0)\|^2] \tag{1}$$

During generation, the flow ODE is solved in reverse, mapping noise $\mathbf{X}_1$ at $\tau = 1$ to the clean latent $\mathbf{X}_0$ at $\tau = 0$.

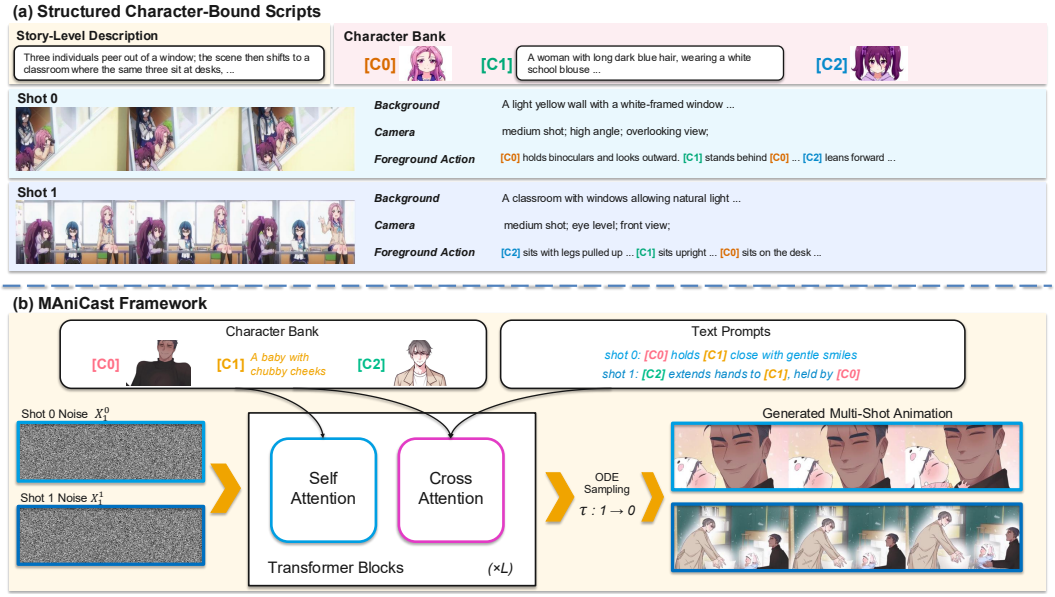

Figure 2: (a) Structured Character-Bound Scripts combine story- and shot-level instructions with a multi-modal character bank. (b) MAniCast generates multi-shot animations with these scripts.

4 METHOD

4.1 STRUCTURED CHARACTER-BOUND SCRIPTS

Common text-controlled methods describe characters and narrative content in a single unstructured prompt, leaving identity-to-role associations implicit and ambiguous in multi-character animation. Structured Character-Bound Scripts separate character definitions from narrative content and use placeholders to retrieve each character's text description or reference image from a shared character bank.

As illustrated in Figure 2(a), each script has three elements:

- **Story-Level Description**: A summary of the overall story across all shots.
- **Shot-Level Instructions**: Provide fine-grained control by dividing each shot into three components:
  - *Foreground Action*: Describes character movements with explicit placeholders (e.g., "[C0] is talking to [C1]"), mapping actions to the correct identities.
  - *Background*: Describes the environmental setting separately from foreground actions, reducing ambiguity between scene attributes and character appearance.
  - *Camera Settings*: Provide structured cinematic controls, including **Shot Type** (e.g., Close-up, Long shot), **Shot Angle** (e.g., High, Low), and **Shot Position** (e.g., Over-the-shoulder, Point-of-view). These standardized attributes specify the intended cinematic composition of each shot.
- **Character Bank**: Stores multi-modal character definitions for casting. Each identity has a unique placeholder (e.g., "[C0]") corresponding to either a reference image or a detailed text description.

This structured script formulates animation creation as character casting and composition, with explicit control over narrative roles and interactions.

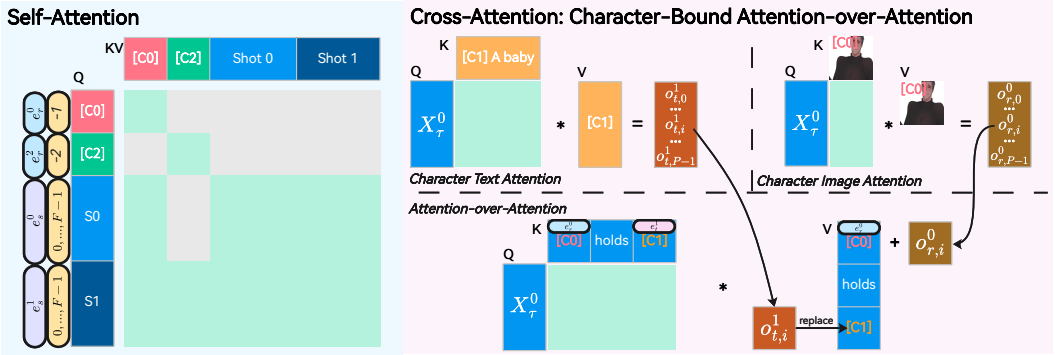

Figure 3: Left: Visual conditioning within self-attention, where green and gray blocks represent unmasked and masked attention, respectively. Right: Character-Bound Attention-over-Attention for incorporating Structured Character-Bound Scripts.

4.2 Framework

As shown in Figure 2(b), MAniCast iteratively denoises multi-shot video latents conditioned on Structured Character-Bound Scripts:

$$\mathbf{X}_0^{0:S-1} = \mathcal{D}_{\boldsymbol{\theta}}\left(\mathbf{X}_1^{0:S-1}, \mathbf{T}^{0:S-1}, \mathcal{C}^{0:N-1}\right) \tag{2}$$

Here $\mathbf{X}_0^{0:S-1}$ denotes denoised video tokens for S shots ($s = 0, \ldots, S-1$), decoded by a VAE into the final animation $\mathbf{I}^{0:S-1}$; $\mathbf{X}_1^{0:S-1}$ is the initial Gaussian noise, and $\mathbf{T}^{0:S-1}$ contains encoded shot instructions. The character bank $\mathcal{C}^{0:N-1}$ contains N identities ($n = 0, \ldots, N-1$), each represented by encoded text tokens $\boldsymbol{C}_{\mathrm{t}}^n$ or VAE-encoded reference tokens $\boldsymbol{C}_{\mathrm{r}}^n$. The denoiser $\mathcal{D}_{\boldsymbol{\theta}}$ is a pretrained DiT with L blocks containing self-attention and cross-attention. We use self-attention to incorporate visual references and Character-Bound Attention-over-Attention (AoA) in cross-attention to route text- or image-defined identities to their narrative placeholders.

4.3 Visual Conditioning within Self-Attention

We concatenate VAE-encoded reference tokens $\boldsymbol{C}_{\mathrm{r}}^n$ with video tokens along the temporal dimension for in-context visual conditioning (Wang et al., 2023). In masked self-attention (Figure 3, left), reference tokens attend only to themselves, while video tokens attend across shots and only to the character references assigned by their shot-level instructions. To distinguish shots and references, each shot's video tokens use Rotary Position Embeddings (Su et al., 2024) (RoPE) at temporal indices $0, \ldots, F-1$ and a learnable shot embedding $\boldsymbol{e}_{\mathrm{s}}^s$. Reference tokens use distinct negative RoPE indices (e.g., -1, -2) and a learnable reference-character embedding $\boldsymbol{e}_{\mathrm{r}}^n$. These learned embeddings are added to the query and key features in each self-attention layer. Masked self-attention gives each shot access to its relevant visual features, but does not bind them to the corresponding textual placeholders; AoA provides that binding in cross-attention.

4.4 Cross-Attention with Character-Bound Attention-over-Attention

Character-Bound Attention-over-Attention (AoA) binds text- or image-defined characters to script placeholders through Character Text Attention, Character Image Attention, and a final routing stage.

Character Text & Image Attention. For each character specified in shot s, we compute its response to the video tokens. Let P and M be the numbers of video and text tokens, and d the attention feature dimension. For a text-defined character n, video tokens $\boldsymbol{X}_\tau^s$ attend to textual features $\boldsymbol{C}_{\mathrm{t}}^n$:

$$\boldsymbol{O}_{\mathrm{t}}^n = \mathrm{CrossAttn}\left(Q_{\mathrm{ca}}(\boldsymbol{X}_\tau^s), K_{\mathrm{ca}}(\boldsymbol{C}_{\mathrm{t}}^n), V_{\mathrm{ca}}(\boldsymbol{C}_{\mathrm{t}}^n)\right) \tag{3}$$

Here, $Q_{\mathrm{ca}}(\cdot)$, $K_{\mathrm{ca}}(\cdot)$, and $V_{\mathrm{ca}}(\cdot)$ denote the query, key, and value projections of the cross-attention layer, respectively, and $\mathrm{CrossAttn}(\cdot)$ denotes scaled dot-product cross-attention. In Figure 3, this

textual branch corresponds to $s = 0$ and $n = 1$ (Character [C1]). For an image-defined character n specified in shot s, the video tokens attend to its visual reference tokens $\boldsymbol{C}_{\mathrm{r}}^{n}$ through self-attention:

$$\boldsymbol{O}_{\mathrm{r}}^{n} = \mathrm{SelfAttn}\left(Q_{\mathrm{sa}}(\boldsymbol{X}_{\tau}^{s}), K_{\mathrm{sa}}(\boldsymbol{C}_{\mathrm{r}}^{n}), V_{\mathrm{sa}}(\boldsymbol{C}_{\mathrm{r}}^{n})\right) \tag{4}$$

Here, $Q_{\mathrm{sa}}(\cdot)$, $K_{\mathrm{sa}}(\cdot)$, and $V_{\mathrm{sa}}(\cdot)$ denote the query, key, and value projections of the self-attention layer, respectively, and $\mathrm{SelfAttn}(\cdot)$ denotes scaled dot-product self-attention. The visual branch in Figure 3 corresponds to $s = 0$ and $n = 0$ (Character [C0]). The two branches produce response matrices $\boldsymbol{O}_{\mathrm{t}}^{n}, \boldsymbol{O}_{\mathrm{r}}^{n} \in \mathbb{R}^{P\times d}$; their i-th rows, $(\boldsymbol{o}_{\mathrm{t},i}^{n})^{\top}$ and $(\boldsymbol{o}_{\mathrm{r},i}^{n})^{\top}$, carry textual and visual identity features for video token i. These features still need to be routed to their corresponding shot-level placeholders.

Character-Bound Attention-over-Attention (AoA). AoA routes these identity features by modifying the values at placeholder positions in standard text cross-attention:

$$\mathrm{CrossAttn}\left(Q_{\mathrm{ca}}(\boldsymbol{X}_{\tau}^{s}), K_{\mathrm{ca}}(\boldsymbol{T}^{s}), V_{\mathrm{ca}}(\boldsymbol{T}^{s})\right) \tag{5}$$

The query-key weights determine how video tokens attend to character placeholders (e.g., [C0]) in $\boldsymbol{T}^{s}$, but the original values contain only generic placeholder semantics, not identity-specific features.

We replicate $V_{\mathrm{ca}}(\boldsymbol{T}^{s}) \in \mathbb{R}^{M\times d}$ across the P video tokens to form a query-specific tensor $\widetilde{\mathbf{V}}^{s} \in \mathbb{R}^{P\times M\times d}$, where $\widetilde{\mathbf{V}}_{i,:,:}^{s} = V_{\mathrm{ca}}(\boldsymbol{T}^{s})$. Here i indexes video tokens and p^{n} is character n's placeholder position. For a text-defined character, we replace the value at p^{n} with $\boldsymbol{o}_{\mathrm{t},i}^{n}$:

$$\widetilde{\mathbf{V}}_{i,p^{n},:}^{s} \leftarrow \boldsymbol{o}_{\mathrm{t},i}^{n}, \qquad i = 0, \ldots, P-1 \tag{6}$$

For [C1] in Figure 3, this inserts Character [C1]'s textual response. Because $\boldsymbol{O}_{\mathrm{t}}^{n}$ is a linear combination of $V_{\mathrm{ca}}(\boldsymbol{C}_{\mathrm{t}}^{n})$ in the same feature space as $V_{\mathrm{ca}}(\boldsymbol{T}^{s})$, the substitution preserves the stability of the original forward pass.

Visual responses occupy a different feature space from $V_{\mathrm{ca}}(\boldsymbol{T}^{s})$, so we project $\boldsymbol{O}_{\mathrm{r}}^{n}$ through a zero-initialized linear layer $\boldsymbol{W}_{0}$ and add the projected features residually rather than replacing the text value. This preserves model stability and the pretrained distribution. For each video token i, we add $\boldsymbol{W}_{0}\boldsymbol{o}_{\mathrm{r},i}^{n}$ at p^{n}:

$$\widetilde{\mathbf{V}}_{i,p^{n},:}^{s} \leftarrow \widetilde{\mathbf{V}}_{i,p^{n},:}^{s} + \boldsymbol{W}_{0}\boldsymbol{o}_{\mathrm{r},i}^{n}, \qquad i = 0, \ldots, P-1 \tag{7}$$

In Figure 3, this augments [C0] with its reference-image features while retaining the original query-key weights and pretrained semantic alignment. We also add learnable modality and identity embeddings (e.g., $\boldsymbol{e}_{\mathrm{r}}^{0}$ in blue and $\boldsymbol{e}_{\mathrm{t}}^{1}$ in red) at the placeholder positions.

5 Implementation Details

For training, we curate a high-quality subset of $105,906$ multi-shot animation videos. To ensure data integrity, we filter out video scenes containing more than 4 shots or 5 characters. We initialize our framework from the pretrained model Wan 2.2 T2V-14B (Wan et al., 2025) and perform full-parameter finetuning, at a resolution of 672×384. Training is conducted for $50,000$ iterations using a global batch size of 16 across 16 GPUs, which takes around 514 hours. We employ the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. The learning rate is set to 5.0×10^{-5}, followed by cosine annealing over $50,000$ steps toward a minimum of 1.0×10^{-6}. At inference time, we solve the probability path via a first-order Euler ODE solver based on the Rectified Flow formulation introduced in Sec. 3. We utilize 50 sampling steps with a shift factor of 6.0 and apply Classifier-Free Guidance (CFG) (Ho & Salimans, 2021) with a scale of 4.0 to balance visual fidelity and prompt adherence.

6 Experiments

6.1 Metrics

VBench Video Quality. We use five VBench (Huang et al., 2024b) dimensions to evaluate motion smoothness, imaging quality, aesthetic quality, and subject and background consistency. Each metric is computed per shot, averaged within each story, and then averaged across stories.

Metric	GT	Baselines					MAniCast (Ours)		
		JoyAI-Echo 1.5	StoryMem	Aura	Phantom-Wan-14B	Kaleido-14B	W/o AoA	Text AoA	Full Method
VBench									
Motion Smooth. ↑	0.9844	0.9864	0.9850	0.9715	0.9842	0.9865	0.9864	0.9872	**0.9884**
Imaging Qual. ↑	0.6819	0.6967	0.6922	0.7103	**0.7352**	0.6932	0.6606	0.6582	0.6735
Aesthetic Qual. ↑	0.5454	0.5877	0.5893	**0.5964**	0.5953	0.5902	0.5790	0.5802	0.5933
Subject Cons. ↑	0.9637	0.9415	0.9473	0.9074	0.9543	0.9585	0.9384	0.9405	**0.9592**
Background Cons. ↑	0.9604	0.9558	0.9602	0.9400	0.9635	**0.9656**	0.9595	0.9621	0.9638
Task-Specific Metrics									
Casting F1 ↑	0.9850	0.9076	0.9013	0.9164	0.8896	0.8941	0.9010	0.9191	**0.9560**
ID-bound Event Acc. ↑	0.9420	0.8287	0.8505	0.7912	0.7818	0.7917	0.8188	0.8601	**0.9125**
Cross-shot ID Succ. ↑	0.7503	0.6873	0.6633	0.6277	0.5825	0.5959	0.6839	0.6813	**0.7293**
Text Attr. Fidelity ↑	0.9609	0.9588	**0.9600**	0.8715	0.9255	0.9228	0.8938	0.9425	0.9519
Matched Ref. Sim. ↑	0.5844	0.5296	0.5857	0.6203	0.6367	**0.6403**	0.5291	0.5279	0.5850

Table 1: Quantitative comparison with 5 baseline methods and ablation study on Character-Bound Attention-over-Attention, across 1,000 multi-shot animation sequences.

Task-Specific Metrics. To evaluate MAniCast's performance on character casting, we introduce five new task-specific metrics. We use Grounding DINO (Liu et al., 2024) to detect character candidates in generated frames. We then use Qwen3.8-27B (Qwen Team, 2026) to extract the required characters, actions, and attributes from the input script and character bank and judge whether the generated video meets these requirements. Rather than asking the VLM to score the video directly, we use targeted identity matching, pairwise cross-shot comparisons, and attribute judgments, then compute the metrics by fixed rules to reduce the risk of hallucinated scores. **Shot Casting F1** measures how accurately the characters specified in the script are cast in each shot, penalizing missing or incorrectly included characters; **Identity-bound Event Accuracy** is the fraction of observable events in which the correct character performs the required action; **Cross-shot Identity Success** is the fraction of required shot pairs in which a character retains the same identity; **Text Attribute Fidelity** measures compliance with appearance attributes (e.g., hair color, clothing, and accessories) of text-defined characters; **Matched Reference Similarity** computes DINOv2 (Oquab et al., 2024) cosine similarity between generated-character crops and corresponding reference-character crops after their identities have been confirmed.

6.2 Comparisons on Multi-Shot Animation Creation

To evaluate MAniCast, we compare its performance on a test set of 1,000 multi-shot animation sequences, ensuring no overlap with the animation episodes or movies in the training data. The baselines comprise multi-shot methods without character placeholders (JoyAI-Echo 1.5 (Duan et al., 2026), StoryMem (Zhang et al., 2025a)), single-shot methods without placeholders (Phantom-Wan-14B (Liu et al., 2025), Kaleido-14B (Zhang et al., 2025b)), and a single-shot method with placeholders (Aura (Zhou et al., 2026b)).

Quantitatively, Table 1 shows that MAniCast achieves the highest Shot Casting F1 (0.9560), Identity-bound Event Accuracy (0.9125), and Cross-shot Identity Success (0.7293) among generated methods. These results indicate more reliable assignment of the required characters to shots and actions, together with stronger identity preservation across shots. Qualitatively, Figure 4 reveals two recurring problems. First, generated actions can be inaccurate even when the correct character appears. For instance, in the second example, Shot 2 requires [C2] to stand near the table and then sit down, yet none of the baseline results captures this transition. Second, baselines can omit required characters or generate incorrect attributes; in the third example, required characters are missing from baseline outputs in both shots. Compared with existing baselines, AoA routes text- and image-defined identity features to their narrative placeholders, enabling MAniCast to cast the specified characters more accurately and bind each one to its intended actions across shots.

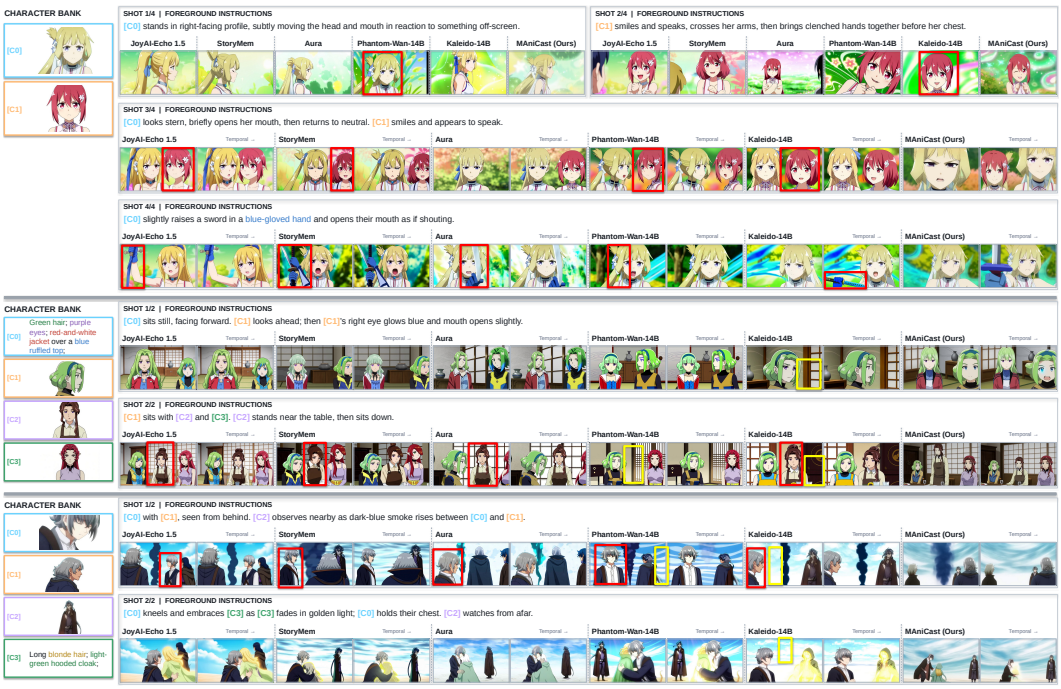

Figure 4: Qualitative comparison of six methods on three multi-shot animation cases. Red boxes highlight errors in character positioning or actions, while yellow boxes indicate missing characters or attribute errors.

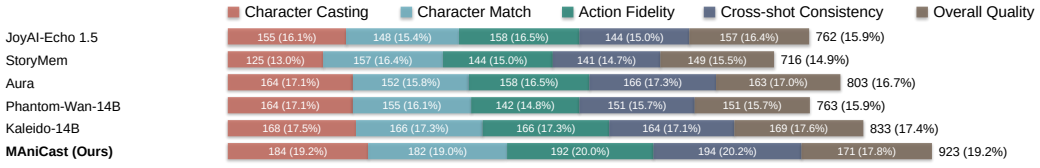

Figure 5: User-study preference results across five evaluation criteria.

6.3 User Study

We collected 120 completed questionnaires, each covering eight randomly selected video cases from a pool of 25 videos. For each video case, participants made five separate choices, selecting the best result among six methods once for each criterion: character casting, character match (with character text or reference image), action fidelity, cross-shot consistency, and overall quality. As shown in Figure 5, MAniCast receives the largest share of selections on all five criteria.

6.4 Ablation Study on Character-Bound Attention-over-Attention

To isolate the contributions of Text and Image AoA, we compare variants without AoA, with Text AoA alone, and with both components (Full AoA). Quantitatively, as shown in Table 1, Text AoA improves Shot Casting F1 from 0.9010 to 0.9191, Identity-bound Event Accuracy from 0.8188 to 0.8601, and Text Attribute Fidelity from 0.8938 to 0.9425, indicating more accurate binding of text-defined characters to their roles and actions. Adding Image AoA further raises these scores to 0.9560, 0.9125, and 0.9519, respectively, while improving Matched Reference Similarity from 0.5279 to 0.5850 and Cross-shot Identity Success from 0.6813 to 0.7293. Qualitatively, Figure 6 illustrates how the two components contribute. In the left example, Text AoA better reproduces the text-defined [C1]'s white mask and golden armor and helmet than the variant without AoA. The right example combines text-defined [C0] and [C1] with image-defined [C2]. Text AoA alone does not assign all three characters to their specified roles, whereas Full AoA shows [C0] resting

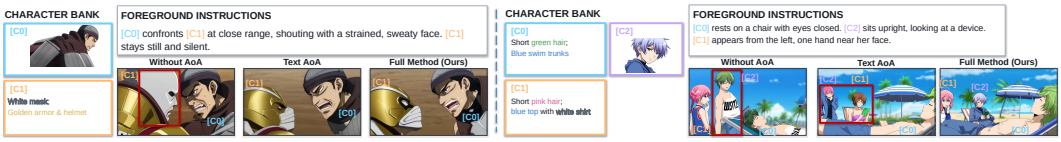

Figure 6: Qualitative ablation of Character-Bound AoA. The left example illustrates adherence to a text-defined character; the right example illustrates casting with mixed text- and image-defined characters.



Figure 7: Animation scenes that challenge automatic character annotation. From left to right: (1) similarly dressed students, (2) backlit group, and (3) small characters with limited visible identity cues.

in the chair, [C2] sitting with a device, and [C1] on the left with one hand near her face, while preserving their distinct appearances without attribute leakage. Together, these results show that Text AoA strengthens adherence to textual character definitions, while combining it with Image AoA enables accurate casting of mixed text- and image-defined characters into their designated actions and positions.

7 Conclusion

We presented *MAniCast*, the first framework to combine modality-unified character casting with explicit identity-to-role binding for multi-shot animation generation. Structured Character-Bound Scripts organize story- and shot-level instructions with a shared character bank in which each identity is defined by either a detailed text description or a reference image, while explicit placeholders assign selected identities to narrative roles and actions across shots. Character-Bound Attention-over-Attention realizes these assignments at the network level by deriving textual or visual identity features and injecting them into the corresponding narrative placeholders. We also introduced *MAniScript*, a multi-shot animation dataset annotated with Structured Character-Bound Scripts, each of which contains a character bank of text- or image-defined identities. Experiments demonstrate that MAniCast outperforms prior text-controlled and reference-based baselines in casting text- and image-defined characters, executing identity-bound actions, and generating complex interactions across shots.

Limitations & Future Work. Firstly, due to training resource constraints, we trained MAniCast on scenes with at most four shots, leaving longer narratives untested. Secondly, our automatic annotation pipeline can be less reliable in crowded scenes, e.g., with more than five characters. As shown in Fig. 7, such scenes often contain groups of people with similar appearances or small character instances that are difficult to distinguish, complicating identity tracking and character binding. To address these limitations, we plan to extend training and evaluation to longer narratives and incorporate video tracking results as additional spatial cues to help the VLM identify characters more reliably in crowded scenes.

Beyond these immediate improvements, Structured Character-Bound Scripts and AoA can be naturally extended from animation creation to general video generation, a direction we plan to explore in future work. Realizing this extension would require larger and more diverse video datasets and a more general-purpose annotation pipeline.

AI USE STATEMENT

We used VLM to assist with MAniScript annotation, including scene grouping, character identity matching, and structured caption generation from existing animation videos. An image-generation model was used only to repaint character reference images. We used LLM to discuss methodological and experimental choices, identify and summarize related work, suggest and edit code, and draft and revise the manuscript. A VLM was additionally used for automated evaluation, as described in the Experiments section. The authors have checked all AI-generated content and take responsibility for the final dataset, implementation, results, and manuscript, including all AI-assisted content.

REFERENCES

Yuval Atzmon, Rinon Gal, Yoad Tewel, Yoni Kasten, and Gal Chechik. Multi-shot character consistency for text-to-video generation, 2025. URL https://openreview.net/forum?id=0zRuk3QdiH.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *International Conference on Learning Representations*, 2026.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, and Junhao Zhuang. Long-horizon audio-visual generation for persistent stories and interactive worlds, 2026. URL https://arxiv.org/abs/2608.23383.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, and Yahui Zhou. SkyReels-A2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation, 2025a. URL https://arxiv.org/abs/2505.04512.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Jiangning Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. PolyVivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. In *Advances in Neural Information Processing Systems*, 2025b.

Binyuan Huang, Yuning Lu, Weinan Jia, Hualiang Wang, Mu Liu, and Daiqing Yang. Rethinking position embedding as a context controller for multi-reference and multi-shot video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yubo Huang, Weiqiang Wang, Sirui Zhao, Tong Xu, Lin Liu, and Enhong Chen. Bind-Your-Avatar: Multi-talking-character video generation with dynamic 3d-mask-based embedding router. *arXiv preprint arXiv:2506.19833*, 2025a.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning, 2025b. URL https://arxiv.org/abs/2501.04698.

Zhitong Huang, Mohan Zhang, and Jing Liao. LVCD: Reference-based lineart video colorization with diffusion models. *ACM Transactions on Graphics*, 43(6):1–11, 2024a.

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*, pp. 21807–21818, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. ShotAdapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Lingen Li, Guangzhi Wang, Zhaoyang Zhang, Yaowei Li, Xiaoyu Li, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. ToonComposer: Streamlining cartoon production with generative post-keyframing. In *International Conference on Learning Representations*, 2026.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Movie Weaver: Tuning-free multi-concept video personalization with anchored prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII*, pp. 38–55, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72969-0. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yihao Meng, Hao Ouyang, Hanlin Wang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Zhiheng Liu, Yujun Shen, and Huamin Qu. AniDoc: Animation creation made easier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18187–18197, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Lu Qiu, Yizhuo Li, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Animeshooter: A multi-shot animation dataset for reference-guided video generation, 2025. URL https://arxiv.org/abs/2506.03126.

Qwen Team. Qwen3.8-27B, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Tencent Hunyuan Team. Script-a-video: Deep structured audio-visual captions via factorized streams and relational grounding, 2026. URL https://arxiv.org/abs/2604.11244.

Jente Vandersanden, Matheus Gadelha, Chun-Hao P. Huang, Hyeonho Jeong, and Yulia Gryaditskaya. EM-Vid: Training-free entity-centric memory for efficient and consistent multi-shot video generation, 2026. URL https://arxiv.org/abs/2605.23610.

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.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. MultiShotMaster: A controllable multi-shot video generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Chuanzhi Xu, Huiqi Liang, Bang Shi, Huiming Zhang, Yifan Xiao, Guangcheng Lin, Haodong Chen, Qiang Qu, Zhicheng Lu, and Weidong Cai. DrawVideo: Generating long video from storyboard keyframe sketches. *arXiv preprint arXiv:2605.23508*, 2026.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory, 2025a. URL https://arxiv.org/abs/2512.19539.

Zhenxing Zhang, Jiayan Teng, Zhuoyi Yang, Tiankun Cao, Cheng Wang, Xiaotao Gu, Jie Tang, Dan Guo, and Meng Wang. Kaleido: Open-sourced multi-subject reference video generation model. *arXiv preprint arXiv:2510.18573*, 2025b.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention, 2025. URL https://arxiv.org/abs/2412.02259.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 110315–110340. Curran Associates, Inc., 2024. doi: 10.52202/079017-3501. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c7138635035501eb71b0adf6ddc319d6-Paper-Conference.pdf.

Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam Woh Ng, Tian Xie, Xiao Han, Yuren Cong, Hang Li, Chuyan Zhu, Aditya Patel, Tao Xiang, and Sen He. Scaling zero-shot reference-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Zixiang Zhou, Zhentao Yu, Yifeng Ma, Hongmei Wang, Wenqing Yu, Cong Wang, Zilin Yang, Rui Chen, Jiarong Ou, Yezhou Liu, Yuan Zhou, and Qinglin Lu. Aura: Consistent multi-subject video generation via vlm-grounded semantic alignment, 2026b. URL https://arxiv.org/abs/2607.04311.