CINEREF: CINEMATIC REFERENCE-GUIDED AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cinematic reference-guided audio-video generation aims to synthesize high-quality videos conditioned on reference images and audios under complex cinematic scenarios. However, current open-source methods primarily treat reference information as passive conditioning, making reference injection fragile under large camera motions and frequent shot transitions. To address this, we propose **CineRef**, a unified framework for cinematic reference-guided audio-video generation. The key insight is that reference information should be explicitly linked to the generated video during training and consistently preserved throughout the entire clip, rather than treated as passive and transient conditioning. Based on this insight, we first propose **Dual-Anchor Reference Alignment**, which combines a reference ID anchor for explicit ID supervision with a face-aware content anchor for preserving natural facial content. We then introduce **Reference Trajectory Consistency** to maintain long-range ID coherence across substantial viewpoint and scale variations along ID-wise trajectory. Finally, we design **Audio-Visual Reference Binding** to establish explicit correspondence between each visual ID and its corresponding audio reference. Extensive experiments on our proposed **CineSet** for audio-video generation demonstrate that CineRef achieves superior ID injection and preservation under diverse camera motions and shot transitions.

1 INTRODUCTION

Recent advances in video generation have enabled high-quality synthesis from multimodal conditions such as text, images, and audio (Brooks et al., 2024; Kong et al., 2024; Seedance et al., 2026; MiniMax, 2026; HaCohen et al., 2026; Jiang et al., 2025). In particular, reference-guided audio-video generation must balance faithful reference conditioning with smooth, high-quality synthesis. Existing methods (Liu et al., 2025; Hu et al., 2025; Chen et al., 2025a; Jiang et al., 2025; Guo et al., 2026) mainly incorporate references through latent concatenation or attention mechanisms, leaving reference injection largely implicit. Under large camera motions and frequent shot transitions, such implicit reference modeling is prone to generation failures such as ID drift, unreliable visual-audio correspondence, and copy-and-paste artifacts (Guo et al., 2026; Chen et al., 2026c).

We argue that robust cinematic generation requires reference information to be explicitly modeled and consistently preserved throughout the video clip, rather than simply regarded as conditioning signals. Without explicit modeling, the weak connection between references and generated content makes reference information difficult to preserve under complex cinematic dynamics. Based on this insight, we present **CineRef**, a unified framework built upon two principles: **explicit reference supervision** and **long-range reference preservation**. A straightforward approach to explicit reference supervision is to directly minimize the ID distance between the reference image and generated frames using a pretrained face recognition model such as ArcFace (Deng et al., 2019). However, naive ID supervision introduces a face-scale bias: the model tends to reduce the loss by enlarging generated faces to falsely facilitate ID matching, rather than preserving the reference ID under natural facial scales. This optimization shortcut results in oversized faces and degraded content fidelity. To address this issue, we propose **Dual-Anchor Reference Alignment**, which combines two complementary anchors. The reference ID anchor directly enforces ID fidelity, while the content anchor applies additional face-aware flow-matching supervision over ground-truth facial regions to preserve natural facial content and spatial distribution.

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.



(a) CineRef supports orbiting camera movements while effectively injecting character ID.



(b) CineRef follows shot transitions and preserves ID across shot transitions.

Figure 1: Reference-guided generation comparison under camera motions and shot transitions.

Explicit reference supervision alone is insufficient for cinematic videos, where large camera motions, viewpoint changes, scale variations, and shot transitions can introduce long-range ID drift. To mitigate this issue, we introduce **Reference Trajectory Consistency**, which constructs ID-wise face trajectories with diverse facial-scale variations and enforces overall ID consistency across the trajectory. Such trajectory-level supervision strengthens ID coherence across large visual variations and extends reference preservation from individual frames to the clip level. Cinematic audio-video generation also requires reliable correspondence among textual, visual, and audio conditions. We therefore design **Audio-Visual Reference Binding**, using RoPE-based alignment for visual-audio binding and paired VLM conditioning for visual-textual binding. In the VAE branch, each reference audio is equipped with a learnable audio marker whose temporal RoPE is aligned with that of the corresponding visual reference, thereby strengthening visual-audio reference binding. In parallel, the VLM branch explicitly associates reference appearance with the corresponding character tag, reinforcing the connection between reference ID and textual content. Extensive experiments under cinematic scenarios demonstrate that CineRef achieves superior performance in ID consistency and visual-audio correspondence. As shown in Figure 1, CineRef exhibits robustness under large camera motions and frequent shot transitions. Our main contributions are summarized as follows:

- We reveal that explicit clip-level reference modeling, rather than passive reference conditioning, is essential for robust cinematic reference-guided audio-video generation under large camera motions and frequent shot transitions.
- We propose **Dual-Anchor Reference Alignment**, which combines explicit ID supervision with face-aware content supervision to jointly improve ID fidelity and content fidelity, together with **Reference Trajectory Consistency** for long-range, clip-level ID preservation.
- We introduce **Audio-Visual Reference Binding**, where the VAE branch strengthens visual-audio reference binding through aligned temporal RoPE with audio markers, while the VLM branch reinforces the association between visual references and textual descriptions.
- In addition to the conventional **OpenS2V** benchmark, we further propose **CineSet** to evaluate reference-guided audio-video generation with diverse camera motions and shot transitions. Extensive experiments on both benchmarks demonstrate consistent improvements over state-of-the-art methods in ID consistency and visual-audio correspondence.

2 RELATED WORK

Cinematic Audio-Visual Generation Recent generative models have evolved from silent video synthesis toward unified audio-visual generation. Proprietary systems such as the Seedance series (Seedance et al., 2026; ByteDance Seed, 2026) and Kling series (Team et al., 2025) demonstrate strong cinematic capabilities, including synchronized audio, multimodal reference control, complex camera motion, and multi-shot storytelling, but their closed-source nature limits systematic study. Meanwhile, open-source models have made rapid progress. JavisDiT++ (Liu et al., 2026a) improves cross-modal synchronization through modality-specific modeling and temporally aligned positional encoding. LTX-2 (HaCohen et al., 2026) adopts asymmetric audio-video streams with bidirectional interaction, while SkyReels-V4 (Chen et al., 2026a) extends joint generation toward multimodal conditioning. MAVIN (Liu et al., 2026b) and UnityShots (Huang et al., 2026) further explore multi-shot generation through ID-aware propagation and cross-shot memory. Despite these advances, preserving reference-specific information under large camera motions and frequent shot transitions remains challenging. CineRef addresses this problem by explicitly modeling and preserving visual and audio references throughout the generated clip.

Reference-Guided Video Generation Reference-guided video generation aims to preserve reference-specific information while following novel textual, spatial, motion, and audio conditions. Phantom (Liu et al., 2025) supports single- and multi-subject generation through cross-modal alignment, while VACE (Jiang et al., 2025) incorporates reference-to-video generation into a unified creation and editing framework. HuMo (Chen et al., 2025a) jointly models text, reference images, and audio for human-centric generation, and MAGREF (Deng et al., 2026) extends reference control to arbitrary and multiple subjects through region-aware guidance and reference feature injection. However, strong reference preservation can introduce content entanglement and copy-and-paste artifacts. Phantom-Data (Chen et al., 2026c) analyzes this problem, while Saber (Zhou et al., 2025) uses masked training and augmentation to mitigate such artifacts. More recently, RefAlign (Wang et al., 2026a) aligns reference representations with a visual foundation model to improve subject consistency and reduce copy-and-paste artifacts. In contrast, CineRef explicitly supervises reference information during training and preserves it throughout the generated clip, balancing reference fidelity with natural content generation.

Audio-Visual Reference Binding Reference-guided audio-video generation further requires reliable correspondence between visual IDs and their associated audio references. HunyuanVideo-Avatar (Chen et al., 2025b) introduces a face-aware audio adapter to isolate character-specific audio conditions, while InterActHuman (Wang et al., 2026b) establishes region-specific multimodal binding by injecting local audio conditions into corresponding identity regions. DreamID-Omni (Guo et al., 2026) further addresses ID-timbre mismatch and speaker confusion through synchronized RoPE and structured captions. CineRef instead introduces Audio-Visual Reference Binding with two complementary branches: the VAE branch aligns audio markers with their corresponding visual references through RoPE, while the VLM branch associates visual references with their corresponding textual descriptions. Together, they establish explicit multimodal reference correspondence for robust generation under complex cinematic dynamics.

3 CINEREF

3.1 PROBLEM FORMULATION

Given visual and audio reference latents $\{\mathbf{z}_i^{\text{ref}}\}_{i=1}^{N}$ and $\{\mathbf{a}_i^{\text{ref}}\}_{i=1}^{M}$, reference-guided audio-video generation aims to jointly predict the video and audio velocity fields from textual latents $\mathbf{c}$, noisy latents $\mathbf{z}_t$, and $\mathbf{a}_t$, where $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_T$ and $\mathbf{a}_t = (1-t)\mathbf{a}_0 + t\mathbf{a}_T$ denote the interpolated video and audio latents at timestep t. The generation process is formulated as

$$\{\hat{\mathbf{v}}_t^{\text{z}}, \hat{\mathbf{v}}_t^{\text{a}}\} = \text{CineRef}_\theta\left(\mathbf{z}_t, \mathbf{a}_t \,\middle|\, \{\mathbf{z}_i^{\text{ref}}\}_{i=1}^{N}, \{\mathbf{a}_i^{\text{ref}}\}_{i=1}^{M}, c, t\right), t \in [0,1]. \tag{1}$$

The standard flow-matching objective is

$$\mathcal{L}_{\text{fm}} = \|\hat{\mathbf{v}}_t^{\text{z}} - \mathbf{u}_t^{\text{z}}\|_2^2 + \|\hat{\mathbf{v}}_t^{\text{a}} - \mathbf{u}_t^{\text{a}}\|_2^2, \tag{2}$$

where $\mathbf{u}_t^{\text{z}} = \mathbf{z}_T - \mathbf{z}_0$ and $\mathbf{u}_t^{\text{a}} = \mathbf{a}_T - \mathbf{a}_0$.

3.2 APPROACH OVERVIEW

This paper aims to generate cinematic reference-guided audio-video content under challenging conditions, including large camera motions and frequent shot transitions. Different from conventional

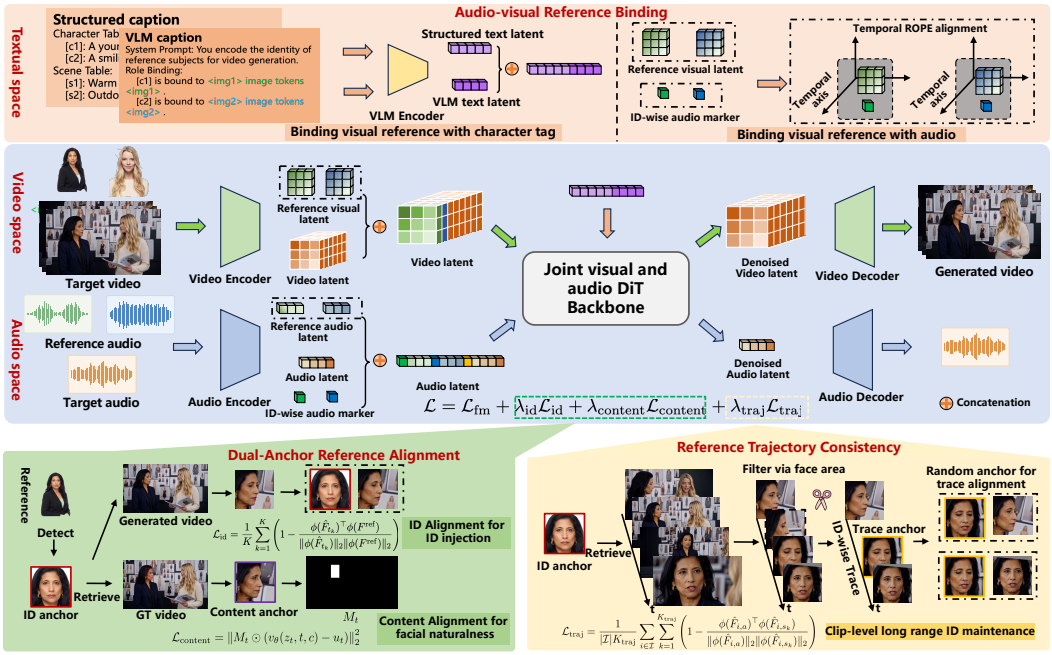


Figure 2: Overview of CineRef. Given visual references, reference audios, and structured conditions, CineRef first establishes multimodal reference correspondence through **Audio-Visual Reference Binding**: the VLM branch pairs visual references with character tags, while the VAE branch aligns learnable audio markers with the corresponding visual references via temporal RoPE. Then, **Dual-Anchor Reference Alignment** combines an ID anchor from generated frames with a content anchor derived from ground-truth facial regions, explicitly enforcing ID fidelity while preserving natural facial content and scale. **Reference Trajectory Consistency** further constructs ID-wise trajectories, samples diverse observations with substantial facial-scale variations, and enforces overall ID consistency across each trajectory for long-range, clip-level reference preservation.

approaches that treat reference images as passive conditions, we argue that reference information should be *explicitly supervised and continuously preserved throughout the generated video clip*. Based on this principle, we propose **CineRef**, a unified framework for cinematic reference-guided audio-video generation. As illustrated in Fig. 2, CineRef forms a coherent reference modeling framework that explicitly constrains reference fidelity, preserves ID information over long-range cinematic variations, and establishes reliable correspondence across visual and audio references.

3.3 Dual-Anchor Reference Alignment

One key challenge in reference-guided generation is to preserve the reference ID without compromising the naturalness of the generated content. A straightforward way to strengthen ID fidelity is to explicitly minimize the ID distance between generated frames and the reference image using a pretrained face recognition model. Given a reference image I^{ref} and a generated video $\hat{V}$, we uniformly sample K frames along the temporal axis and perform face detection and alignment using RetinaFace. We then extract ID embeddings from the sampled generated faces and the reference face using a pretrained ArcFace encoder $\phi(\cdot)$. The reference ID supervision is formulated as

$$\mathcal{L}_{\text{id}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{\phi(\hat{F}_{s_k})^{\top}\phi(F^{\text{ref}})}{\|\phi(\hat{F}_{s_k})\|_2\|\phi(F^{\text{ref}})\|_2}\right), \tag{3}$$

where $\{s_k\}_{k=1}^{K}$ denotes the uniformly sampled frame indices, F^{ref} is the aligned face crop from the reference image, and $\hat{F}_{s_k}$ is the aligned face crop from the generated frame at s_k. However, we observe that directly optimizing $\mathcal{L}_{\text{id}}$ introduces an undesirable *face-scale bias*, as presented in Figure 3. Since the ID loss is evaluated only on detected and aligned face crops, it is insensitive to whether the generated face occupies an unnaturally large region in the full frame. This creates an



Figure 3: Face bias in the training step using only ID anchor.

unintended optimization shortcut, where the model can reduce $\mathcal{L}_{\text{id}}$ by enlarging the generated face to facilitate ID matching, rather than learning to preserve the reference ID under natural variations in facial scale. As training progresses, this bias manifests as systematically oversized faces, reduced expression diversity, and degraded content fidelity.

To address this issue, we propose **Dual-Anchor Reference Alignment**, which introduces two complementary anchors: a **reference ID anchor** and a **content anchor**. The reference ID anchor determines *who* to generate by enforcing similarity to the reference ID, while the content anchor constrains *how* the ID should naturally appear in the target video. Specifically, we construct the content anchor from ground-truth facial regions and introduce additional face-aware supervision into the flow-matching objective. For each ground-truth frame I_s, we detect the corresponding face region and construct a face-aware mask M_s. The frame-wise masks are combined into a video-level face-aware mask M, which is used to define the content supervision:

$$\mathcal{L}_{\text{content}} = \|M \odot (v_\theta(\mathbf{z}_\tau, \tau, \mathbf{c}) - \mathbf{u}_\tau)\|_2^2, \tag{4}$$

where $\mathbf{z}_\tau$ denotes the noisy video latent at flow-matching timestep τ, $\mathbf{c}$ denotes the conditioning information, $v_\theta(\cdot)$ is the predicted flow, and $\mathbf{u}_\tau$ is the corresponding flow-matching target. Since M is derived from the ground-truth facial regions, $\mathcal{L}_{\text{content}}$ places additional supervision on facial content within its natural spatial extent, counteracting the face-enlargement shortcut induced by the ID objective. The two anchors provide complementary supervision: the reference ID anchor encourages strong ID fidelity, while the content anchor reinforces facial content at its natural location and scale. Together, they preserve the reference ID without sacrificing natural facial scale, pose, expression, or overall content fidelity.

3.4 Reference Trajectory Consistency

Explicit frame-level supervision alone is insufficient for cinematic videos, where the same subject may undergo substantial appearance variations due to camera motion, viewpoint changes, facial-scale variations, and shot transitions. Although Dual-Anchor Reference Alignment improves reference fidelity at the frame level, it does not explicitly enforce ID consistency across substantially different observations of the same subject. Consequently, large visual changes throughout the clip can still lead to long-range ID drift. We introduce **Reference Trajectory Consistency**, which explicitly enforces ID consistency across challenging observations along each subject trajectory. Given a generated video, we first detect and track each target ID across all valid frames. For the i-th reference ID, we define its ID-wise trajectory as $\mathcal{T}_i = \left\{\hat{F}_{i,s}\right\}_{s\in\mathcal{V}_i}$, where $\mathcal{V}_i$ denotes the set of valid frames in which the i-th ID is successfully detected and tracked, and $\hat{F}_{i,s}$ is the aligned face crop of ID i at frame s. Rather than uniformly enforcing consistency over all observations in $\mathcal{T}_i$, we focus on trajectory points with substantial facial-scale changes, which are more challenging for ID preservation. We first select an observation at frame a as the ID anchor. For each valid observation, we obtain its detected face area $A_{i,s}$ and measure the relative variation with respect to the anchor: $\Delta A_{i,s} = \frac{|A_{i,s}-A_{i,a}|}{A_{i,a}}$. We then construct a candidate set by retaining observations whose facial-area variation exceeds a predefined threshold τ_A: $\mathcal{C}_{i,a} = \{s \in \mathcal{V}_i \mid \Delta A_{i,s} > \tau_A\}$. This filtering removes redundant observations with similar facial scales and emphasizes trajectory points exhibiting substantial visual variations. From the candidate set, we sample K_{traj} observations, $\mathcal{S}_{i,a} = \{s_1, s_2, \ldots, s_{K_{\text{traj}}}\} \subseteq \mathcal{C}_{i,a}$. The trajectory consistency loss is then defined as

$$\mathcal{L}_{\text{traj}} = \frac{1}{|\mathcal{I}|K_{\text{traj}}} \sum_{i\in\mathcal{I}} \sum_{k=1}^{K_{\text{traj}}} \left(1 - \frac{\phi(\hat{F}_{i,a})^\top \phi(\hat{F}_{i,s_k})}{\|\phi(\hat{F}_{i,a})\|_2 \|\phi(\hat{F}_{i,s_k})\|_2}\right), \tag{5}$$

where $\mathcal{I}$ denotes the set of valid reference IDs and $\phi(\cdot)$ is the pretrained ArcFace encoder.

Unlike frame-wise ID supervision, Reference Trajectory Consistency explicitly couples challenging observations within each ID-wise trajectory. By aligning observations with substantial facial-scale variations to a shared ID anchor, the proposed objective strengthens ID coherence across large visual changes and extends reference preservation from individual frames to the clip level.

Combining the original flow-matching objective with Dual-Anchor Reference Alignment and Reference Trajectory Consistency, the overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{fm}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{content}}\mathcal{L}_{\text{content}} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}}, \tag{6}$$

where λ_{id}, λ_{content}, and λ_{traj} control the strengths of the three reference supervision terms.

3.5 Audio-Visual Reference Binding

As formulated in Sec. 3.1, CineRef conditions generation on visual references $\{\mathbf{z}_i^{\text{ref}}\}_{i=1}^{N}$ and audio references $\{\mathbf{a}_i^{\text{ref}}\}_{i=1}^{M}$. Since the two modalities may contain different numbers of references, their correspondence can become ambiguous, especially when the certain visual reference lack paired audio reference. We therefore pad missing audio slots with zero-valued inputs to form N aligned visual-audio reference slots, providing an explicit slot-wise correspondence for subsequent reference binding. The reference latents are then concatenated with the corresponding noisy latents:

$$\{\hat{\mathbf{v}}_t^{\text{z}}, \hat{\mathbf{v}}_t^{\text{a}}\} = \text{CineRef}_\theta\left([\mathbf{z}_1^{\text{ref}}, \ldots, \mathbf{z}_N^{\text{ref}}, \mathbf{z}_t], [\mathbf{a}_1^{\text{ref}}, \ldots, \mathbf{a}_N^{\text{ref}}, \mathbf{a}_t] | c, t\right), t \in [0, 1]. \tag{7}$$

A key challenge is to establish reliable correspondence among visual references, audio references, and semantic conditions. Existing structured captions can associate an audio reference with a character tag and further connect the tag with the corresponding linguistic content. However, such binding remains largely implicit and unreliable under large camera motions and substantial visual changes. To address this, we propose **Audio-Visual Reference Binding** to explicitly reinforce two complementary associations. The VLM branch binds each visual reference to its corresponding character tag via structured caption c_s and VLM caption c_v. By jointly encoding the reference appearance and character tag, the semantic representation explicitly captures which visual ID corresponds to each character, thereby strengthening the association between reference appearance and generated content. The VAE branch directly binds each visual reference to its corresponding audio reference. For each reference audio $\mathbf{a}_i^{\text{ref}}$, we prepend a learnable audio marker $\mathbf{m}_i$ and align its temporal RoPE coordinate with that of the corresponding visual reference. Since the marker is attached to the corresponding audio segment, this alignment establishes an explicit association between the visual and audio references. Together, the VLM branch strengthens visual-reference-to-character binding, while the VAE branch strengthens visual-to-audio reference binding, resulting in reliable reference blinding. The final generation process is formulated as follows:

$$\{\hat{\mathbf{v}}_t^{\text{z}}, \hat{\mathbf{v}}_t^{\text{a}}\} = \text{CineRef}_\theta\left([\mathbf{z}_1^{\text{ref}}, \ldots, \mathbf{z}_N^{\text{ref}}, \mathbf{z}_t], [\mathbf{a}_1^{\text{marker}}, \mathbf{a}_1^{\text{ref}}, \ldots, \mathbf{a}_N^{\text{marker}}, \mathbf{a}_N^{\text{ref}}, \mathbf{a}_t] | [\mathbf{c}_\mathbf{s}, \mathbf{c}_\mathbf{v}], t\right). \tag{8}$$

4 Experiments

4.1 Experiment Setup

Datasets & Evaluation Metrics We evaluate CineRef on both the conventional reference-guided video generation benchmark OpenS2V and our proposed CineSet. OpenS2V evaluates general reference-guided generation quality, while CineSet focuses on audio-video generation under challenging cinematic dynamics with diverse camera motions and shot transitions. CineSet contains 100 samples, including 60 single-subject and 40 multi-subject samples, with 140 reference images and 140 reference audios in total. For each sample, we construct both **single-shot** and **multi-shot** variants. The multi-shot variant contains three 5-second shots separated by hard cuts, while the single-shot variant contains one continuous 10-second shot. Both variants include diverse camera motions. All videos are generated at 1344×768 resolution and 24 FPS. More details are provided in Appendix A.3. On CineSet, we evaluate reference-guided audio-video generation from multiple complementary aspects. We use **FaceSim** to measure visual ID preservation across generated frames and **SpkSim** to evaluate the similarity between generated and reference speaker timbres. We further report **Sync-C** and **LSE-D** to assess audio-visual synchronization, where higher Sync-C indicates more confident lip-audio alignment and lower LSE-D indicates a smaller synchronization distance.

Table 1: Subject-driven video generation comparison on the OpenS2V benchmark.

Method	TotalScore↑	Aesthetics↑	MotionSmoothness↑	MotionAmplitude↑	FaceSim-Cur↑	GmeScore↑	NexusScore↑	NaturalScore↑
				Closed-source				
Vidu2.0	51.95%	41.48%	90.45%	13.52%	35.11%	67.57%	43.37%	65.88%
Pika2.1	51.88%	46.88%	87.06%	24.71%	30.38%	69.19%	45.40%	63.32%
Kling1.6	56.23%	44.59%	86.93%	**41.60%**	40.10%	66.20%	45.89%	**74.59%**
Saber-14B	57.91%	42.42%	96.12%	21.12%	49.89%	67.50%	**47.22%**	72.55%
				Open-source				
VACE-14B	57.55%	**47.21%**	94.97%	15.02%	55.09%	67.27%	44.08%	67.04%
SkyReels-A2-P14B	52.25%	39.41%	87.93%	25.60%	45.95%	64.54%	43.75%	60.32%
Phantom-14B	56.77%	46.39%	96.31%	**33.42%**	51.46%	**70.65%**	37.43%	69.35%
MAGREF-480P	52.51%	45.02%	93.17%	21.81%	30.83%	**70.47%**	43.04%	66.90%
VINO	57.85%	45.92%	94.73%	12.30%	52.00%	69.69%	42.67%	71.99%
Kaleido	56.59%	46.15%	**98.11%**	12.67%	34.56%	67.51%	41.70%	**82.18%**
BindWeave	57.61%	45.55%	95.90%	13.91%	53.71%	67.79%	46.84%	66.85%
RefAlign-14B	**60.42%**	46.84%	**97.61%**	22.48%	**55.23%**	68.32%	**48.52%**	73.63%
LTX-Ingredients	43.33%	30.67%	93.96%	20.74%	19.76%	59.06%	36.60%	58.39%
CineRef	**61.40%**	**48.34%**	90.83%	14.77%	**74.13%**	65.04%	44.53%	68.32%

Implementation Details We use the publicly available LTX-2.3 [1] as our base audio-video generation model and fine-tune both the video and audio branches with LoRA. The LoRA rank and scaling factor are both set to 384. We use AdamW with a learning rate of 1×10^{-4}. We adopt multi-resolution training at 360p, 480p, 540p, 720p, and 960p, with all videos sampled at 24 FPS. Each reference audio is cropped or padded to 5 seconds. Reference images are randomly resized during training to improve robustness to different reference scales. For each video clip, we randomly sample four target frames for face-based supervision. Each target latent is decoded together with its preceding two latent frames, yielding a 17-frame video segment. The last decoded frame is used as the target frame. We use RetinaFace for face detection and alignment with a detection threshold of 0.5 to filter low-confidence detections. The aligned face crops are resized to 256×256 and fed into ArcFace to extract ID embeddings. We discard reference-target pairs whose ArcFace cosine similarity is below 0.5. We set $\lambda_{\text{id}} = 0.5$, $\lambda_{\text{content}} = 0.3$, and $\lambda_{\text{traj}} = 0.3$. For Reference Trajectory Consistency, we sample $K_{\text{traj}} = 3$ observations per trajectory and maintain trajectories for at most three valid IDs, i.e., $|\mathcal{I}| \leq 3$. Observations whose face area differs from that of the trajectory anchor by more than 20% are selected as trajectory candidates. During inference, we transfer the trained LoRA weights to the corresponding distilled LTX-2.3 model and set the LoRA strength to 1.0. Final videos are generated at 768p resolution using 8-step inference.

4.2 Main Results

Subject-driven video generation. We compare CineRef with a broad range of representative subject-driven video generation methods, including closed-source models such as Vidu2.0 (Bao et al., 2024), Pika2.1 (Pika, 2024), Kling1.6 (Team et al., 2025), and Saber-14B (Zhou et al., 2025), as well as open-source methods including VACE-14B (Jiang et al., 2025), SkyReels-A2-P14B (Fei et al., 2025), Phantom-14B (Liu et al., 2025), MAGREF-480P (Deng et al., 2026), VINO (Chen et al., 2026b), Kaleido (Zhang et al., 2025), BindWeave (Li et al., 2026), and RefAlign-14B (Wang et al., 2026a). We further include LTX-Ingredients (HaCohen et al., 2026), which shares the same LTX-2.3 backbone as CineRef, to isolate the effect of our reference modeling. As shown in Table 1, CineRef achieves the highest **TotalScore** of 61.40% and the highest **FaceSim-Cur** of 74.13%, substantially outperforming the second-best FaceSim-Cur of 55.23%. This demonstrates strong reference ID injection and preservation. Meanwhile, CineRef achieves the highest **Aesthetics** score of 48.34% while maintaining competitive naturalness, suggesting that the improved reference fidelity is not obtained through a simple copy-and-paste shortcut. The comparison with LTX-Ingredients is particularly informative: under the same LTX-2.3 backbone, CineRef improves FaceSim-Cur from 19.76% to 74.13% and TotalScore from 43.33% to 61.40%, together with a substantial improvement in aesthetics. These results highlight the effectiveness of our explicit reference supervision and long-range reference preservation beyond the capability of the underlying backbone.

Audio-visual video generation. Since most subject-driven baselines do not natively support joint audio-video generation, we further compare CineRef with DreamID-Omni and MiniMax-H3 on CineSet. Under the single-shot setting in Table 2, CineRef achieves the best **FaceSim-Cur** and

[1] https://github.com/Lightricks/LTX-2

Table 2: Audio-visual generation comparison on CineSet under single-shot and multi-shot settings.

Setting	Dataset	Method	SpkSim ↑	Sync-C ↑	LSE-D ↓	FaceSim-Cur ↑	FaceSim-Arc ↑	GmeScore ↑	AesScore ↑
Single-shot	CineSet-single	MiniMax-H3	**0.7011**	**5.3967**	**7.4870**	**0.7218**	**0.6865**	**0.5203**	**5.5941**
		DreamID-Omni	0.2356	0.8760	10.9440	0.5122	0.4851	0.5130	5.4820
		LTX-Ingredients	0.2194	3.6030	7.6450	0.2717	0.2535	**0.5366**	4.6650
		CineRef	**0.5345**	**4.4484**	**7.4619**	**0.8359**	**0.8112**	0.4918	**5.5099**
	CineSet-multi	MiniMax-H3	**0.5981**	**3.2430**	10.0655	**0.6933**	**0.6564**	**0.4998**	**5.7218**
		DreamID-Omni	0.2712	0.8870	11.5970	0.4246	0.4052	0.4807	5.5080
		LTX-Ingredients	0.2571	**3.3450**	**8.3110**	0.1662	0.1558	**0.5119**	4.6620
		CineRef	**0.4466**	3.0417	**8.0814**	**0.7301**	**0.7067**	0.4844	**5.5272**
Multi-shot	CineSet-single	MiniMax-H3	**0.6988**	**4.7109**	**7.2815**	**0.7362**	**0.6985**	0.5125	**5.6779**
		DreamID-Omni	0.3611	3.5070	9.1910	0.6629	0.6400	**0.5278**	**5.5430**
		LTX-Ingredients	0.1827	2.9890	8.0170	0.2449	0.2351	**0.5140**	4.6910
		CineRef	**0.5134**	**3.9013**	**7.2194**	**0.7978**	**0.7707**	0.4794	5.4378
	CineSet-multi	MiniMax-H3	**0.6006**	3.3291	9.3634	**0.7096**	**0.6728**	**0.5085**	**5.6924**
		DreamID-Omni	0.3265	2.1110	9.9710	0.5104	0.4880	**0.4948**	**5.5040**
		LTX-Ingredients	0.2272	**3.3540**	**7.9720**	0.1669	0.1516	0.4781	4.5750
		CineRef	**0.4106**	**3.8334**	**7.4437**	**0.7225**	**0.6943**	0.4818	5.4845

FaceSim-Arc on both subsets. On CineSet-single, it reaches 0.8359 and 0.8112, respectively, clearly outperforming the second-best results. CineRef also obtains the lowest **LSE-D** on both subsets, indicating accurate lip–audio synchronization. Compared with LTX-Ingredients, it substantially improves visual ID similarity and speaker similarity while maintaining competitive semantic consistency and visual quality.

Under the more challenging multi-shot setting, CineRef again achieves the best **FaceSim-Cur** and **FaceSim-Arc** on both subsets, reaching 0.7978/0.7707 on CineSet-single and 0.7225/0.6943 on CineSet-multi. It also obtains the best **Sync-C** of 3.8334 and the lowest **LSE-D** of 7.4437 on CineSet-multi. These results demonstrate robust cross-shot ID preservation and strong audio-visual synchronization. The comparison with LTX-Ingredients further shows that the gains mainly come from our reference modeling rather than the LTX-2.3 backbone itself.

Inference efficiency. In addition to generation quality, we evaluate the inference efficiency of CineRef against MiniMax-H3 under the same hardware configuration. As shown in Figure 4, CineRef is consistently faster in both single- and multi-subject settings. For single-subject generation, CineRef reduces the average inference time from 1619 s to 1128 s, corresponding to a reduction of about 30%. The advantage becomes more pronounced in the multi-subject setting, where the average inference time decreases from 1931 s to 1057 s, a reduction of about 45%. Similar trends are observed for the median inference time. These results demonstrate that CineRef achieves strong reference-guided audio-video generation while maintaining substantially lower inference cost, particularly in the more challenging multi-subject setting.

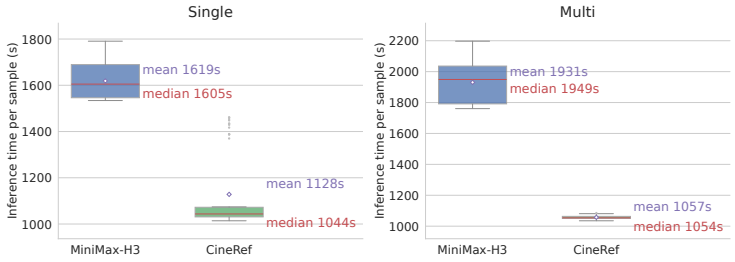


Figure 4: Comparison of Inference speed.

Dual Anchor Matters. Figure 5 illustrates the face-scale bias caused by using the ID anchor alone. As training progresses, the model increasingly enlarges facial regions to reduce the ID matching loss, rather than learning to preserve identity under natural facial scales. This optimization shortcut leads

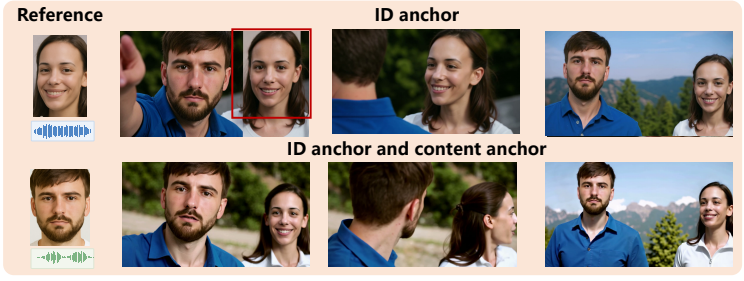


Figure 5: Dual anchor mitigates copy-and-paste artifacts caused by ID-only supervision while improving reference and content fidelity.

to oversized faces and degraded content fidelity. The content anchor alleviates this bias by introducing additional face-aware flow-matching supervision on ground-truth facial regions, encouraging the model to preserve natural facial content and spatial distribution. Together, the ID and content anchors balance ID fidelity with content fidelity.

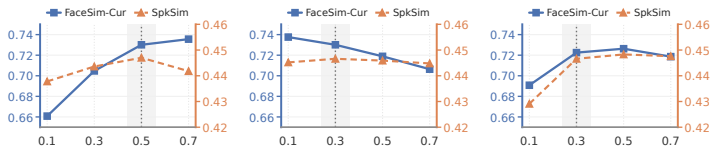
5 Ablation Study

Module ablation. Table 3 evaluates the contribution of each ablated component in CineRef. The ID anchor provides the largest gain in ID preservation, improving FaceSim-Cur from 0.1684 to 0.6447 in the single-shot setting and from 0.1708 to 0.5718 in the multi-shot setting. The content anchor further improves the scores to 0.6876 and 0.6143, respectively. Reference Trajectory Consistency is particularly effective for multi-shot generation, increasing FaceSim-Cur from 0.6143 to 0.6897, which demonstrates its benefit for cross-shot ID preservation. The ID and content anchors also gradually improve SpkSim, while Audio-Visual Reference Binding brings the largest gain, increasing SpkSim from 0.3417 to 0.4466 for single-shot generation and from 0.3126 to 0.4106 for multi-shot generation. Overall, these results confirm the complementary effects of explicit reference supervision, trajectory consistency, and audio-visual reference binding.

Table 3: Module ablation of CineRef on CineSet-multi.

Module				Single-shot		Multi-shot	
ID	Content	Traj	A/V	FaceSim-Cur ↑	SpkSim ↑	FaceSim-Cur ↑	SpkSim ↑
				0.1684	0.2547	0.1708	0.2286
✓				0.6447	0.3018	0.5718	0.2714
✓	✓			0.6876	0.3295	0.6143	0.2968
✓	✓	✓		0.7162	0.3417	0.6897	0.3126
✓	✓		✓	0.6918	0.4219	0.6197	0.3847
✓		✓	✓	0.7135	0.4298	0.7046	0.3934
✓	✓	✓	✓	**0.7301**	**0.4466**	**0.7225**	**0.4106**

Hyperparameter ablation. We study the sensitivity of CineRef to the three reference supervision weights in Fig. 6. As shown in Fig. 6a, increasing λ_{id} steadily improves FaceSim-Cur, indicating that stronger ID supervision enhances reference fidelity. However, when the weight becomes too large, the aesthetic score starts to drop. This suggests a trade-off between ID preservation and natural generation, and we therefore set $\lambda_{\text{id}} = 0.5$.



(a) λ_{id}(single shot) (b) λ_{content}(single shot) (c) λ_{traj}(multi shot)

Figure 6: Sensitivity analysis of the loss weights.

A similar trade-off is observed for λ_{content}. Increasing this weight improves visual quality, but gradually reduces FaceSim-Cur, since stronger content supervision places more emphasis on preserving natural facial content. We choose $\lambda_{\text{content}} = 0.3$ as a balanced setting.

For λ_{traj}, the effect is more pronounced in the multi-shot setting. Increasing it from 0.1 to 0.3 clearly improves cross-shot ID consistency, while further increasing the weight provides only limited gains and slightly hurts single-shot performance. We therefore set $\lambda_{\text{traj}} = 0.3$. Based on these results, we use $\lambda_{\text{id}} = 0.5$, $\lambda_{\text{content}} = 0.3$, and $\lambda_{\text{traj}} = 0.3$ in all experiments.

6 Conclusion

This paper reveals that existing reference-guided audio-video generation methods are limited by implicit reference modeling, where reference information is mainly treated as conditioning and can become difficult to preserve under complex cinematic dynamics. We propose **CineRef**, which explicitly models reference-to-generation correspondence during training and preserves it throughout the generated clip. Specifically, **Dual-Anchor Reference Alignment** balances ID fidelity and content fidelity, while **Reference Trajectory Consistency** strengthens long-range ID coherence under large viewpoint and scale variations. We further introduce **Dual-Branch Audio-Visual Reference Binding** to establish explicit correspondence among visual, audio, and textual conditions. Extensive experiments demonstrate the effectiveness of CineRef in improving ID consistency, audio-visual correspondence, and overall generation quality under challenging cinematic scenarios.

AI USE STATEMENT

We used generative AI tools to assist with language editing and wording refinement. These uses are separate from the multimodal and generative models employed as components of our proposed data construction pipeline, which are described explicitly in the paper. The scientific ideas, methodological design, experiments, data analysis, and conclusions were developed and verified by the authors. All AI-assisted content was manually reviewed and revised for correctness and consistency with the actual methodology and results. We take full responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided detailed implementation settings and hyperparameters in the **Implementation Details** section of the main paper to facilitate reproduction of our results. The complete data construction pipeline, including **Reference Image Construction**, **Scene Image Construction**, **Reference Audio Construction**, and **Structured Caption Generation**, is described in Appendix A.2. The source code will be made available upon acceptance to further support reproducibility.

REFERENCES

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024.

ByteDance Seed. Seedance 2.5. ByteDance Seed, 2026.

Guibin Chen, Dixuan Lin, Jiangping Yang, Youqiang Zhang, Zhengcong Fei, Debang Li, Sheng Chen, Chaofeng Ao, Nuo Pang, Yiming Wang, Yikun Dou, Zheng Chen, Mingyuan Fan, Tuanhui Li, Mingshan Chang, Hao Zhang, Xiaopeng Sun, Jingtao Xu, Yuqiang Xie, Jiahua Wang, Zhiheng Xu, Weiming Xiong, Yuzhe Jin, Baoxuan Gu, Binjie Mao, Yunjie Yu, Jujie He, Yuhao Feng, Shiwen Tu, Chaojie Wang, Rui Yan, Wei Shen, Jingchen Wu, Peng Zhao, Xuanyue Zhong, Zhuangzhuang Liu, Kaifei Wang, Fuxiang Zhang, Weikai Xu, Wenyan Liu, Binglu Zhang, Yu Shen, Tianhui Xiong, Bin Peng, Liang Zeng, Xuchen Song, Haoxiang Guo, Peiyu Wang, Max W. Y. Lam, Chien-Hung Liu, and Yahui Zhou. Skyreels-v4: Multi-modal video-audio generation, inpainting and editing model, 2026a. URL https://arxiv.org/abs/2602.21818.

Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. Vino: A unified visual generator with interleaved omnimodal context. *arXiv preprint arXiv:2601.02358*, 2026b.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning, 2025a. URL https://arxiv.org/abs/2509.08519.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025b.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yunsheng Jiang, Li Li, Xinghui Li, Liyang Chen, Siyu Zhou, et al. Phantom-data: Towards a general subject-consistent video generation dataset. In *International Conference on Learning Representations*, volume 2026, pp. 73018–73037, 2026c.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob 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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Nbl43eAVaE.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian HE, and Xiangwang Hou. DreamID-omni: Unified framework for controllable human-centric audio-video generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bZjSinGaUo.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jiahao Wang, Xu He, Zhenchao Tang, Meng Chu, Xin Tao, Pengfei Wan, and Jiaya Jia. Unityshots: Memory-driven multi-shot audio-video generation with boundary-aware gating. *arXiv preprint arXiv:2606.21661*, 2026.

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

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Zhaoyang Li, Dongjun Qian, Kai Su, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, Zehuan Yuan, et al. Bindweave: Subject-consistent video generation via cross-modal integration. In *International Conference on Learning Representations*, volume 2026, pp. 107960–107979, 2026.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=hRRWfFpKRp.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, and Boxin Shi. MAVIN: Multi-shot audio-visual generation with customized narrative control. *arXiv preprint arXiv:2606.29473*, 2026b.

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. *arXiv preprint arXiv:2502.11079*, 2025.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. MiniMax Research, 2026.

Pika. Pikascenes. https://pika.art/ingredients/, 2024.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang

Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

Lei Wang, Yuxin Song, Ge Wu, Haocheng Feng, Hang Zhou, Jingdong Wang, Yaxing Wang, and Jian Yang. Refalign: Representation alignment for reference-to-video generation. *arXiv preprint arXiv:2603.25743*, 2026a.

Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, Yuan Zhang, Mingyuan Gao, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=rJilRU8D3c.

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*, 2025.

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. *arXiv preprint arXiv:2512.06905*, 2025.

A Appendix

A.1 Baseline and Evaluation Metrics

Baseline settings. For **DreamID-Omni**, we use the default setting of 50 denoising steps with classifier-free guidance (CFG). Since DreamID-Omni does not natively support multi-shot inference, we generate each shot independently using its corresponding shot-level description. Global video attributes are kept consistent across shots, while shot-specific descriptions control the content

Table 4: Evaluation metrics used on CineSet. ↑ and ↓ indicate that higher and lower values are better, respectively. Sync-C and LSE-D are computed from the lip–audio distance curve produced by SyncNet.

Category	Metric	Better	Description
Audio ID	SpkSim	↑	Cosine similarity between speaker embeddings of generated and reference speech.
Visual ID	FaceSim-Cur	↑	Face embedding similarity using CurricularFace (R101).
Visual ID	FaceSim-Arc	↑	Face embedding similarity using ArcFace (InsightFace).
Lip-sync	Sync-C	↑	Lip–audio synchronization confidence.
Lip-sync	LSE-D	↓	Lip–audio distance at the best temporal alignment.
Semantic	GME Score	↑	Similarity between the video and the corresponding textual description using GME-Qwen2-VL-7B.
Visual	AES	↑	Mean aesthetic score over sampled video frames.

and dynamics of individual shots. Since **MiniMax-H3** natively supports multi-shot generation, we use the default inference hyperparameters with 50 denoising steps and CFG.

CineSet evaluation. For CineSet, we evaluate reference-guided audio-video generation from five complementary aspects: speaker identity, visual identity, lip synchronization, semantic consistency, and visual quality. **SpkSim** measures the cosine similarity between speaker embeddings of the generated and reference speech, focusing on whether the generated voice preserves the reference timbre rather than the spoken content. For visual identity, we report both **FaceSim-Cur** and **FaceSim-Arc**, computed using CurricularFace and ArcFace, respectively. Using two independent face recognition models provides complementary measurements of ID preservation and reduces dependence on a single embedding space. For audio-visual synchronization, **Sync-C** measures the confidence of the best lip–audio alignment, while **LSE-D** measures the corresponding alignment distance. The two metrics are complementary: higher Sync-C indicates a more distinctive synchronization match, whereas lower LSE-D indicates a smaller lip–audio discrepancy. Beyond reference fidelity, **GmeScore** evaluates semantic consistency between the generated video and the corresponding textual description, capturing whether the generated content follows the intended scene and actions. Finally, **AesScore** measures the perceptual aesthetics of generated frames. Higher values indicate better performance for all metrics except LSE-D, for which lower is better.

OpenS2V evaluation. Following the OpenS2V benchmark, we report seven component metrics together with the overall **TotalScore**. **Aesthetics** measures the perceptual and visual quality of the generated video, while **MotionSmoothness** evaluates temporal smoothness and motion continuity. **MotionAmplitude** measures the magnitude of motion in the generated content, reflecting whether the model can produce sufficiently dynamic videos rather than nearly static outputs. **FaceSim** evaluates facial ID similarity between the generated subject and the reference image. **GmeScore** measures the semantic relevance between the generated video and the input condition, while **NexusScore** evaluates subject consistency across the generated video. **NaturalScore** measures the naturalness of the generated subject and penalizes artifacts such as unnatural appearance or copy-and-paste behavior. **TotalScore** summarizes the overall generation performance across these complementary aspects. Higher values indicate better performance for all OpenS2V metrics.

A.2 Dataset Construction

A.2.1 Reference Image Construction

Reference image construction plays an important role in reference-guided video generation. A straightforward approach is to segment the foreground subject from the original video frames and directly use it as the reference image for the corresponding clip. However, references extracted from the same clip can be overly similar to the target frames, encouraging the model to exploit a copy-and-paste shortcut rather than learning robust reference-to-generation correspondence. To alleviate this issue, we construct reference images through a hierarchical retrieval process that considers both character identity and appearance consistency across long-form videos.

Given a long-form video, we first perform face-based identity tracking and aggregate ArcFace embeddings within each trace to obtain a trace-level identity representation. We then cluster temporally separated traces belonging to the same character based on these representations. For each trace, we sample several frames and extract the foreground subject using a segmentation model. Qwen-VL

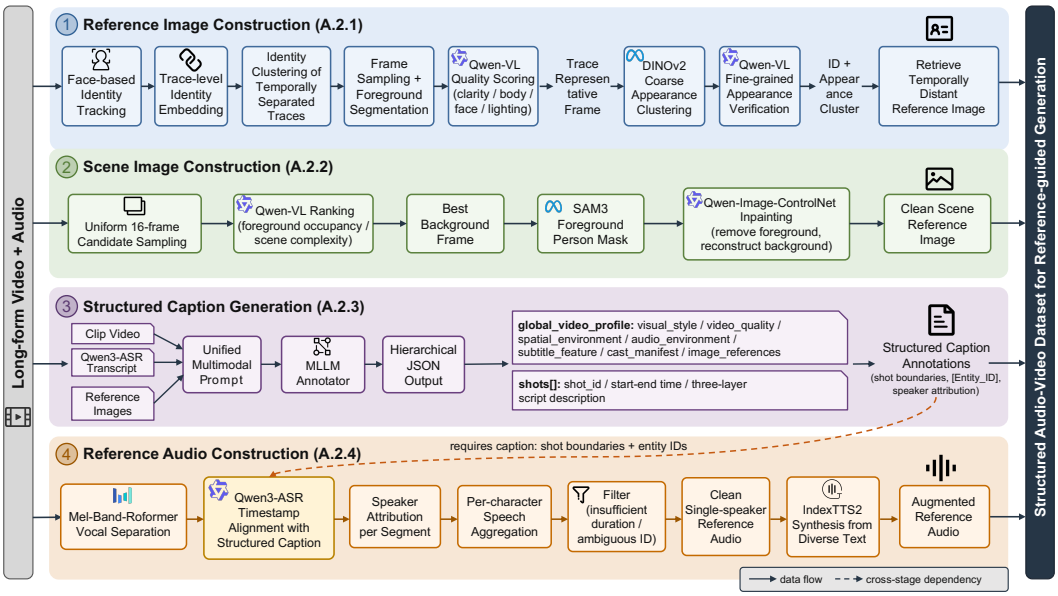


Figure 7: Overview of Dataset Construction.

then evaluates the sampled frames based on clarity, body completeness, face completeness, and lighting quality, and selects the highest-scoring frame as the trace representative.

Identity consistency alone is insufficient for reference retrieval, since the same character may appear with different clothing, makeup, hairstyles, or other visual styling throughout the video. A reference with a mismatched appearance configuration can weaken appearance following during generation. We therefore further partition each identity cluster according to appearance consistency. Specifically, representative frames are first grouped using DINOv2 features to obtain coarse appearance clusters, reducing the search space for subsequent comparison. Qwen-VL then performs fine-grained appearance verification within each group, merging traces with consistent character styling while separating those with different appearance configurations.

Through this hierarchical identity-and-appearance association, we retrieve high-quality reference images from temporally distant traces that match both the character ID and target appearance. This reduces direct appearance overlap between the reference and target clip while preserving consistent character styling, discouraging copy-and-paste shortcuts and improving reference following.

A.2.2 Scene Image Construction

In addition to reference images, scene images provide important spatial context for audio-video generation. However, directly extracting scene images from source videos is non-trivial, as most frames contain foreground subjects and cannot serve as clean scene references. We therefore adopt a three-stage strategy consisting of **candidate sampling**, **scene ranking**, and **foreground removal**. To reduce the computational cost of processing all frames, we first uniformly sample 16 frames from each video clip as candidates. We then evaluate each candidate based on foreground occupancy and scene complexity using Qwen3.5-397B, and select the frame with the highest score. Finally, SAM3 is used to segment the foreground person and obtain its mask, after which Qwen-Image-ControlNet-Inpainting removes the foreground region and reconstructs the missing background. The resulting image is used as the scene reference.

A.2.3 Reference Audio Construction

Reference audio should provide clean and reliable timbre information for each character. We construct it in two stages: **character speech extraction** and **speech augmentation**. First, we separate vocals from the original audio using Mel-Band-Roformer. We then align the Qwen3-ASR timestamps with the structured caption annotations described below. This alignment identifies the

speaking character for each speech segment. Speech segments belonging to the same character are subsequently aggregated. Characters with insufficient speech duration or ambiguous identity correspondence are filtered out. This yields clean single-speaker reference audio.

The extracted speech may still have limited linguistic diversity. We therefore use it as a timbre prompt for IndexTTS2 to synthesize additional speech from diverse text. This improves linguistic diversity and phoneme coverage while preserving the character's timbre. The resulting audio provides richer and more robust references for generation.

A.2.4 Structured Caption Generation

Conventional video captions often compress an entire clip into a single coarse description. This makes shot-level changes difficult to model and obscures the distinction between persistent attributes and transient dynamics. We therefore introduce **Structured Caption Generation**, where an MLLM is used to convert each video clip into a hierarchical structured annotation.

Each sample contains two levels. The first is a **global profile** shared across all shots. It records persistent information such as visual style, scene environment, audio environment, subtitle attributes, and the static appearance of each character. It also stores the correspondence between reference images and their associated entities or scenes. The second level is a **shot list** obtained from shot-boundary segmentation. Each shot contains its temporal boundary together with a description of camera configuration, lighting, character actions, dialogue, and other local dynamics. This decomposition explicitly separates cross-shot invariants from shot-level variables.

To provide more explicit supervision for cross-shot consistency and audio-visual correspondence, we further structure the annotation from four complementary aspects.

(i) Static–dynamic decoupling via entity anchoring. Each core entity is assigned a persistent composite tag, such as `[middle-aged-man_A]`, which is reused across all shots. The tag contains only clip-persistent attributes, while actions, emotions, and temporary props are described separately at the shot level. Fine-grained static appearance is recorded once in the global cast manifest. This decouples character identity from transient dynamics and strengthens cross-shot subject consistency.

(ii) Global-to-local context injection. The global profile provides shared visual, spatial, and audio context for all shot-level descriptions. These attributes include visual style, lighting characteristics, scene environment, ambient sound, and acoustic properties. Conditioning each shot on this shared context encourages consistent rendering across shot transitions.

(iii) Explicit cinematic and action supervision. Each shot is annotated with its camera configuration, lighting condition, and character actions. We further describe the temporal relation between character actions and speech when applicable. This provides fine-grained supervision for cinematic dynamics and audio-conditioned motion.

(iv) Hard audio-visual temporal alignment. Speech is explicitly associated with its corresponding entity and delimited by fixed boundary tokens, `[entity_ID:voice-state]<S>... <E>`. The voice state distinguishes on-screen speech, voice-over, and inner monologue according to their visual correspondence. Synchronized non-verbal sounds and foley events are also explicitly annotated. This provides direct supervision for speaker attribution, lip synchronization, and audio-visual correspondence.

Overall, the structured caption represents each video as a combination of persistent global attributes, shot-level dynamics, and temporally aligned audio-visual events. This provides more precise and decomposed supervision for cinematic audio-video generation.

A.3 CineSet Benchmark Construction

We construct **CineSet**, a benchmark for reference-guided multi-subject video generation conditioned on character reference images, reference speech, and structured storyboards. Each sample contains reference images for one or more characters, the corresponding reference audios, and a storyboard specifying both global video attributes and shot-level dynamics. CineSet contains 100 samples, including 60 single-subject samples and 40 multi-subject samples. Each multi-subject sample contains

Table 5: Distribution of camera motion patterns in the multi-shot and single-shot CineSet scripts. **The multi-shot variant is counted by shot**: the single-subject subset contains 180 shots from 60 samples, while the multi-subject subset contains 120 shots from 40 samples. The single-shot variant is counted by sample (60 single-subject and 40 multi-subject samples). Since a single shot may contain multiple motion patterns, the row counts can sum to more than the total number of shots; the single-subject subset contains an average of 1.36 motion types per shot. Neither variant contains samples that remain fully static throughout the entire video.

	Multi-shot (by shot)		**Single-shot (by sample)**	
Motion Pattern	Single (n=180)	Multi (n=120)	Single (n=60)	Multi (n=40)
Push-in / Pull-out / Zoom	93	62	24	20
Lateral / Tracking Motion	51	52	11	12
Handheld Motion	46	24	25	12
Pan	41	21	4	6
Orbit / Rotation	7	6	21	14
Crane / Elevation	6	3	7	6
Static Camera	1	1	–	–

two characters appearing across different shots, providing a challenging setting for evaluating cross-shot ID preservation. In total, CineSet includes 140 reference images and 140 reference audios.

To disentangle the effects of camera motion and shot transitions, we construct **multi-shot** and **single-shot** variants for the same set of samples. The multi-shot variant consists of three 5-second shots separated by hard cuts, resulting in a total duration of 15 seconds. The single-shot variant removes all hard cuts and merges the original actions and dialogue into one continuous 10-second shot. Both variants are designed with diverse camera motions, and no sample remains fully static throughout the video. The statistics are presented in Table 5. In the multi-shot variant, 298 of 300 shots explicitly contain camera motion. Push-in, pull-out, and zoom are the most frequent patterns, followed by lateral/tracking and handheld motions. Since a single shot may involve multiple motion patterns, these categories are not mutually exclusive. For the single-shot variant, we construct a pool of 30 camera-motion trajectories, including 20 general trajectories and 10 single-subject-specific trajectories, and distribute them across the benchmark to ensure diverse and balanced motion coverage. All videos are generated at 1344×768 resolution and 24 FPS. This controlled design enables systematic evaluation of ID consistency, speaker similarity, and audio-visual synchronization under diverse camera motions and shot-transition patterns.