CastDirector: Directing Joint Audio-Video Generation with Grounded Screenplays

Anonymous authors
Paper under double-blind review

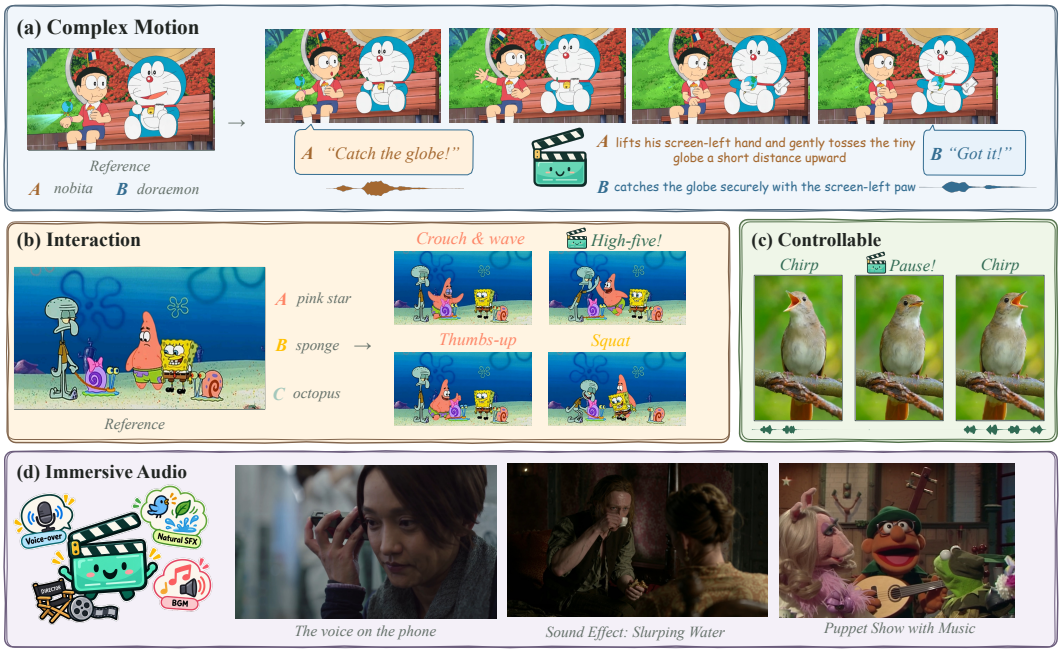

Figure 1: Screenplay-driven generation with CastDirector: (a) cartoon toss-and-catch with assigned dialogue; (b) character-specific gestures; (c) bird vocalization with a directed pause; (d) phone speech, drinking sounds, and puppet-show music. Videos with sound: anonymous demo page.

Abstract

Cinematic audio-video generation must assign dialogue and actions to the intended characters while producing immersive, scene-consistent audio. Existing systems can produce coherent clips while assigning speech or actions to the wrong character or omitting requested events. We introduce CastDirector, a screenplay-driven system for coordinating character performances and cinematic soundtracks. A grounded screenplay specifies the cast, speaker identities, dialogue, actions, and camera and audio cues as ordered events. To construct training screenplays from audiovisual data, our annotation pipeline extracts transcripts and visual actions independently, then verifies and combines them with the original clip to bind dialogue and actions to characters. The proposed multimodal Director jointly encodes this screenplay and the first frame, conditioning both generation streams to connect character actions, voices, and sound events to the visible scene. We first align a pretrained LTX-2.3 generator with the Director's features, then adapt the Director with LoRA through the audio-video generation loss. Under matched screenplay content, experiments show improved speaker binding, higher human ratings for action, dialogue, and audio, and gains in audio quality and speech accuracy. Qualitative examples showcase coordinated performances and immersive audio combining character voices, environmental ambience, sound effects, and scene-appropriate music.

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.

1 INTRODUCTION

Joint audio-video systems such as OVI (Low et al., 2025), JavisDiT (Liu et al., 2026), LTX-2.3 (HaCohen et al., 2026), and daVinci (SII-GAIR et al., 2026) can synthesize video, speech, sound effects, and music from a conditioning frame and a text prompt. For cinematic scenes, these elements must follow a shared dramatic plan: the intended character delivers each line and performs each action, interactions unfold in order, and voices and sound events fit the visible scene. Current image-to-video-and-audio (I2VA) systems can produce plausible clips while assigning speech or actions to the wrong character, omitting an interaction, or losing the requested event order; Fig. 4 illustrates failures in object ownership and character-specific exits. These failures motivate a directing interface that binds screenplay instructions to the characters and setting in the first frame.

We introduce **CastDirector**, which uses a grounded screenplay to specify character performances and a multimodal Director to guide joint audio-video generation. The screenplay lets users specify the cast and voices and organize character-specific dialogue, actions, camera instructions, and audio cues into ordered events (Table 1). To guide their execution, the Director, a vision-language model, jointly encodes the screenplay and first frame to condition both generation streams (Fig. 2). To supply training screenplays for this system, our evidence-separated annotation pipeline independently transcribes audio and describes actions from silent video, then verifies both against the original clip to assign dialogue and actions to characters (Fig. 3).

The Director addresses two requirements for executing this screenplay. **First, the screenplay must be interpreted together with the first frame.** A text-only prompt encoder cannot directly associate character descriptions with the visible cast. Voice and sound-effect instructions also need the context of the characters, objects, and environment depicted in the shot. We therefore use Qwen3-VL (Bai et al., 2025) to jointly encode the first frame and screenplay. Modality-specific adapters inject its features into both streams of LTX-2.3 (HaCohen et al., 2026), making visual context available to audio generation as well as to character actions. **Second, the Director's representation must serve screenplay-conditioned generation.** We first train the generator to consume a frozen Director's features, then adapt the Director with LoRA (Hu et al., 2022) through the audio-video generation loss. This trains the conditioning representation on structured screenplays and the resulting synthesis task.

We assess character and event execution through automatic measurements, blinded human ratings, and qualitative comparisons, alongside evaluation on Verse-Bench (Wang et al., 2025b). On the proposed Director-Bench, a 40-case benchmark for character and event execution, CastDirector improves speaker assignment and receives higher mean ratings for action, dialogue, and audio than the native backbone. Examples show coordinated gestures, cartoon and animal characters, and immersive soundtracks combining scene-consistent voices, ambience, effects, and music (Fig. 1 and Section 4.6).

Our main contributions are:

- A multimodal Director for character-specific screenplay execution that grounds both generation streams in the visible cast and scene through modality-specific conditioning. Generator alignment followed by Director adaptation optimizes this conditioning through the joint audio-video generation loss.
- A grounded screenplay representation of cast, dialogue, actions, and audiovisual cues, supported by an evidence-separated pipeline for constructing character-specific training annotations from audiovisual data.
- Strong performance on character binding and audiovisual generation: CastDirector improves speaker assignment and turn-level synchronization, achieves the highest observed mean human ratings for action, dialogue, and audio on Director-Bench, and obtains leading audio-quality and speech-accuracy results in our Verse-Bench comparison.

2 RELATED WORK

Character Animation. EMO (Tian et al., 2024), OmniHuman (Lin et al., 2025), MuseTalk (Zhang et al., 2024), and InfiniteTalk (Yang et al., 2025) animate characters from supplied speech. MultiTalk (Kong et al., 2025) and AnyTalker (Zhong et al., 2025) extend audio-driven generation to multiple people. Implicit character decoupling (Xue et al., 2025), SocialDirector (Ouyang et al.,

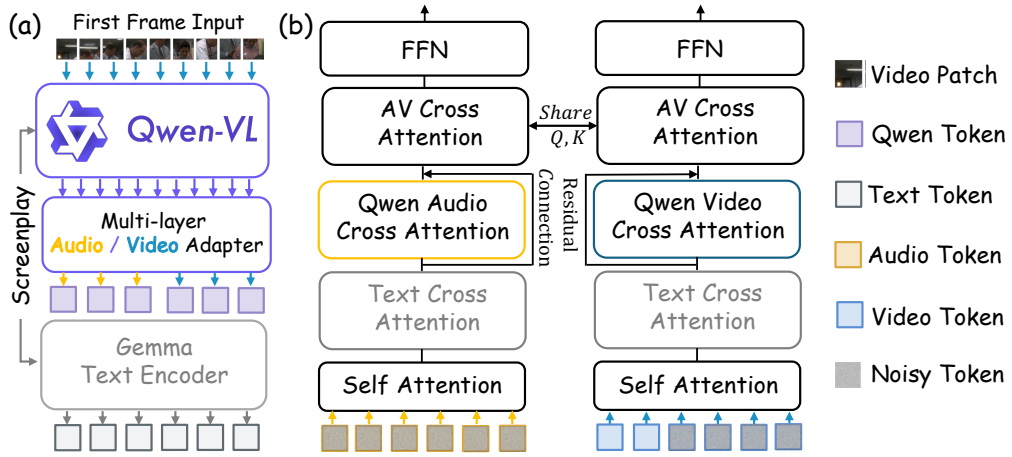

Figure 2: Overview of CastDirector. **(a) Screenplay encoding.** The Qwen-VL Director jointly encodes the screenplay and first frame, and a multi-layer adapter produces modality-specific audio and video contexts. Gemma separately encodes the same screenplay. **(b) Dual-stream injection.** New residual Qwen cross-attention branches inject these contexts into both streams after text cross-attention and before audio-video cross-attention in each block. Textured tokens denote noisy latents, while first-frame video tokens remain clean.

2026), and MindFlow (Chen et al., 2026a) address shared-frame characters, social interaction, and dyadic animation, respectively. While audio-driven animation takes speech as input, our I2VA setting jointly generates the soundtrack and video from a first frame and screenplay. The screenplay must therefore specify who speaks and acts, as well as how the scene should sound.

Joint Audio-Video Generation. UniVerse-1 (Wang et al., 2025b), OVI (Low et al., 2025), NAVA (Ji et al., 2026), JOVA (Huang et al., 2025a), and Harmony (Hu et al., 2026) combine audio and video generation through pretrained experts, cross-modal fusion, or joint attention. JavisDiT (Liu et al., 2026) uses a hierarchical spatio-temporal prior; daVinci (SII-GAIR et al., 2026) adopts a single-stream design; LTX-2 (HaCohen et al., 2026) and MOVA (OpenMOSS Team et al., 2026) use dual-stream backbones with bidirectional cross-attention. These advances provide the generation backbones on which we study screenplay execution and character binding. We build primarily on LTX-2.3 and also evaluate an OVI integration.

Language and Multimodal Conditioning. T5 (Raffel et al., 2020) and Gemma (Gemma Team et al., 2025) provide pretrained language features, while decoder-only LLM conditioning (Wang et al., 2025a), Playground v3 (Liu et al., 2024), OmniHuman-1.5 (Jiang et al., 2025), HOMIE (Cai et al., 2026), and HunyuanVideo 1.5 (Wu et al., 2025) explore richer language and multimodal representations. VideoDirectorGPT (Lin et al., 2023) and DreamRunner (Wang et al., 2026) organize prompts into scene plans. Seed-TTS (Anastassiou et al., 2024), MoonCast (Ju et al., 2025), and CoVoMix2 (Zhang et al., 2025) study voice conditioning formats for speech or dialogue generation. Building on these ideas, we design a multimodal director that jointly encodes structured prompts and their associated character images, enabling fine-grained control over character interactions in generated audio-visual scenes.

3 METHOD

Given a first frame I and a grounded screenplay S, CastDirector generates a video V and soundtrack A that follow the specified character dialogue, actions, and audiovisual cues in the requested order. We first define the screenplay, then describe the Director's joint visual-text encoding and its adaptation to generation (Fig. 2). Our backbone is LTX-2.3 (HaCohen et al., 2026), a dual-stream diffusion transformer (Peebles & Xie, 2023) with native Gemma (Gemma Team et al., 2025) text conditioning and bidirectional audio-video cross-attention. We preserve these operations and the flow-matching objective (Lipman et al., 2023); first-frame video tokens remain clean and are excluded from the loss (Section A.4).

Table 1: The hierarchical grounded screenplay format, illustrated with a two-character scene. [] denote parent tags, and <> denote nested child tags. Section A.1 gives the complete record.

Section / nested field	Concrete value
[SCENE]	Dim concrete corridor; dark green tones and wall lamps.
[VISIBLE_MAIN_CHARACTER]	Two visible characters in this case; ...
<count>	2
<first>	Woman with long dark hair and a dark jacket; ...
<appearance>	Long dark hair; dark jacket.
<posture>; <frame_position>	Standing, facing the man; frame-left.
<actions>	Moves leftward and exits the frame.
<second>	Second: man in a dark coat and scarf; frame-right
[EVENT_SEQUENCE]	Two chronological events in this case; ...
<first>	The man apologizes while the two characters face each other; ...
<dialogue>	One spoken turn.
<voice>	Adult male, soft and apologetic.
<character_id>	speaker_1: man in dark coat and scarf.
<words: English>	"I'm sorry."
<event>	The woman stands facing the man in the corridor.
<lighting>; <camera_motion>	Dim green lighting; slow pull-out.
<second>, <third>, ...	Second: the woman exits frame-left; the man turns left; ...
[AUDIO_CAPTION]	Low ambient hum, soft male whisper, ...
[SUBJECT]	Indoor, people, ...
[ATMOSPHERE]	Tense, cinematic, ...

3.1 Grounded Screenplay

The screenplay S is tagged text with six sections: scene, visible cast, event sequence, audio caption, subject, and atmosphere. Cast entries describe appearance, posture, and frame position. Each ordered event names its actor and may contain dialogue, lighting, and camera instructions. Dialogue specifies a voice description, a persistent speaker_n ID paired with an identity phrase, and language-tagged verbatim words. Cast entries use ordinal tags; speaker IDs identify recurring voices, and identity phrases connect dialogue and actions to the described cast. Thus speaker numbering need not match cast order. The native text encoder and Director receive the same complete screenplay (Fig. 2(a)), preserving a shared, editable specification for both modalities. Table 1 illustrates the bindings; Section A.1 gives the full serialization.

3.2 Evidence-Separated Screenplay Annotation

Preparing a screenplay requires accurate words, visible actions, and speaker assignments. When used alone to annotate a clip end to end, Gemini (Kavukcuoglu et al., 2026) can misrecognize spoken words, add unsupported details, or describe actions that do not occur. We therefore supply independently extracted audio and visual priors to guide its multimodal reasoning (Fig. 3). For a training clip $D = (V, A)$, the audio prior R_{aud} and visual prior R_{vis} are fused into screenplay S:

$$R_{\text{aud}} = \text{ASR}(A), \qquad R_{\text{vis}} = \text{Keye}(V), \qquad S = \text{Gemini}(V, A;\, R_{\text{aud}}, R_{\text{vis}}). \tag{1}$$

Audio prior. Qwen3-ASR (Shi et al., 2026) extracts the transcript and language label directly from the soundtrack. This audio prior anchors the spoken content in acoustic evidence, reducing reliance on Gemini's standalone speech recognition during end-to-end annotation.

Visual prior. Keye-VL-2.0 (Kwai Keye Team et al., 2026) independently reads silent video to describe the visible cast and chronological actions. Its prompt excludes speech and mouth-motion descriptions and requires concrete actors instead of ambiguous pronouns, providing appearance and motion evidence without assigning dialogue.

Multimodal verification and binding. Gemini (Kavukcuoglu et al., 2026) receives the original clip together with the audio prior and visual prior. The audio prior specifies the spoken words and language, while the original soundtrack provides voice attributes and temporal evidence for speaker



Figure 3: Evidence-separated annotation. Qwen3-ASR (Shi et al., 2026) provides the audio prior, comprising transcripts and language labels; Keye-VL-2.0 (Kwai Keye Team et al., 2026) provides the visual prior from silent video. Gemini (Kavukcuoglu et al., 2026) verifies and fuses both with the original clip, adding voice attributes and speaker assignments.

assignment. It treats the visual descriptions as fallible, checking each proposed action against the clip. For dialogue, it first describes the heard voice, then associates the turn with a compatible visible character using lip evidence, or marks an off-screen or device source. This separates voice description from speaker assignment and avoids choosing a speaker merely because that character is visually prominent. Each turn records voice attributes, a speaker ID with an identity phrase, and language-tagged words in that order. The resulting screenplay combines character-specific events with scene-level audio and atmosphere. Prompt constraints, decoding settings, and the validation and regeneration procedure are detailed in Section A.3.

3.3 Joint First-Frame and Screenplay Encoding

A screenplay identifies roles and events, but executing them requires associating its descriptions with the visible cast and setting. This context matters for both modalities: video must assign actions to the intended characters, while audio must combine the specified voices and sound events with the depicted scene. We therefore use Qwen3-VL-8B (Bai et al., 2025) as a multimodal Director Q_ϕ that jointly reads I and S. Its input contains the same first-frame view seen by the video VAE, a task instruction, and the full screenplay. The Director returns feature representations from each of its L layers and a token-validity mask:

$$\{H^\ell\}_{\ell=1}^{L},\ M = Q_\phi(I, S), \qquad H^\ell \in \mathbb{R}^{T\times d_q}, \quad M \in \{0,1\}^T, \tag{2}$$

where T is the sequence length, d_q the hidden width, and M the token-validity mask. These features carry joint visual-text context; padded positions are zeroed and masked out of downstream attention. The corridor example in Table 1 and Fig. 5 illustrates the intended grounding. The first frame places the woman in the left foreground and the man in a dark coat and scarf on the right, providing visual referents for the screenplay's cast entries. The identity phrase paired with `speaker_1` assigns "I'm sorry" to the man, although he is the second cast entry; the subsequent event assigns the exit to the woman. Joint encoding lets the Director interpret these dialogue and action instructions against the same visible cast, while the corridor, voice description, and audio caption provide shared context for the soundtrack. The first frame establishes the scene and characters; the screenplay specifies their subsequent performance (Section A.1). We then map the Director's features to modality-specific contexts and inject them into both generation streams.

Modality-specific contexts. Audio and video use different attention widths and may benefit from different levels of the Director's representation. For each stream $m \in \{v, a\}$, a learned layer mixture, LayerNorm, and projection produce context C_m:

$$\pi_m^\ell = \frac{\exp(\alpha_m^\ell)}{\sum_{j=1}^{L}\exp(\alpha_m^j)}, \qquad C_m = P_m\Big(\mathrm{LN}_m\Big(\sum_{\ell=1}^{L}\pi_m^\ell H^\ell\Big)\Big) \in \mathbb{R}^{T\times d_m}. \tag{3}$$

Here d_m is the stream's attention width, and the mixture logits α_m^ℓ start at zero. Both streams initially read the same layer average and learn their own mixture and projection.

Dual-stream injection. Each backbone block b receives a new Director cross-attention branch after native text attention and before audio-video exchange. For stream tokens X_m^b at this in-

sertion point, the update is

$$X_m^b \leftarrow X_m^b + \mathrm{CA}_{m,b}^{\mathrm{dir}}\big(\mathrm{RMSNorm}(X_m^b), C_m;\, M\big). \tag{4}$$

Both streams therefore access the same visually grounded screenplay before exchanging information. In particular, C_a gives audio generation access to character and scene context alongside the requested voice and sound descriptions. Following ControlNet (Zhang et al., 2023), we initialize the new attention output projections to zero, preserving the pretrained generator at initialization. The Director is encoded once per clip and its contexts are reused throughout denoising; inference and resource costs appear in Sections A.5 and A.6.

3.4 Adapting the Director to Screenplay-Conditioned Generation

Joint encoding supplies multimodal features, but the pretrained Director has not been optimized to provide conditions for this generator's screenplay execution. We adapt the Director through the audio-video flow-matching loss, allowing synthesis errors to update the representation of the frame and screenplay. Training proceeds in two phases on the same corpus so that the generator first learns to read a fixed representation before the Director itself is updated.

Phase 1: Director Alignment. We freeze the Director, train rank-64 LoRA on the query, key, value, and output projections of native LTX attention, and fully train the new layer mixtures, projections, and Director cross-attention branches. This establishes the interface between the pretrained multimodal encoder and the generator.

Phase 2: Director Adaptation. We adapt the Director with rank-32 LoRA and continue training the LTX LoRA. The layer mixtures, projections, and Director cross-attention branches introduced in Phase 1 remain fully trainable throughout Phase 2 (Fig. 7). Gradients from the audio-video generation loss update these modules together with the Director and LTX LoRA, allowing the conditioning representation and its interface to the generator to adapt jointly. This optimizes the Director for screenplay-conditioned synthesis without a separate language-modeling objective.

Both phases use the same audio-video flow-matching objective, with clean first-frame conditioning and joint Director-context dropout of 0.1 for classifier-free guidance (Ho & Salimans, 2022). The loss and optimization details are provided in Section A.4.

4 Experiments

4.1 Experimental Setup

Model and training. Our primary system combines LTX-2.3-dev (22B) with Qwen3-VL-8B. We train at 1088×640 and 24 fps on clips of 113 to 169 frames, using a global batch size of 128 on 32 H200 GPUs. The filtered and annotated corpus contains 1,459,594 samples totaling 2,177 hours, including film and television data. Director Alignment (Phase 1) and Director Adaptation (Phase 2) use the same complete training corpus. Data processing, training details, and inference are given in Sections A.2, A.4 and A.5.

Baselines. We compare with MOVA (OpenMOSS Team et al., 2026), NAVA (Ji et al., 2026), daVinci (SII-GAIR et al., 2026), OVI (Low et al., 2025), and LTX-2.3 (HaCohen et al., 2026), using their official inference code and default sampling settings. Outputs are aligned to the evaluation frame rate and duration. Section 4.5 evaluates our design on both LTX-2.3 and OVI.

4.2 Evaluation Setup

Inputs and general quality. We reannotate 395 Verse-Bench Set-2/3 reference clips as screenplays and evaluate generation using the UniVerse-1 (Wang et al., 2025b) toolkit. CastDirector receives these screenplays, while Gemini 3.5 rewrites the same content into each baseline's official prompt format. Manual checks verify matching dialogue, character actions, event order, and audio descriptions across methods. We report DINOv3 image identity (ID) (Siméoni et al., 2025), audio Fréchet distance (FD) (Kilgour et al., 2019), PaSST KL (Koutini et al., 2022), CLAP score (CS) (Elizalde et al., 2023), AudioBox enjoyment/usefulness/quality (CE/CU/PQ) (Tjandra et al., 2025), word error rate (WER)

Table 2: Director-Bench complex-scene generation. Automatic metrics evaluate fidelity, speech, turn-localized synchronization, and speaker assignment. The three rightmost columns report human ratings from 1 to 5, averaged equally over cases, for action (Act.), dialogue (Dlg.), and audio (Aud.). Higher is better. **Bold** and underline denote best and second best results.

Method	FID↓	FVD↓	ID↑	WER↓	Turn Sync-C↑	Speaker Match↑	Act.↑	Dlg.↑	Aud.↑
OVI	134.85	40.19	0.3486	0.4823	1.9479	0.6316	2.30	2.40	2.30
MOVA	123.31	33.82	**0.5009**	0.1136	2.1513	0.6621	3.18	3.86	3.21
NAVA	118.36	39.56	0.4139	0.0912	2.6182	0.7043	3.11	3.74	3.27
daVinci	127.69	37.90	0.4498	0.0801	2.1939	0.7062	3.26	3.79	3.32
LTX-2.3	114.34	34.79	0.4951	0.1192	3.4091	0.7500	3.67	3.74	3.82
CastDirector	**108.19**	**28.28**	0.4992	**0.0769**	**3.8979**	**0.8491**	**4.09**	**4.30**	**4.28**

Table 3: Verse-Bench Set-2/3 comparison on all 395 cases, with shared reference-derived screenplay content adapted to each method's official prompt format.

Method	ID↑	FD↓	KL↓	CS↑	CE↑	CU↑	PQ↑	WER↓	LSE-C↑
OVI	0.843	0.938	1.992	0.259	3.775	5.931	5.985	0.107	6.079
MOVA	0.826	0.953	1.716	0.285	3.986	6.088	6.327	0.141	7.043
NAVA	**0.847**	0.922	1.669	0.258	3.974	6.254	6.626	0.105	7.675
daVinci	0.744	0.988	2.009	0.260	**4.219**	6.102	6.422	0.101	6.903
LTX-2.3	0.806	0.964	1.814	0.296	3.810	5.822	6.398	0.087	**8.909**
CastDirector	0.825	**0.761**	**1.357**	**0.331**	4.169	**6.344**	**6.692**	**0.074**	8.801

with SenseVoice (An et al., 2024), and SyncNet confidence (LSE-C) (Chung & Zisserman, 2016; Prajwal et al., 2020). Prompt construction and metric accounting appear in Sections A.8 and A.9.

Character and event execution. Director-Bench contains 40 manually verified screenplay cases spanning live action, animation, animals, English, Chinese, Japanese, and Korean dialogue, off-screen speech, and silence. Reference clips and all training clips from the same source videos are excluded from our training sets. We report FID (Heusel et al., 2017), FVD (Unterthiner et al., 2018), ArcFace ID (Deng et al., 2019), and WER, alongside two turn-level diagnostics: Turn Sync-C measures SyncNet confidence for the requested speaker within each dialogue turn, and Speaker Match checks whether that speaker has the highest confidence among visible candidates. Both turn-level metrics use the same 31 cases with visible frontal human faces, excluding off-screen, back-facing, and non-human cases. Both metrics average turns within each case, then average cases equally. Unalignable or unscorable turns receive zero (Section A.10). A blinded study with 20 participants rates action order and completion, dialogue order and completion, and the suitability of sound effects and vocal timbre on a scale from 1 to 5 (Section A.11). This benchmark complements Verse-Bench by evaluating character binding and screenplay execution.

4.3 Character Binding and Screenplay Execution

In our blinded study, CastDirector receives the highest observed mean human ratings for action, dialogue, and audio among the compared methods (Table 2). Relative to LTX-2.3, action rises from 3.67 to 4.09, dialogue from 3.74 to 4.30, and audio from 3.82 to 4.28. These criteria cover execution by the specified characters as well as the suitability of sound effects and vocal timbre.

Automatic measurements provide complementary evidence for dialogue binding. Speaker Match improves from 0.7500 to 0.8491, Turn Sync-C from 3.4091 to 3.8979, and WER decreases from 0.1192 to 0.0769. CastDirector also has the lowest FID and FVD in this comparison, while ArcFace ID improves slightly over LTX-2.3 and remains below MOVA. The gains in speaker assignment and synchronization are consistent with the higher human dialogue ratings.

4.4 General Audio-Video Quality

In the Verse-Bench comparison (Table 3), CastDirector obtains the best FD, PaSST KL, CLAP, AudioBox CU and PQ, and WER among the listed methods, with second-best CE and LSE-C. Under

Table 4: Ablations on LTX-2.3 and OVI. Director (text) matches Phase 1 except that its Director receives no image. Frozen Director (continued) matches Phase 2 training steps with the Director frozen. Both phases use the same data. Lower WER and higher values of the other metrics are better. Bold marks the best value within each backbone.

Setting	LTX-2.3					OVI				
	Verse-Bench		Director-Bench			Verse-Bench		Director-Bench		
	WER	LSE-C	WER	Turn Sync-C	Speaker Match	WER	LSE-C	WER	Turn Sync-C	Speaker Match
Baseline	0.087	**8.909**	0.1192	3.4091	0.7500	0.107	6.079	0.4823	1.9479	0.6316
+ Screenplay	0.092	8.480	0.0826	3.5091	0.7672	0.124	7.007	0.2917	2.2171	0.6765
+ Director (text)	0.082	8.621	0.0778	3.5926	0.7793	0.122	7.148	0.2892	2.3217	0.6871
+ Frozen Director (Phase 1)	0.084	8.548	0.0782	3.6566	0.7985	0.126	7.220	0.2924	2.4383	0.6955
+ Frozen Director (continued)	0.109	8.420	0.1089	3.4638	0.7740	0.145	7.109	0.3095	2.3069	0.6753
+ Adaptation (Phase 2)	**0.074**	8.801	**0.0769**	**3.8979**	**0.8491**	**0.092**	**7.439**	**0.2759**	**2.5509**	**0.7160**

the shared LTX-2.3 refinement pipeline, image identity improves from 0.806 to 0.825, although several baselines remain stronger on this metric. LSE-C decreases slightly from 8.909 to 8.801. The results show gains in audio quality and speech accuracy alongside the character-binding improvements, without a uniform gain on every metric.

4.5 Screenplay and Director Ablation

We compare five shared settings on LTX-2.3 and OVI: the native backbone with rewritten prompts, screenplay conditioning through the native text encoder, a text-only Director, the multimodal Director after alignment (Phase 1), and Director Adaptation (Phase 2). Table 4 also includes continuation controls with a frozen Director. Full metrics appear in Sections A.12 and A.13.

Visual conditioning. Compared with the text-only Director, the multimodal Director in Phase 1 raises Director-Bench Speaker Match from 0.7793 to 0.7985 on LTX-2.3 and from 0.6871 to 0.6955 on OVI, with higher Turn Sync-C on both. These results support joint visual-text conditioning for character binding, although WER increases slightly on both benchmarks.

LTX-2.3 adaptation. Phase 2 further raises Speaker Match from 0.7985 to 0.8491 and Turn Sync-C from 3.6566 to 3.8979. The complete model lowers WER on both benchmarks and improves Verse-Bench LSE-C from 8.548 in Phase 1 to 8.801, slightly below the native baseline's 8.909.

Transfer to OVI. We attach a Qwen3-VL-4B Director to OVI's audio and video towers using learned depth mappings and the same two-phase training principle (Section A.7). Speaker Match rises from 0.6316 at baseline to 0.7160 after Phase 2, while Turn Sync-C rises from 1.9479 to 2.5509. The final model reduces Director-Bench WER from 0.4823 to 0.2759 and improves Verse-Bench LSE-C from 6.079 to 7.439. Verse-Bench WER reaches 0.092, below the baseline's 0.107.

Controlling for training duration. Phase 2 improves Speaker Match over frozen continuation on LTX-2.3 (0.8491 vs. 0.7740) and OVI (0.7160 vs. 0.6753). The Phase 1 checkpoint and all other training settings are matched within each backbone (Section A.13).

4.6 Qualitative Evaluation

Figure 4 compares CastDirector with four baselines on two scenes using the same conditioning frame and directing content, serialized for each model as described in Section A.8. Full screenplays appear in Section A.16. In panel (a), CastDirector keeps the green tool with the receiver in black as the giver in white exits right. LTX-2.3 and MOVA also move the receiver away, NAVA sends the giver left, and daVinci loses the tool from view. In panel (b), CastDirector and MOVA show the requested orange character exiting left while the blue character stays. LTX-2.3 and daVinci move the blue character out of view, and NAVA retains the orange character. These examples illustrate character-specific action execution. Dialogue ratings and turn-level metrics assess speech and synchronization separately.

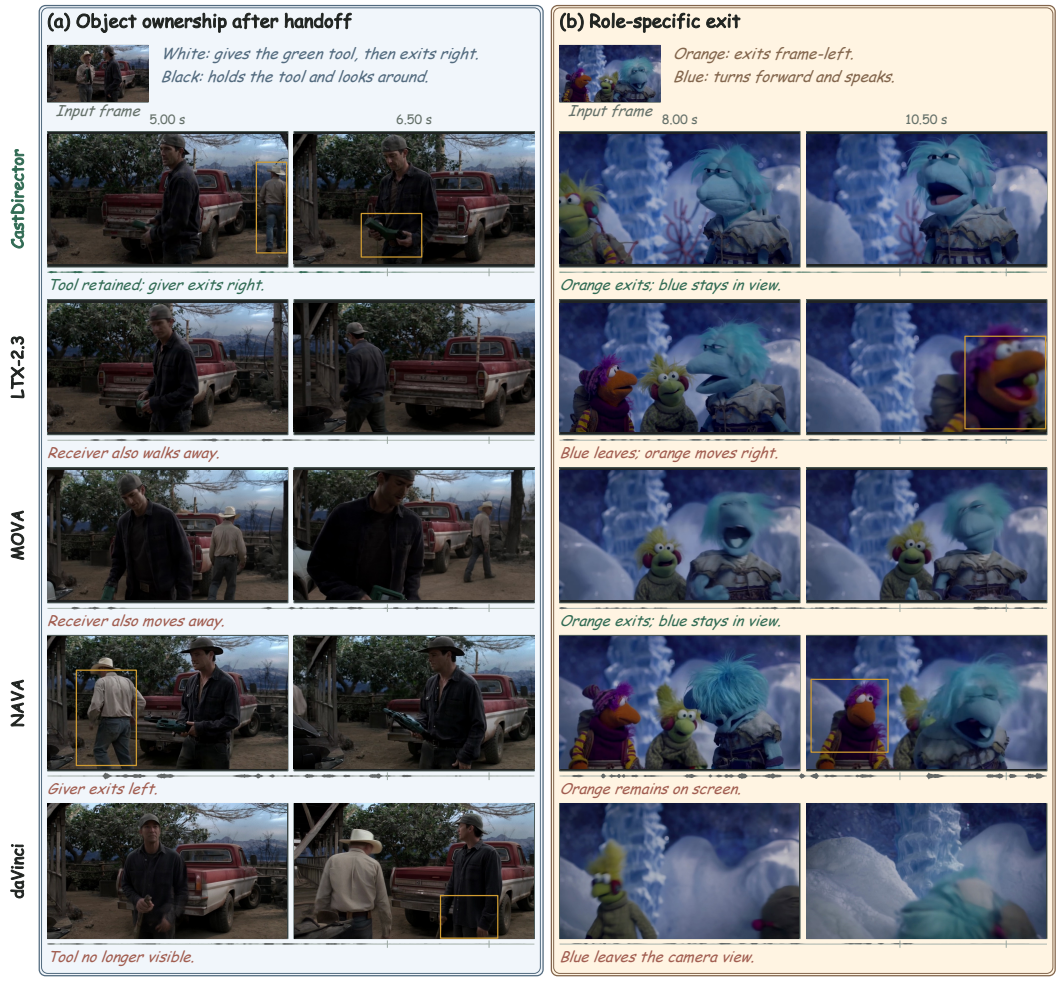

Figure 4: Qualitative comparisons with four baselines. Each panel shows the conditioning frame, requested actions, and generated frames at two shared timestamps. (a) The character in white gives the green tool to the character in black, then exits frame-right. The receiver should retain the tool. (b) The orange character exits frame-left while the blue character turns forward and speaks. Yellow boxes highlight relevant characters or objects. Annotations summarize the observed outcomes. Full prompts for both panels appear in Section A.16.

Broader audiovisual examples. The supplementary demos showcase coordinated character performances and immersive audio. Full videos with sound are available on our anonymous demo page[1], complementing the still frames in Fig. 1.

5 Conclusion

We presented CastDirector for binding screenplay instructions to visible characters in cinematic audio-video generation. Its grounded screenplay makes dialogue, actions, event order, and audiovisual cues explicit. The multimodal Director jointly encodes this screenplay and the first frame, supplies conditions to both generation streams, and is adapted through the generation loss to support screenplay execution. Results show improved speaker binding and human ratings for action, dialogue, and audio, with examples of coordinated performances and scene-consistent sound. These advances bring joint audio-video generation a step closer to realizing its potential for film and television production.

[1] https://completely-anonymous-z.github.io/video-generation-demo/

AI USE STATEMENT

Large language models (LLMs, including models from the GPT, Opus, and Fable families) were used to assist code writing and manuscript polishing. These assistants were not used to formulate the research ideas or design the experiments. Separately, Qwen3-ASR, Keye-VL, and Gemini were used in the screenplay annotation pipeline. Gemini also rewrote the shared screenplay content into each baseline's official prompt format.

REPRODUCIBILITY STATEMENT

We will release the inference code as open source as soon as possible. For now, please refer to our demo page[2] for generation results.

DATA STATEMENT

Our training corpus draws overwhelmingly on publicly downloadable academic datasets: OpenHumanVid (Li et al., 2024), MovieNet (Huang et al., 2020), PTVD (Li et al., 2023), HowTo100M (Miech et al., 2019), SpeakerVid-5M (Zhang et al., 2026), TalkVid (Chen et al., 2026b), and VGGSound (Chen et al., 2020), with a small supplement collected and curated internally. The material includes substantial film and television content as well as virtual human data, and undergoes manual review and AI-based filtering and annotation. We obtain the public datasets from their official research distributions and use them subject to the providers' terms and access conditions.

REFERENCES

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. *arXiv preprint arXiv:2407.04051*, 2024.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *Interspeech*, 2023.

Yiyang Cai, Nan Chen, Rongchang Xie, Junwen Pan, Chunyang Jiang, Cheng Chen, Wen Zhou, Zhenbang Sun, Wei Xue, Wenhan Luo, et al. Homie: Human-object centric video personalization via multimodal intelligent enchancement. *arXiv preprint arXiv:2607.18217*, 2026.

Hejia Chen, Haoxian Zhang, Xu He, Xiaoqiang Liu, Pengfei Wan, Shoulong Zhang, and Shuai Li. MindFlow: Harmonizing cognitive semantics and acoustic dynamics for facial animation generation in dyadic conversations. *arXiv preprint arXiv:2606.27779*, 2026a.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725. IEEE, 2020.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Guanbin Li, Ser-Nam Lim, Harry Yang, and Benyou Wang. TalkVid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3492–3500, 2026b.

[2] https://completely-anonymous-z.github.io/video-generation-demo/

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*, pp. 251–263. Springer, 2016.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. PaddleOCR 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

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.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio and video generation through cross-task synergy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16085–16095, 2026.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2020.

Xiaohu Huang, Hao Zhou, Qiangpeng Yang, Shilei Wen, and Kai Han. Jova: Unified multimodal learning for joint video-audio generation. *arXiv preprint arXiv:2512.13677*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025b.

Longbin Ji, Guan Wang, Xuan Wei, Chenye Yang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Jingzhou He. Native audio-visual alignment for generation. *arXiv preprint arXiv:2605.30073*, 2026.

Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. OmniHuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint arXiv:2508.19209*, 2025.

Zeqian Ju, Dongchao Yang, Kai Shen, Yichong Leng, Zhengtao Wang, Songxiang Liu, Xinyu Zhou, Tao Qin, Xiangyang Li, Jianwei Yu, and Xu Tan. MoonCast: High-quality zero-shot podcast generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 6283–6317, 2025.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 1–10. IEEE, 2025.

Koray Kavukcuoglu, Jeff Dean, Oriol Vinyals, and Noam Shazeer. Gemini 3.5: Frontier intelligence with action. `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/`, May 2026. Google Blog. Accessed: September 1, 2026.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech 2019*, pp. 2350–2354. ISCA, 2019. doi: 10.21437/INTERSPEECH.2019-2219.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70990–71013, 2025.

Khaled Koutini, Jan Schlueter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, pp. 2753–2757, 2022.

Kwai Keye Team, Bin Wen, Changyi Liu, Chengru Song, Chongling Rao, Guowang Zhang, Han Li, Haonan Fan, Hengrui Ju, Jiankang Chen, et al. Kwai Keye-VL-2.0 technical report. *arXiv preprint arXiv:2606.10651*, 2026.

Chen Li, Xutan Peng, Teng Wang, Yixiao Ge, Mengyang Liu, Xuyuan Xu, Yexin Wang, and Ying Shan. PTVD: A large-scale plot-oriented multimodal dataset based on television dramas. *arXiv preprint arXiv:2306.14644*, 2023.

Chunyu Li, Jiaye Li, Ruiqiao Mei, Haoyuan Xia, Hao Zhu, Jingdong Wang, and Siyu Zhu. HalloLive: Real-time streaming joint audio-video avatar generation with asynchronous dual-stream and human-centric preference distillation. *arXiv preprint arXiv:2604.23632*, 2026.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. *arXiv preprint arXiv:2412.00115*, 2024.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13848–13858, 2025.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

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.

Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, 2026.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J. Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

Yanzuo Lu, Ronglai Zuo, and Jiankang Deng. RAVEN: Real-time autoregressive video extrapolation with consistency-model GRPO. *arXiv preprint arXiv:2605.15190*, 2026.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision*, pp. 2630–2640. IEEE, 2019.

MiniMax. MiniMax-H3: A general-purpose omni-modal generative system, 2026. URL `https://github.com/MiniMax-AI/MiniMax-H3`. Open-weight release; accessed 2026-08-31.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. MOVA: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Liangyang Ouyang, Ruicong Liu, Caixin Kang, Yifei Huang, and Yoichi Sato. SocialDirector: Training-free social interaction control for multi-person video generation. *arXiv preprint arXiv:2605.10079*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492. ACM, 2020. doi: 10.1145/3394171.3413532.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Le Shen, Qian Qiao, Tan Yu, Ke Zhou, Tianhang Yu, Yu Zhan, Zhenjie Wang, Ming Tao, Shunshun Yin, and Siyuan Liu. SoulX-FlashTalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation. *arXiv preprint arXiv:2512.23379*, 2025.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026.

SII-GAIR, Sand.ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. OmniForcing: Unleashing real-time joint audio-visual generation. *arXiv preprint arXiv:2603.11647*, 2026.

Ye Tao, Lupeng Liu, Xuenan Xu, Jiasun Feng, Jiarui Wang, Ying Qin, Shuiyang Mao, Wei Liu, and Shuai Wang. Foley-omni: A unified multimodal generation model from task-level audio synthesis to complete video soundtrack generation. *arXiv preprint arXiv:2606.03672*, 2026.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote portrait alive: Generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Andrew Z. Wang, Songwei Ge, Tero Karras, Ming-Yu Liu, and Yogesh Balaji. A comprehensive study of decoder-only LLMs for text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28575–28585. IEEE, 2025a.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. UniVerse-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025b.

Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. DreamRunner: Fine-grained compositional story-to-video generation with retrieval-augmented motion adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10503–10511, 2026.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Jingyun Xue, HongFa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Towards multiple character image animation through enhancing implicit decoupling. In *International Conference on Learning Representations*, 2025.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. InfiniteTalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 37:47455–47487, 2024.

Tan Yu, Qian Qiao, Le Shen, Ke Zhou, Jincheng Hu, Dian Sheng, Bo Hu, Haoming Qin, Jun Gao, Changhai Zhou, et al. SoulX-FlashHead: Oracle-guided generation of infinite real-time streaming talking heads. *arXiv preprint arXiv:2602.07449*, 2026.

Leying Zhang, Yao Qian, Xiaofei Wang, Manthan Thakker, Dongmei Wang, Jianwei Yu, Haibin Wu, Yuxuan Hu, Jinyu Li, Yanmin Qian, and Sheng Zhao. CoVoMix2: Advancing zero-shot dialogue generation with fully non-autoregressive flow matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 128915–128938, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3813–3824, 2023.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S^3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 192–201, 2017.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. SpeakerVid-5M: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. In *International Conference on Learning Representations*, 2026.

Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. MuseTalk: Real-time high-fidelity video dubbing via spatio-temporal sampling. *arXiv preprint arXiv:2410.10122*, 2024.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, et al. AnyTalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

A Appendix

A.1 Grounded Screenplay Details

The grounded screenplay is serialized as tagged text. Figure 5 shows the conditioning image for case 08 of Director-Bench, paired with the complete screenplay below; Table 1 summarizes its hierarchical fields.



Figure 5: First-frame input paired with the corridor screenplay. The woman occupies the left foreground, facing the man in a dark coat and scarf on the right, against a dim green-toned corridor with a wall lamp. The screenplay binds the man's dialogue and the woman's later exit to these visible characters; the requested speech and motion are specified by the screenplay.

Connecting screenplay fields to the image. The `[SCENE]` section describes the corridor and lighting visible in the first frame. Within `[VISIBLE_MAIN_CHARACTER]`, the `<first>` entry describes the woman on the left and the `<second>` entry describes the man on the right; appearance, posture, and frame position provide complementary cues for associating each entry with the image. The `[EVENT_SEQUENCE]` then specifies how these characters should act and speak. Its first event pairs `speaker_1` with the identity phrase "man in dark coat, scarf, holding a firearm" and the line "I'm sorry." Thus the speaker ID refers to the second cast entry: the explicit identity phrase establishes the association, independently of cast ordering. The next event asks the woman to exit frame-left while the man turns his head to his own left, distinguishing a frame-relative exit from an anatomical turn. These are instructions for the generated sequence, rather than actions already established by the still image.

Shared visual and acoustic context. The dialogue's `<voice>` field requests a soft, apologetic male delivery, while `[AUDIO_CAPTION]` specifies a low ambient hum, a soft male whisper, and a dramatic synth pad. The atmosphere field further describes the intended tense, cinematic mood. Joint encoding exposes the Director to these acoustic instructions together with the visible speaker and setting, producing the shared context from which the audio and video adapters construct their conditions (Section 3.3).

Complete serialized screenplay.

```
[SCENE]
  dimly lit concrete corridor; dark green tones,
  wall lamps; medium shot
[END_SCENE]
[VISIBLE_MAIN_CHARACTER]
  <count>2</count>
```

```
  <first>
    <appearance>woman with long dark hair,
    dark jacket</appearance>
    <posture>standing, facing the man</posture>
    <frame_position>frame-left</frame_position>
    <actions>moves leftward, exiting the frame</actions>
  </first>
  <second>
    <appearance>man in dark coat, scarf,
    holding a firearm</appearance>
    <posture>standing upright</posture>
    <frame_position>frame-right</frame_position>
    <actions>looks at the woman, then turns his
    head to his own left</actions>
  </second>
[END_VISIBLE_MAIN_CHARACTER]
[EVENT_SEQUENCE]
  <first>
    <dialogue>
      <voice>adult male, soft and apologetic</voice>
      <character_id>speaker_1: man in dark coat,
      scarf, holding a firearm</character_id>
      <words: English>I'm sorry.</words>
    </dialogue>
    <event>The woman with long dark hair and dark
    jacket stands facing the man in the
    dimly lit corridor.</event>
    <lighting>dim green-toned lighting with wall
    lamps</lighting>
    <camera_motion>The camera pulls out with small
    amplitude at slow speed.</camera_motion>
  </first>
  <second>
    <event>The woman moves leftward and exits the
    frame, while the man turns his head
    to his own left.</event>
    <lighting>unchanged dim lighting</lighting>
    <camera_motion>The camera continues to pull out
    at slow speed.</camera_motion>
  </second>
[END_EVENT_SEQUENCE]
[AUDIO_CAPTION]
  low ambient hum, soft male whisper,
  dramatic synth pad
[END_AUDIO_CAPTION]
[SUBJECT]["indoor", "people", "action"][END_SUBJECT]
[ATMOSPHERE]"tense, cinematic, somber"[END_ATMOSPHERE]
```

The prompt requires all six outer sections and stable speaker IDs. The production validators concentrate on failure modes that can be checked reliably from text: malformed/truncated tags, invalid event/dialogue nesting, ASR violations, and conflicts between voice and identity attributes. Invalid records are regenerated rather than silently truncated, because losing a late event would change the requested control target.

A.2 Raw Audiovisual Data Processing

We begin with approximately 5,000 hours of audiovisual material drawn overwhelmingly from OpenHumanVid (Li et al., 2024), MovieNet (Huang et al., 2020), PTVD (Li et al., 2023), HowTo100M (Miech et al., 2019), SpeakerVid-5M (Zhang et al., 2026), TalkVid (Chen et al., 2026b), and VGGSound (Chen et al., 2020), with a small internally collected and curated supplement. This initial pool contains substantial film and television content alongside virtual human and other audiovisual data. The collected material is curated through manual review and AI-based filtering operators; below we summarize the shared quality-control stages rather than dataset-specific extrac-

tion details. The public datasets are obtained through their official research distributions, subject to the providers' terms and access conditions. Before any screenplay annotation, every candidate clip passes through separate visual and audio quality gates. A clip advances to annotation only when both modalities satisfy all of their respective requirements.

Visual filtering. We first apply an aesthetic classifier following the LAION (Schuhmann et al., 2022) aesthetic-prediction setup to screen out visually poor material. A heuristic black-border detector identifies and crops extraneous borders so that subsequent operators act on the content-bearing image region. We then use PaddleOCR (Cui et al., 2025) to detect burned-in subtitles and remove clips containing them. To avoid geometric distortion, we inspect the pixel sample aspect ratio (SAR) and retain only clips satisfying $|\mathrm{SAR} - 1| \leq 0.01$, which permits at most a 0.01 deviation from square pixels. Finally, we derive motion statistics from CoTracker3 (Karaev et al., 2025) point trajectories and reject clips dominated by abrupt, large-amplitude global motion, thereby filtering severe camera shake.

Audio filtering. We align the audio-side procedure with the filtering protocol of Foley-Omni (Tao et al., 2026). Specifically, we score each original soundtrack with the unified AudioBox Aesthetics (Tjandra et al., 2025) audio-quality predictor, which assesses speech, music, and general sound within a common model, and retain clips whose audio-quality score is at least 0.6. This gate is applied to the complete soundtrack rather than to a separated speech or effects stem.

Data-volume accounting. The joint audiovisual filtering stage reduces the raw pool from approximately 5,000 to 3,000 hours. The surviving clips are then sent to the evidence-separated annotation pipeline described in Sections A.3 and 3.2. During annotation, clips are excluded because of unclosed or malformed label structures, Gemini API refusals, or failed Qwen3-ASR or Keye-VL annotations. The final training corpus contains 1,459,594 successfully filtered and annotated samples totaling 2,177 hours. Director Alignment and Director Adaptation use this same complete corpus. The raw and intermediate pool durations are approximate; the final sample count and duration describe the shared training inventory.

Train and evaluation separation. We freeze the evaluation inventories before constructing the final training manifests and enforce isolation at both the source and identity levels. For every test case used in this paper, its reference clip, conditioning first frame, and every candidate training clip from the same source video are excluded from training. We additionally encode every detectable human identity in the test sets and all indexed training face tracks with ArcFace (Deng et al., 2019), then search the training index for identity overlap. This audit detects no identity overlap between the test and training faces examined. The source-level exclusion continues to apply to cases without a detectable human face and to non-human cases for which an ArcFace identity comparison is not meaningful.

Content and annotation limitations. The source datasets retain their original academic curation, while our additional gates remove low-quality audiovisual material, burned-in subtitles, severe camera shake, malformed annotations, and items rejected by the annotation services. These stages are not an exhaustive guarantee that all sensitive cinematic content or social stereotypes are absent, and such residual content remains a limitation of the corpus.

The screenplays are automatically produced by ASR, vision-language, and multimodal models and can inherit their language, visibility, speaker-attribution, and descriptive biases. We mitigate these errors through evidence-separated inputs, closed-schema validation, consistency checks, regeneration, and rejection of failed records, as detailed in Section A.3. At this scale, we perform periodic human spot checks (Section A.3). The automatically generated screenplays serve as structured training supervision and are not exhaustively verified per clip.

The retained corpus also spans a broad range of scene subjects. Figure 6 summarizes the distribution of its most frequent subject tags and pairs it with representative first frames from the curated corpus. Because each clip may carry several subject tags, the chart reports shares of tag assignments rather than mutually exclusive shares of videos.

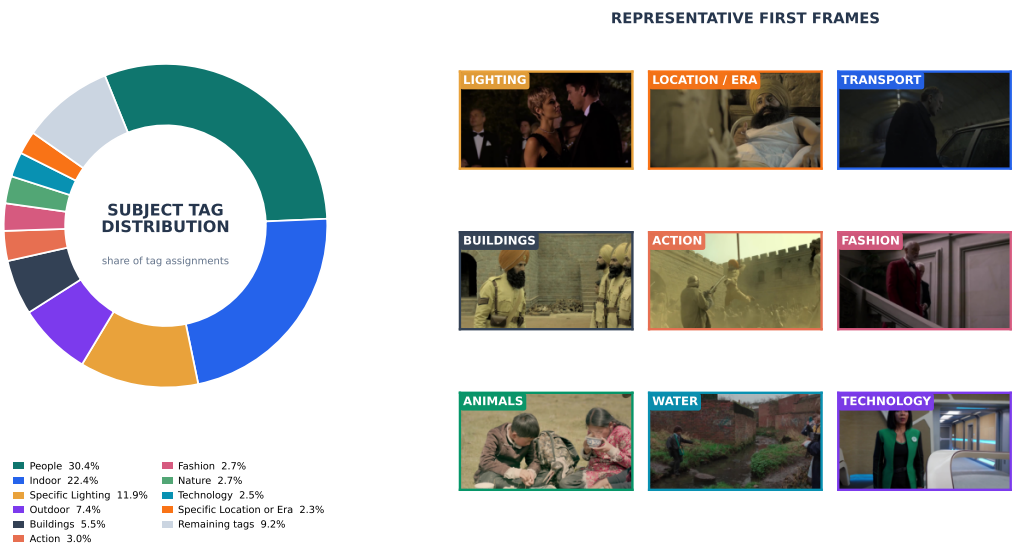

Figure 6: Subject-tag distribution and representative first frames. The donut chart shows the share of all subject-tag assignments accounted for by the ten most frequent tags, with the remaining tags aggregated into one slice. These percentages describe tag assignments rather than videos because the taxonomy is multi-label. The right panel presents a 3×3 grid of representative first frames for nine distinct subject tags selected from the curated corpus.

A.3 Annotation Pipeline Implementation

Training requires screenplays for millions of clips. Gemini (Kavukcuoglu et al., 2026), when used alone for end-to-end annotation without external priors, can produce fluent descriptions while misrecognizing speech or introducing unsupported visual details. Our pipeline instead provides an audio prior and a visual prior before multimodal verification (Eq. (1) and Fig. 3). Qwen3-ASR (Shi et al., 2026) supplies the audio prior, consisting of the transcript and language label. Keye-VL-2.0 (Kwai Keye Team et al., 2026) supplies the visual prior, describing the visible cast and chronological actions from silent video. Gemini then uses both priors and the full clip to construct the screenplay, checking proposed actions against the video and assigning dialogue using the heard voice and visible lip evidence. The following paragraphs describe the prompt constraints, decoding settings, and validators.

Visual prior. We run Kwai Keye-VL-2.0-30B-A3B with the audited v3.2 visual prompt. The model samples the video at 2 fps; each frame is allocated between 128×28^2 and 512×28^2 input pixels, with a clip-level budget of $180 \times 1024 \times 28^2$ pixels. Decoding uses temperature 0.1, top-$p = 0.001$, top-$k = 1$, and at most 6144 tokens. The production deployment uses one tensor-parallel-size-1 service per H200, with eight services per node and six nodes. This affects throughput, not annotation semantics.

The Keye prompt prohibits audio, dialogue, subtitle transcription, mouth-motion descriptions, and chain-of-thought output. It requests exactly two blocks: `VISIBLE_MAIN_CHARACTER`, with a count and closed fields for appearance, posture, frame position, and action, and a chronological event-sequence block. Events must name a concrete actor rather than use pronouns. Frame-left/right is permitted only for frame position; body-side labels must be anatomical, and turns preferentially name a visible target instead of an ambiguous direction.

The Keye parser removes leaked thinking text, normalizes common angle-tag spacing errors, and extracts only supported blocks. The v3.2 gate requires both blocks and their canonical end tags, nonempty bodies, a positive character count, at least one appearance/ordinal field, balanced critical angle tags, and no truncated tail. A failed item is retried once on another endpoint. Valid outputs are deduplicated by absolute video path and exported as {`video, keye_event_auxiliary`}; a later successful retry takes precedence.

Audio prior. The input manifest stores the audio prior as a Qwen ASR transcript in `qwenasr` and a detected language label. TV and movie manifests use Qwen3-ASR (Shi et al., 2026); other sources use the corresponding Qwen ASR annotations available in their manifests. An empty transcript records that ASR returned no spoken text; it is retained separately from missing annotation metadata. Auxiliary records are joined by canonicalized absolute video path only; basename matching is forbidden because independent datasets may reuse filenames.

Constrained Gemini fusion. Gemini samples the original clip at 5 fps and receives text before the video payload, with an audio-prior block and, when available, a visual-prior block. Multi-person, movie, and TV subsets use Gemini-3.5-Flash (Kavukcuoglu et al., 2026); the remaining subsets use Gemini-3-Flash.[3] Production decoding uses temperature 0, top-$p = 1$, top-$k = 1$, and the high reasoning effort setting, which we found necessary for reliable speaker attribution from voice. These settings favor repeatability, although service-side decoding is not guaranteed to be bit-exact.

The fusion prompt instructs Gemini to preserve the spoken words and language supplied by the audio prior and to treat the visual prior as a set of candidate observations to verify. It explicitly requires checking proposed motions against the clip, discarding unsupported actions, and performing a two-pass audio check before choosing a speaker. Each dialogue turn must contain `<voice>`, `<character_id>`, and `<words: LANGUAGE>` in that order. The voice field includes a gender cue, optional age class, and a tone/emotion phrase. When audio and imagery disagree, audio controls the acoustic fields and visible evidence controls the scene/event fields. An unavailable visual prior is treated as a soft failure; SyncNet and legacy human-face count priors are never injected.

Validation, regeneration, and resumption. Programmatic checks reject malformed or truncated tags, invalid event nesting, missing dialogue fields, violations of the audio prior, and coarse voice-identity attribute conflicts. The final gate requires structured-description-only output and an event sequence with angle-bracket nested fields and at least one event. For every spoken turn it checks the voice/character/words triplet, the language label, absence of bare speech, and coarse gender/age compatibility between the heard voice and character description. It also rejects dialogue unsupported by the audio prior, inconsistent language labels, Latin words introduced into Chinese speech, and duplicated Chinese characters beyond their transcript counts. A content failure triggers one fresh generation; exhausted failures are recorded, not admitted to training. These checks constrain text-verifiable errors; visual actions and speaker binding rely on the preceding multimodal verification. An append-only JSONL and SQLite WAL state database make success-key resumption deterministic and prevent duplicate API work.

Human spot checks. During annotation, we periodically inspect Gemini-produced screenplays against their source clips to monitor quality. Only a small fraction of the inspected samples exhibit errors such as incorrect speaker assignments or left-right confusions. These errors are difficult to detect reliably with automated models, and affected samples are discarded when identified during human inspection.

A.4 LTX Implementation Details

Primary model. Our main system uses the 22B-parameter LTX-2.3-dev backbone and a Qwen3-VL-8B Director. We train at 1088×640 resolution and 24 fps with length buckets from 113 to 169 frames. The native video and audio VAE latents and Gemma text conditions are computed offline; the Director contexts are computed online from the VAE-aligned first frame and complete grounded screenplay. The effective global batch size is 128 on 32 H200 GPUs, with one sample per GPU and four gradient accumulation steps.

Training. The curated training corpus comprises 1,459,594 samples totaling 2,177 hours, drawn overwhelmingly from public academic audiovisual datasets with a small internally collected supplement. It contains substantial film and television content together with virtual human and other audiovisual material, and is cleaned through manual review and AI-based filtering operators. The raw

[3]The cited model card documents the 3.x series. Gemini 2.5 (Gemini Team, 2025) provides the most recent published technical report for the model family. We record the exact served model identifiers and access dates in our internal annotation logs.

audiovisual filtering procedure and the data-volume accounting through annotation are detailed in Section A.2. Director Alignment and Director Adaptation use the same complete corpus and the same screenplays produced by the evidence-separated annotation pipeline in Section 3.2. The adaptation phase uses learning rates of 1×10^{-5} for the DiT updates, 2×10^{-5} for the injection updates, and 2×10^{-6} for Qwen LoRA. In both phases, every sample follows the strict I2VA construction in Section 3.4.

The primary Director provides features of width 4096 and supports a maximum sequence length of 2560. Video and audio each have their own learned feature mixture, LayerNorm, width projection, and per-block cross-attention. Attention masks are forwarded to LTX and padded feature positions are zeroed. The output projection of every newly added attention module is initialized to zero.

During Director Alignment (Phase 1), LTX attention uses rank-64 LoRA with scale 64 (427.8M parameters); all Qwen weights are frozen, while the two layer mixtures (33.6M) and the Director cross-attention branches of both streams in all 48 blocks (3.23B for video and 1.21B for audio) are trained in full. This configuration has 4.90B trainable parameters, dominated by the newly inserted full-width attention modules. Parameter counts are global totals across trainable modules, rather than the parameter shards held by an individual GPU. During Director Adaptation (Phase 2), the Director receives rank-32 LoRA with scale 64 (87.3M parameters), and the LTX LoRA continues training. Both streams' layer mixtures, LayerNorms, width projections, and Director cross-attention branches remain fully trainable, initialized from their Phase 1 weights. As illustrated in Fig. 7, the multimodal Director and the LTX backbone are adapted through LoRA, while the multi-layer audio/video adapter and the two streams' Director cross-attention branches receive full parameter updates. The native Gemma text encoder remains frozen. The audio-video flow-matching loss jointly updates the Director LoRA, the full conditioning interface, and the LTX LoRA. The final learning rates are 10^{-5} (DiT), 2×10^{-5} (injection), and 2×10^{-6} (Qwen), with 500 warmup steps and gradient clipping at 0.5. Director dropout is 0.1. It never drops the clean first-frame video condition.

Input-length audit. Across the 16 training metadata files, screenplay lengths have a median of 1,870 characters, a 95th percentile of 2,778, a 99th percentile of 3,172, and a maximum of 3,683. We additionally process the 64 longest screenplays by character count with the Qwen processor, including the 1088×640 image, task instruction, and chat template. These complete inputs contain 1,435 to 1,512 tokens, all below the 2,560-token limit. The token check covers these 64 inputs, while the character statistics cover all 16 metadata files.

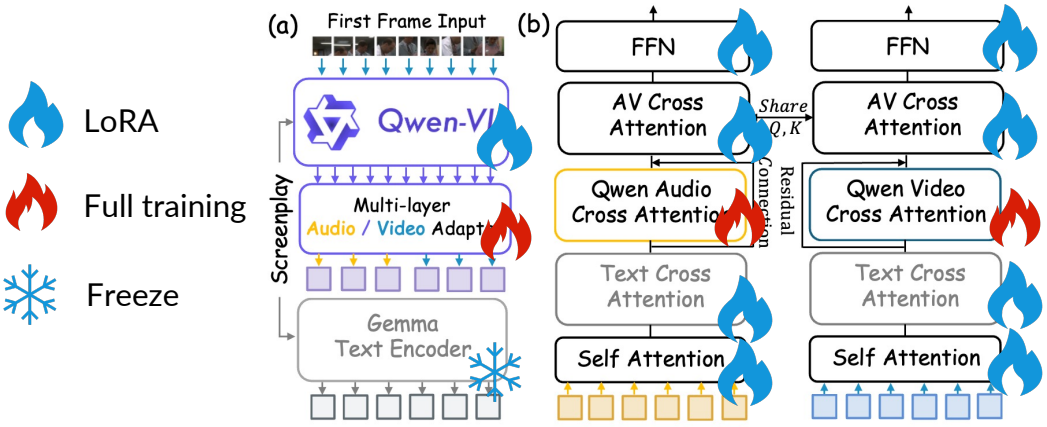

Figure 7: Phase 2: Director Adaptation. Blue flames denote LoRA updates to Qwen-VL and the LTX backbone; red flames denote full training of the multi-layer audio/video adapter and the Director cross-attention branches in both streams. The adapter includes the learned layer mixtures, normalization, and width projections. The snowflake marks the frozen Gemma text encoder. The audio-video generation loss trains the Director's conditioning features jointly with the modules that map and inject them into the generator.

Objective. We retain the backbone's flow-matching objective (Lipman et al., 2023). For clean stream latents Z_m, Gaussian noise ϵ_m, and a flow time $t \in [0, 1]$ shared by both streams, the interpolated latent $Z_{m,t}$ and its target velocity U_m for each generated token are

$$Z_{m,t} = (1 - t)\, Z_m + t\, \epsilon_m, \qquad U_m = \epsilon_m - Z_m, \tag{5}$$

and the network is trained to predict U_m from $Z_{m,t}$. Equation equation 5 applies to all audio tokens and to the video tokens after the conditioning frame. The video tokens that encode the first frame remain clean throughout denoising and are excluded from the video loss.

Every sample is trained as image-to-audio-video generation: the clean first frame is present in the video stream. Let $\widehat{U}_m$ be the predicted velocity and G_v mask out the clean first-frame video tokens. The loss is

$$\mathcal{L} = \frac{\|G_v \odot (\widehat{U}_v - U_v)\|_2^2}{\sum G_v} + \frac{1}{|U_a|}\|\widehat{U}_a - U_a\|_2^2. \tag{6}$$

The video and audio terms are combined without additional weighting. The flow time t is shared by both streams and drawn from a shifted logit-normal distribution whose shift μ depends on the number of video tokens; at 1088×640 the eight length buckets covering 113 to 169 frames use μ from 4.24 to 5.94. Dropping both Director contexts never removes the clean first-frame video tokens, so a sample never degenerates into text-to-audio-video training.

A.5 Inference Procedure

Director encoding. Given the conditioning image and the screenplay, the first frame is resized and center-cropped to the generation resolution, exactly as the video VAE input, and passed through the Director once together with the task instruction and the screenplay. The Director features are fused by the two learned mixtures into C_v and C_a. This is a single vision-language forward pass per clip; its cost does not depend on the number of denoising steps, and the resulting contexts are reused by every step and every block.

Stage 1: directed generation. CastDirector denoises video and audio latents jointly at the training resolution of 1088×640 and 24 fps for 30 Euler steps, with the clean first-frame latent held fixed in the video stream. Guidance follows the native LTX configuration: classifier-free guidance with scale 4.0 and spatiotemporal guidance (STG) with scale 1.0 applied to block 29 of both streams. The conditional pass receives the Gemma encoding of the screenplay, the Director contexts C_v and C_a, and the first frame. The unconditional pass receives the Gemma encoding of a generic negative prompt ("worst quality, inconsistent motion, blurry, jittery, distorted"), omits the Director contexts, and keeps the first frame. Guidance therefore contrasts both the native text condition and the Director contexts, while preserving the first-frame condition in both passes.

Stage 2: upsampling and refinement. The Stage-1 video latent is upsampled $2\times$ spatially by the official LTX-2.3 latent upsampler. The upsampled video latent and the Stage-1 audio latent are then re-noised to the first sigma of the official four-step distilled schedule and refined jointly by the official LTX-2.3-dev transformer with the official distilled LoRA, without classifier-free guidance, STG, or any Director parameters; the clean first frame remains a conditioning input. The final 2176×1280 video is decoded by the video VAE, and the audio latent is decoded by the audio VAE and vocoder. All numbers reported for CastDirector and for the LTX-2.3 baseline use this same two-stage pipeline, so their differences arise in Stage 1, including its conditioning interface and training.

A.6 Efficiency and Resource Cost

We isolate the two costs introduced by the Director instead of timing the complete generation pipeline, whose shared VAE, upsampling, and video encoding stages would obscure the difference. All measurements use batch size one and bfloat16 on one NVIDIA H200, with model loading excluded from the timed region. For the denoising microbenchmark, we preload all conditions and time one conditional Euler step at a matched resolution of 544×320 for 121 frames, which keeps both variants resident without weight offloading. Classifier-free guidance and STG are disabled so that each measurement contains exactly one DiT forward. We discard five warm-up runs and report the

Table 5: Isolated inference cost on one NVIDIA H200. Values are mean ± sample standard deviation. The denoising rows compare one matched conditional DiT forward; Qwen3-VL encoding is an additional once-per-clip operation used only by CastDirector.

Measurement	LTX-2.3	CastDirector	Difference
One-step denoise time (s)	0.2409 ± 0.0022	0.3711 ± 0.0024	$+0.1302$
One-step peak memory (GiB)	36.012 ± 0.000	47.552 ± 0.000	$+11.540$
Qwen3-VL encoding time (s)	n/a	16.34 ± 2.31	$+16.34$
Qwen3-VL peak memory (GiB)	n/a	18.313 ± 0.000	$+18.313$

mean and sample standard deviation over 20 measured runs. The Qwen3-VL feature encoder is profiled separately with the full structured prompt and the actual 1088×640 conditioning frame, using three warm-up runs followed by 15 measured runs. Every timed region is bounded by CUDA synchronization, and peak allocated memory is reset before each run.

The injected branches add 0.1302 seconds and 11.540 GiB to one matched DiT forward. Qwen3-VL encoding is performed once and its features are reused across all denoising steps. The Qwen encoder is moved off the GPU before the DiT runs, so its 18.313 GiB peak is a separate phase and must not be added to the CastDirector denoising peak.

A.7 OVI Implementation Details

The OVI integration preserves the same semantic interface but respects the topology of the generator. It uses Qwen3-VL-4B with feature width 2560.

The exact DiT-preprocessed first frame is resized to 1280×720 for the Director. Because OVI has separate video and audio towers, each tower learns a depth mapping from the Director features to its target blocks. Every target block receives an independent zero-initialized Director cross-attention gate. Director Alignment (Phase 1) trains rank-64 LoRA on OVI's native attention plus the new injection modules. Director Adaptation (Phase 2) applies rank-16 LoRA with scale 32 to the Director.

A.8 Evaluation Inputs and Prompt Construction

Director-Bench contains 40 image-to-audio-video cases spanning multi-character scenes, live action, animation, animals, English, Chinese, Japanese, and Korean speech, voiceover, and silence. Each case is built from a real reference clip: its first frame is the conditioning image, its screenplay is produced by the annotation pipeline of Section 3.2 and manually verified, and the clip itself serves as the reference for FID and FVD. The reference clips, their first frames, and all clips from the same source videos are excluded from every training set used in this paper. Prompts include character-specific dialogue and actions as well as ordered events, exposing plausible-looking but incorrectly directed generations.

Prompt construction and parity. Gemini 3.5 (Kavukcuoglu et al., 2026) rewrites the same screenplay into each baseline's specific input format: a single English paragraph for LTX-2.3, three-part English prose for daVinci, multi-paragraph prose for MOVA, a Chinese cinematic description with tagged speech for NAVA, and English prose with tagged speech and a trailing audio description for OVI. Spoken lines are copied verbatim, and a programmatic check rejects any rewrite in which a line is missing or altered. We also manually reviewed every method's inputs to verify that they preserve the same scene facts, plot events, character-specific actions, and sound descriptions, including speaker assignments and event order. CastDirector receives the screenplay itself. The rewriting and manual content checks apply to both benchmarks (Section 4.2).

A.9 Verse-Bench Accounting

Our Verse-Bench evaluation comprises 395 videos from Set-2 and Set-3 and uses the official image+text-to-audio-video toolkit of UniVerse-1 (Wang et al., 2025b) in a screenplay-conditioned setting. Reference clips are annotated as screenplays using the training annotation pipeline (Section 3.2).

Table 6: Verse-Bench metric eligibility for the 395-case evaluation.

Metric	Dir.	Eligibility
ID	↑	generated video exists
FD, KL	↓	Set-2 audio prompt and reference audio
CS	↑	Set-2 audio prompt
CE, CU, PQ	↑	generated Set-2 audio
WER	↓	reference speech text
LSE-C	↑	speech text and a face detected by SyncNet

CastDirector receives these directly, while Gemini 3.5 rewrites them into each baseline's specific input format (Section A.8). This comparison uses the manually verified shared screenplay content in place of the original generation prompts. It measures how well each system realizes the same dialogue, character actions, event order, and audio descriptions. Prompt serialization follows each model's expected interface, while the requested content is matched across methods. All Verse-Bench results in this paper use this protocol. All methods successfully generate outputs for all 395 cases. Each metric is computed on the applicable cases according to the eligibility conditions in Table 6. We report the official image identity (ID), audio Fréchet distance (FD), PaSST KL, CLAP score (CS), AudioBox enjoyment (CE), usefulness (CU), quality (PQ), WER, and LSE-C.

We extract mono PCM audio from each mp4 before official scoring. ID is the mean DINOv3 (Siméoni et al., 2025) cosine similarity between the conditioning image and generated frames. FD compares generated and ground-truth Set-2 audio in CLAP (Elizalde et al., 2023) feature space, while KL compares their PaSST (Koutini et al., 2022) category distributions. CS is CLAP similarity between generated audio and the audio prompt. CE, CU, and PQ are the three reported AudioBox Aesthetics (Tjandra et al., 2025) outputs. WER uses SenseVoice (An et al., 2024) against `speech_prompt.text`, and LSE-C uses SyncNet.

A.10 Director-Bench Scoring Implementation

General metrics and eligibility. FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) use the complete benchmark. ArcFace (Deng et al., 2019) ID uses live-human cases with a detectable reference face, and WER uses cases with reference speech and Whisper large-v3 (Radford et al., 2023) transcriptions. After Unicode NFKC normalization, lowercasing, and punctuation/symbol removal, English uses whitespace words and Chinese uses characters. We report $\sum_i e_i / \sum_i r_i$, where e_i and r_i are edit and reference-unit counts.

Turn Sync-C and Speaker Match. The two turn-level metrics separate two questions that clip-level lip synchronization conflates: whether the requested character is synchronized during a dialogue turn, and whether that character is more synchronized than the other visible characters. Both metrics use a fixed subset of 31 cases featuring live human actors with visible frontal faces. Off-screen, back-facing, and non-human cases are excluded from these two metrics. The same subset is used for every method, and generation failures do not remove a case or a requested turn from scoring. For clip i, let $\mathcal{E}_i$ contain all requested dialogue turns and let turn j request speaker s_{ij}. WhisperX (Bain et al., 2023) monotonically aligns the requested sentence to ASR word timestamps to obtain a padded temporal window W_{ij}. If generated speech cannot be aligned, the turn remains in $\mathcal{E}_i$ and receives zero for both metrics.

The spatial association is computed independently. S3FD (Zhang et al., 2017) detections are linked over time, gaps are interpolated, and box centers and scales are median-filtered before extracting face and mouth crops. CLIP (Radford et al., 2021) compares sampled face crops with the visible speaker profiles in the screenplay. A global one-to-one assignment $\pi_i(s)$ binds every visible speaker s to at most one track, preventing two characters from claiming the same face. We manually reviewed every candidate face box and every speaker-to-track assignment across all evaluated outputs to verify the identity of each candidate. SyncNet scoring follows this spatial verification. After separating the vocal stem, SyncNet is evaluated on every bound track inside every aligned

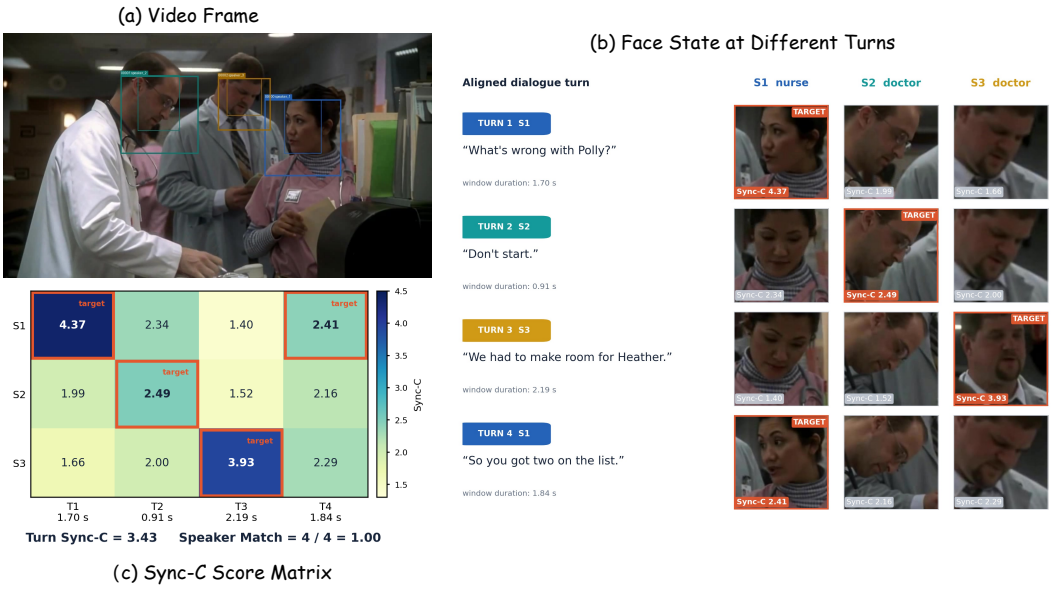

Figure 8: Illustration of Turn Sync-C and Speaker Match on a hospital dialogue case, assembled from assets exported by our evaluation viewer. Panel (a) shows an evaluated video frame with the detected face and mouth regions and the one-to-one mapping between tracks and screenplay speakers. Panel (b) expands each aligned dialogue turn into its text, temporal-window duration, and the three face states sampled at the turn midpoint. Orange borders indicate the requested speaker and each crop is annotated with its SyncNet confidence. Panel (c) arranges the same speaker-by-turn confidences as a matrix, with orange cells marking the requested sequence S1, S2, S3, and S1. The requested speaker is the top-scoring candidate in every turn, yielding Speaker Match 1.00. Averaging the four requested-speaker confidences equally gives Turn Sync-C 3.30.

window. We denote its confidence by

$$c_{ijs} = \mathrm{SyncNetC}\big(A_i[W_{ij}],\, V_{i,\pi_i(s)}[W_{ij}]\big)\,, \tag{7}$$

Here $A_i[W_{ij}]$ denotes the vocal audio within the aligned window, and $V_{i,\pi_i(s)}[W_{ij}]$ denotes the mouth crops from the corresponding bound face track.

For a scorable turn, the predicted speaker is the candidate with the largest confidence, $\hat{s}_{ij} = \arg\max_s c_{ijs}$. Its synchronization score is $q_{ij} = c_{ij\hat{s}_{ij}}$, and its match indicator is $a_{ij} = \mathbf{1}[\hat{s}_{ij} = s_{ij}]$. If alignment fails or the requested speaker has no bound or scorable track, we set $q_{ij} = a_{ij} = 0$. Turn Sync-C and Speaker Match, shortened to Match in some displays, average all requested turns with equal weight:

$$C_i^{\mathrm{turn}} = \frac{1}{|\mathcal{E}_i|}\sum_{j\in\mathcal{E}_i} q_{ij}, \qquad M_i^{\mathrm{spk}} = \frac{1}{|\mathcal{E}_i|}\sum_{j\in\mathcal{E}_i} a_{ij}. \tag{8}$$

Each benchmark score is the uniform mean of its clip scores over the same 31 cases. There is no minimum-confidence gate for a valid finite matrix: even when all candidates have weak evidence, Match follows the top-scoring candidate. Failed turns remain in both denominators, and a clip with no scorable turns receives zero for both metrics. The output retains turn-level status counts and eligible coverage so that these decisions can be audited. Speaker Match is a relative diagnostic based on SyncNet confidence and should be interpreted together with Turn Sync-C. A correct speaker ranking can occur even when all candidates have low synchronization confidence.

As Fig. 8 shows, panel (b) exposes the per-turn face states from which the compact matrix in panel (c) is formed. A high confidence on the orange requested-speaker crop raises Turn Sync-C, but Speaker Match additionally requires that crop to outrank the other two candidates in the same row of panel (b), equivalently the other cells in the same column of panel (c). Consequently, a generator that animates several mouths together may retain a plausible Turn Sync-C yet lose Match when a non-speaking character receives the highest confidence.

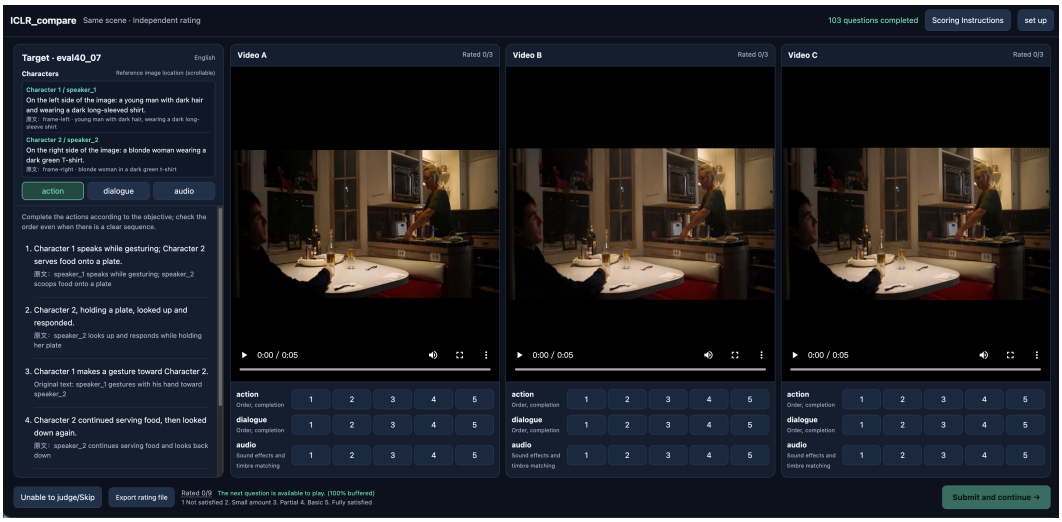

Figure 9: Human evaluation interface. Target instructions and persistent character descriptions appear on the left. Three anonymous outputs and their action, dialogue, and audio rating controls appear on the right. Each criterion is rated independently from 1 to 5.

A.11 Human Evaluation on Director-Bench

Interface and procedure. The study involves 20 participants, including university students from language programs. All participants receive the same training on the evaluation task and scoring criteria. Tasks are randomly assigned to participants. Each task presents three outputs for the same case, sampled from the six methods in Table 2, with method identities hidden and their positions randomized. The interface (Fig. 9) places the target instructions beside the videos and keeps the scoring controls below each video. Character descriptions remain visible above the action, dialogue, and audio tabs, identifying each character by appearance and position. Chinese translations accompany the original instructions. Raters can replay the clips, consult each target dimension, and score the outputs independently before continuing to the next task. They may skip a task they cannot judge.

Criteria and aggregation. Each output receives one score per applicable criterion. Action assesses the order and completion of requested actions by the specified characters. Dialogue assesses the order and completion of lines by the specified speakers. Audio assesses whether sound effects and vocal timbre match the scene and voice descriptions. Scores range from 1 (not satisfied) to 5 (fully satisfied), with intermediate levels indicating slight, partial, and substantial fulfillment. Dialogue is inapplicable when no spoken line is requested. Every method covers all 40 cases, with at least five ratings for each method-case combination. The study yields 2,386 ratings each for action and audio and 1,817 for dialogue. For each method and criterion, we first average the available ratings within each applicable case and then average the case means with equal weight to obtain Table 2. Skipped tasks contribute no ratings, and cases without requested speech are excluded from dialogue averaging.

A.12 Full Ablation Results

Table 7 reports all automatic metrics for the five settings shared by LTX-2.3 and OVI in Table 4. The additional continuation controls are detailed in Section A.13. The OVI implementation is detailed in Section A.7. Human-evaluation criteria are excluded because those results belong to a separate blinded study. The shared settings are Baseline, Screenplay, Director (text), Director (Phase 1), and Adaptation (Phase 2). Screenplay uses the native text encoder with the structured input. Director (text) encodes the screenplay without an image, while Director (Phase 1) jointly encodes the screenplay and first frame. All other training and inference settings are matched between these two variants. The generator retains its native first-frame condition in both variants.

On Director-Bench, adding the image to the Director input raises LTX-2.3 Speaker Match from 0.7793 to 0.7985 and Turn Sync-C from 3.5926 to 3.6566. For OVI, the corresponding scores rise

Table 7: Full metrics for the five ablation settings shared by LTX-2.3 and OVI on Verse-Bench and Director-Bench. The additional continuation controls appear in Table 8. The text-only Director omits the image from its input, while Phase 1 uses both modalities. For FID, FVD, FD, KL, and WER, lower values are better. Higher values are better for all other metrics. Within each backbone, **bold** and underline denote the best and second-best results.

(a) Verse-Bench

Setting	ID↑	FD↓	KL↓	CS↑	CE↑	CU↑	PQ↑	WER↓	LSE-C↑
OVI									
Baseline	0.843	0.938	1.992	0.259	3.775	5.931	5.985	0.107	6.079
+ Screenplay	0.850	0.935	1.999	**0.266**	3.875	6.102	6.326	0.124	7.007
+ Director (text)	0.846	0.947	1.994	0.259	3.922	6.116	6.329	0.122	7.148
+ Director (Phase 1)	0.849	0.942	1.989	0.260	3.941	6.122	6.320	0.126	7.220
+ Adaptation (Phase 2)	**0.860**	**0.921**	**1.875**	0.263	**3.961**	**6.186**	**6.350**	**0.092**	**7.439**
LTX-2.3									
Baseline	0.806	0.964	1.814	0.296	3.810	5.822	6.398	0.087	**8.909**
+ Screenplay	0.812	0.828	1.605	0.291	3.922	5.979	6.500	0.092	8.480
+ Director (text)	0.819	0.772	1.498	0.312	4.013	6.193	6.571	0.082	8.621
+ Director (Phase 1)	**0.828**	**0.758**	1.363	0.329	4.112	6.323	6.680	0.084	8.548
+ Adaptation (Phase 2)	0.825	0.761	**1.357**	**0.331**	**4.169**	**6.344**	**6.692**	**0.074**	8.801

(b) Director-Bench

Setting	FID↓	FVD↓	ID↑	WER↓	Turn Sync-C↑	Speaker Match↑
OVI						
Baseline	134.85	40.19	0.3486	0.4823	1.9479	0.6316
+ Screenplay	131.65	**40.11**	0.4113	0.2917	2.2171	0.6765
+ Director (text)	133.21	42.57	0.4005	0.2892	2.3217	0.6871
+ Director (Phase 1)	131.14	41.87	0.4128	0.2924	2.4383	0.6955
+ Adaptation (Phase 2)	**127.79**	40.85	**0.4177**	**0.2759**	**2.5509**	**0.7160**
LTX-2.3						
Baseline	114.34	34.79	0.4951	0.1192	3.4091	0.7500
+ Screenplay	113.27	32.55	0.4860	0.0826	3.5091	0.7672
+ Director (text)	114.76	33.51	0.4826	0.0778	3.5926	0.7793
+ Director (Phase 1)	113.92	31.42	0.4975	0.0782	3.6566	0.7985
+ Adaptation (Phase 2)	**108.19**	**28.28**	**0.4992**	**0.0769**	**3.8979**	**0.8491**

from 0.6871 to 0.6955 and from 2.3217 to 2.4383. The text-only variant has slightly lower WER on both benchmarks for both backbones, despite its lower turn-level binding and synchronization scores. Phase 2 further improves these two metrics and yields the lowest WER in all four backbone and benchmark combinations. It updates Director LoRA together with the generator LoRA and conditioning modules, so this comparison measures the complete adaptation phase. On LTX-2.3, Phase 2 improves seven of the nine Verse-Bench metrics over Phase 1, while Phase 1 retains better ID and FD. For OVI, the final variant improves all reported Verse-Bench metrics over its native baseline.

A.13 Continued Training with a Frozen Director

Controlled comparison. We evaluate whether improvements after Phase 1 can be explained by additional training while the Director remains frozen, using both LTX-2.3 and OVI. Within each backbone, the control and Phase 2 start from the same Phase 1 checkpoint and use the same training data, number of optimizer updates, and all other hyperparameters. The control keeps the Director frozen while continuing the generator LoRA and conditioning-module updates used in the corresponding Phase 2. The only experimental difference is whether the Director receives LoRA updates. Both runs use the same inference and evaluation settings. Both backbones are evaluated on Verse-Bench and Director-Bench.

Table 8 includes the Phase 1 checkpoints as references. For LTX-2.3 on Director-Bench, Phase 2 improves Speaker Match from the frozen continuation's 0.7740 to 0.8491 and Turn Sync-C from

Table 8: Continuation with and without Director adaptation on LTX-2.3 and OVI. Within each backbone, the two continuation runs start from the same Phase 1 checkpoint and match all other hyperparameters and training steps. Phase 1 is included as the starting reference. **Bold** and underline indicate the best and second-best values among the three settings within each backbone.

(a) Verse-Bench

Setting	ID↑	FD↓	KL↓	CS↑	CE↑	CU↑	PQ↑	WER↓	LSE-C↑
OVI									
Phase 1	0.849	0.942	1.989	0.260	3.941	6.122	6.320	0.126	7.220
Frozen Director (continued)	0.844	0.948	1.888	0.257	3.910	6.118	6.256	0.145	7.109
Adaptation (Phase 2)	**0.860**	**0.921**	**1.875**	**0.263**	**3.961**	**6.186**	**6.350**	**0.092**	**7.439**
LTX-2.3									
Phase 1	**0.828**	**0.758**	1.363	0.329	4.112	6.323	6.680	0.084	8.548
Frozen Director (continued)	0.818	0.771	1.4977	0.322	4.005	6.157	6.548	0.109	8.420
Adaptation (Phase 2)	0.825	0.761	**1.357**	**0.331**	**4.169**	**6.344**	**6.692**	**0.074**	**8.801**

(b) Director-Bench

Setting	FID↓	FVD↓	ID↑	WER↓	Turn Sync-C↑	Speaker Match↑
OVI						
Phase 1	131.14	41.87	0.4128	0.2924	2.4383	0.6955
Frozen Director (continued)	132.77	42.05	0.4001	0.3095	2.3069	0.6753
Adaptation (Phase 2)	**127.79**	**40.85**	**0.4177**	**0.2759**	**2.5509**	**0.7160**
LTX-2.3						
Phase 1	113.92	31.42	0.4975	0.0782	3.6566	0.7985
Frozen Director (continued)	116.24	33.67	0.4896	0.1089	3.4638	0.7740
Adaptation (Phase 2)	**108.19**	**28.28**	**0.4992**	**0.0769**	**3.8979**	**0.8491**

3.4638 to 3.8979. It also lowers WER from 0.1089 to 0.0769, with better FID, FVD, and identity scores. On Verse-Bench, WER decreases from 0.109 to 0.074 and LSE-C rises from 8.420 to 8.801. Phase 2 also improves the other seven Verse-Bench metrics over the frozen continuation. The LTX-2.3 frozen continuation performs worse than its starting Phase 1 checkpoint on all reported metrics in this experiment.

For OVI on Verse-Bench, Phase 2 improves all nine metrics over the frozen continuation, including lower WER (0.092 versus 0.145) and higher LSE-C (7.439 versus 7.109). The OVI frozen continuation improves KL over Phase 1 (1.888 versus 1.989), while its other eight metrics are worse. On Director-Bench, Phase 2 improves all six metrics over the OVI frozen continuation. Speaker Match rises from 0.6753 to 0.7160 and Turn Sync-C from 2.3069 to 2.5509, while WER decreases from 0.3095 to 0.2759. Together, these controls support adapting the Director during continued training for speech accuracy and character binding on both backbones at a fixed number of updates.

A.14 Global and Face-Localized Text Attention as Binding Diagnostics

We use native audio-to-text attention to examine one possible mechanism behind the different baseline binding accuracy of LTX-2.3 and OVI. This is a diagnostic association rather than a causal attribution. The example in Fig. 10 contains two visible characters in a bedroom and three ordered dialogue turns. The man initially holds a newspaper, folds it while the woman speaks, and then delivers the final line. Both backbones receive the same conditioning first frame, semantic content, seed, and sampling configuration through their respective native prompt serializations.

We reuse cached query and key tensors and recompute the attention in every native block as

$$A_{\ell qhk} = \mathrm{softmax}_k\left(\frac{Q_{\ell qh} K_{\ell hk}^{\mathsf{T}}}{\sqrt{d_h}}\right), \tag{9}$$

where ℓ, q, h, and k index the block, query, head, and text key, respectively. For the global audio diagnostic, we first average over all audio queries and heads,

$$a_{\ell k}^{\text{audio}} = \frac{1}{Q_\ell H_\ell} \sum_{q=1}^{Q_\ell} \sum_{h=1}^{H_\ell} A_{\ell q h k}, \tag{10}$$

renormalize over the retained text keys within each block, and then average the distributions over blocks. Following the time-chain view used by our diagnostic tool, we retain the text from the sentence containing the first utterance through the final utterance. The panels use the earliest saved positive late-denoising call and average all native audio blocks. Speech mass is the sum over alphanumeric subword keys inside the three utterance spans; quotation marks, speech delimiters, and punctuation-only keys are excluded. This definition is independent of the display threshold.

For complementary spatial analysis, we also compute face-localized video-to-text attention. We detect the face box in each frame and map it to the spatiotemporal support B_q of every video query. For face box B_f, the query weight is its spatial coverage multiplied by its temporal overlap with the face track,

$$w_q = \max_f \left(\frac{|B_q \cap B_f|}{|B_q|} \cdot \tau(q, f) \right), \qquad a_{\ell k}^{\text{face}} = \frac{\sum_{q,h} w_q A_{\ell q h k}}{\sum_{q,h} w_q}, \tag{11}$$

where $\tau(q, f)$ is the temporal-overlap fraction. Fractional coverage preserves queries that cross a face boundary, and the maximum prevents double counting when tracks overlap. We exclude the clean conditioning-frame video queries by default, renormalize over text keys in each block, and average over blocks. This face-localized calculation applies only to spatial video queries; the audio panels in Fig. 10 use the global computation in Eq. (10) and do not use face crops.

The word colors in Fig. 10 use one shared scale for both baseline panels: the floor and cap are the pooled P40 and P98 of the displayed word scores, and the contrast exponent is $\gamma = 1.6$. Consequently, color differences can be compared directly across the two models rather than only within a panel.

Within this shared speaking-sequence scope, the LTX-2.3 baseline assigns 34.23% of its native audio-to-text attention to spoken lexical content, compared with 22.71% for the OVI baseline. LTX-2.3 also distributes high attention across more words and all three turns, while OVI leaves much of the first and third turns below the shared display floor. This behavior is consistent with LTX-2.3's stronger native Speaker Match and provides a possible explanation for its better starting-point binding accuracy. It does not establish that attention allocation alone causes the metric difference.

Across repeated trained-model runs, the mean attention assigned to speech text increases by 8.72% for LTX-2.3 and 7.24% for OVI, where the relative gain is $g = (M_{\text{after}} - M_{\text{before}})/M_{\text{before}}$. The aggregate result indicates that screenplay training increases speech-focused attention in both backbones, with a slightly larger relative increase for LTX-2.3. We report only these average changes because the corresponding trained per-token traces were not retained, and we do not infer a token-specific redistribution from the aggregate values.

A.15 Additional Comparison with H3

We additionally compare with the H3-Base FL2VA model (MiniMax, 2026) on both evaluation sets (Table 9). H3 receives the conditioning first frame and the output of its official Context-IR module, while CastDirector receives the first frame and screenplay itself. We manually verified that the Context-IR input preserves the same dialogue, characters, actions, event order, and sound descriptions as the other methods.

H3 generation settings. We run the local FL2VA model through vLLM-Omni on eight H200 GPUs. Generation uses a 768-pixel short edge, preserves the conditioning image's aspect ratio, and runs at 24 fps for 50 sampling steps. The video and audio flow shifts are 12 and 3, respectively, and the default random seed is 42. The requested duration follows each case's metadata, using the frame count divided by the frame rate when no explicit duration is provided. The reported outputs come from the 768p Base model without the H3-Regenerate-2K stage.

H3 obtains better Director-Bench FID and ArcFace ID, as well as Verse-Bench DINOv3 identity and PaSST KL. CastDirector achieves higher Turn Sync-C (3.8979 versus 3.2670) and Speaker Match

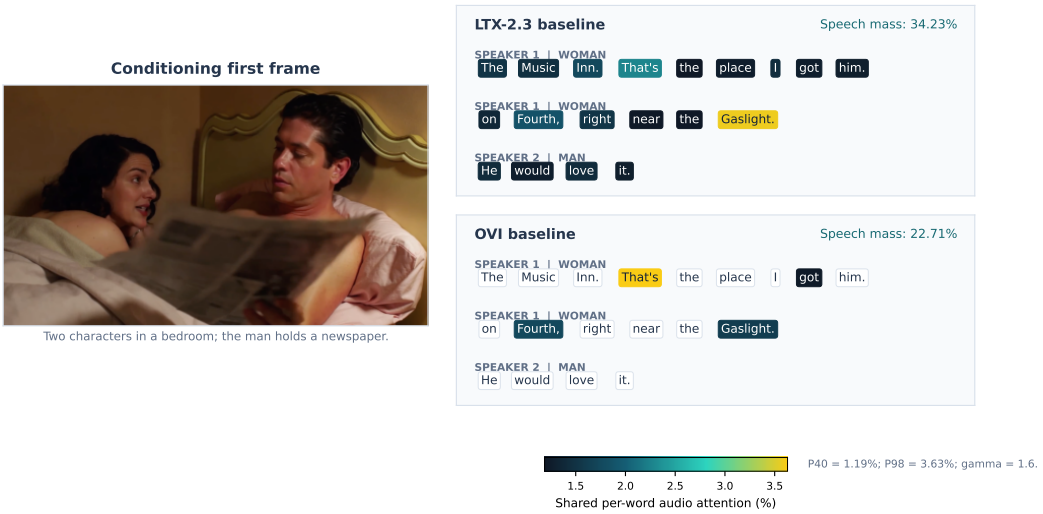

Figure 10: Native audio-to-text attention on the newspaper dialogue example, shown with its conditioning first frame. The conditioning frame occupies the left column, while the right column compares measured attention from the unmodified LTX-2.3 and OVI baselines on the same two-character, three-turn content. Speech mass is computed from the underlying attention values and is unaffected by highlighting thresholds. Word colors encode absolute attention share using a single pooled scale across both baseline panels. Values are averaged over all native audio blocks for the earliest saved positive late-denoising call. The comparison is an attention diagnostic, not a causal attribution.

Table 9: Head-to-head comparison with H3 on both benchmarks. **Bold** and underline mark the best and second-best values in each row among the three systems. H3 leads on several visual metrics, while CastDirector leads on the reported turn-level binding and synchronization metrics.

Benchmark	Metric	H3	LTX-2.3	CastDirector
Director-Bench	FID↓	**102.04**	114.34	108.19
	FVD↓	34.47	34.79	**28.28**
	ID↑	**0.5636**	0.4951	0.4992
	WER↓	0.1564	0.1192	**0.0769**
	Turn Sync-C↑	3.2670	3.4091	**3.8979**
	Speaker Match↑	0.6894	0.7500	**0.8491**
Verse-Bench	ID↑	**0.870**	0.806	0.825
	FD↓	0.892	0.964	**0.761**
	KL↓	**1.346**	1.814	1.357
	CS↑	0.275	0.296	**0.331**
	CE↑	4.012	3.810	**4.169**
	CU↑	6.164	5.822	**6.344**
	PQ↑	6.563	6.398	**6.692**
	WER↓	0.102	0.087	**0.074**
	LSE-C↑	8.260	**8.909**	8.801

(0.8491 versus 0.6894), with lower Director-Bench WER (0.0769 versus 0.1564). These results provide an additional comparison of generation quality and dialogue binding; differences in training and inference configurations prevent attributing the ranking to any single architectural choice.

A.16 Full Prompts for the Qualitative Comparison

The following grounded screenplays are the complete prompts for the two cases in Fig. 4, discussed in Section 4.6. CastDirector receives these structured descriptions directly, while Gemini 3.5 rewrites the same content into each baseline's specific input format (Section A.8). Line wrapping is applied only for readability, with all fields and dialogue preserved. In panel (a), the figure's labels "White" and "Black" refer to the men in the light cowboy hat and dark cap, respectively. In panel (b), "Orange" and "Blue" refer to the orange and blue puppets.

A.16.1 Object Ownership after Handoff

This prompt corresponds to panel (a) of Fig. 4.

```
Structured Description:
[SCENE]
  overcast outdoor ranch; red and white pickup truck, wooden fence; medium shot
[END_SCENE]
[VISIBLE_MAIN_CHARACTER]
  <count>2</count>
  <first>
    <appearance>man in light cowboy hat, beige long-sleeve shirt, blue jeans</appearance>
    <posture>standing, then walking away</posture>
    <frame_position>frame-left</frame_position>
    <actions>hands a green tool to the man in the dark cap, then walks away</actions>
  </first>
  <second>
    <appearance>man in dark cap, dark long-sleeve shirt, dark jeans</appearance>
    <posture>standing, then holding tool, shifting stance</posture>
    <frame_position>frame-right</frame_position>
    <actions>receives the green tool, holds it, looks around, shifts body position</actions>
  </second>
[END_VISIBLE_MAIN_CHARACTER]
[EVENT_SEQUENCE]
  <first>
    <dialogue>
      <voice>elderly male, dry and casual</voice>
      <character_id>speaker_1: man in light cowboy hat</character_id>
      <words: English>Back in the right place.</words>
    </dialogue>
    <event>speaker_1 hands the green tool to speaker_2</event>
    <lighting>overcast daylight</lighting>
    <camera_motion>The camera holds a static shot.</camera_motion>
  </first>
  <second>
    <dialogue>
      <voice>adult male, dry and conversational</voice>
      <character_id>speaker_2: man in dark cap</character_id>
      <words: English>Yeah, right next to the skill saw, in case the zombies come.</words>
    </dialogue>
    <event>speaker_1 walks away to the right while speaker_2 holds the tool and looks
    around</event>
    <lighting>unchanged overcast daylight</lighting>
    <camera_motion>The camera holds a static shot.</camera_motion>
  </second>
[END_EVENT_SEQUENCE]
[AUDIO_CAPTION]
  outdoor ranch ambience, rustling clothes, clear male dialogue
[END_AUDIO_CAPTION]
[SUBJECT]["outdoor", "people", "technology"][END_SUBJECT]
[ATMOSPHERE]"casual, slightly humorous"[END_ATMOSPHERE]
```

A.16.2 Role-Specific Exit

This prompt corresponds to panel (b) of Fig. 4.

```
Structured Description:
[SCENE]
  dimly lit icy cave; blue ice formations, snow; medium close-up
[END_SCENE]
[VISIBLE_MAIN_CHARACTER]
  <count>3</count>
  <first>
    <appearance>orange puppet with purple hair, purple scarf, yellow-purple striped sleeves,
    backpack</appearance>
    <posture>standing upright</posture>
    <frame_position>frame-left</frame_position>
    <actions>faces right, looking at the blue puppet</actions>
  </first>
  <second>
    <appearance>green puppet with yellow hair, red earmuffs, green sweater</appearance>
```

```
    <posture>standing upright</posture>
    <frame_position>frame-center</frame_position>
    <actions>faces right, looking at the blue puppet</actions>
  </second>
  <third>
    <appearance>blue puppet with light blue hair, tunic with ruffled edges</appearance>
    <posture>standing upright</posture>
    <frame_position>frame-right</frame_position>
    <actions>turns head left and right, speaking to the other puppets</actions>
  </third>
[END_VISIBLE_MAIN_CHARACTER]
[EVENT_SEQUENCE]
  <first>
    <dialogue>
      <voice>adult female, slightly annoyed and questioning</voice>
      <character_id>speaker_1: orange puppet with purple hair</character_id>
      <words: English>Hey, how come you didn't wait for me?</words>
    </dialogue>
    <event>speaker_1 speaks while looking at speaker_3</event>
    <lighting>cool blue ambient light with soft shadows</lighting>
    <camera_motion>The camera holds a static shot on the three puppets.</camera_motion>
  </first>
  <second>
    <dialogue>
      <voice>adult male, raspy and dismissive</voice>
      <character_id>speaker_3: blue puppet with light blue hair</character_id>
      <words: English>Well, if you're gonna walk slow, don't walk when I see Joe.</words>
    </dialogue>
    <event>speaker_3 turns his head to speak to speaker_1 and speaker_2</event>
    <lighting>unchanged cool blue lighting</lighting>
    <camera_motion>The camera holds a static shot.</camera_motion>
  </second>
  <third>
    <dialogue>
      <voice>adult male, raspy and excited</voice>
      <character_id>speaker_3: blue puppet with light blue hair</character_id>
      <words: English>Oh, I like that, Flomo.</words>
    </dialogue>
    <event>speaker_3 turns forward and speaks excitedly as speaker_1 exits frame-left</event>
    <lighting>unchanged cool blue lighting</lighting>
    <camera_motion>The camera holds a static shot.</camera_motion>
  </third>
[END_EVENT_SEQUENCE]
[AUDIO_CAPTION]
  raspy male and clear female puppet voices, soft cave ambience
[END_AUDIO_CAPTION]
[SUBJECT]["people", "indoor", "other"][END_SUBJECT]
[ATMOSPHERE]"playful, whimsical, fantasy"[END_ATMOSPHERE]
```

A.17 Future Work

CastDirector inherits the cost of a bidirectional diffusion generator and adds one Director forward pass, so it currently generates offline rather than interactively. On one H200, the Director adds 0.1302 seconds to a matched conditional DiT forward, while its once-per-clip Qwen3-VL encoding takes 16.34 ± 2.31 seconds; the corresponding peak-memory costs are reported in Table 5. Two lines of work suggest a route to real-time directed generation. The first reduces sampling steps by distillation: improved distribution matching distillation (Yin et al., 2024) shortens the trajectory for image synthesis, and adversarial distribution matching (Lu et al., 2025) extends the idea to video. The second reformulates generation autoregressively, extending a clip causally instead of denoising it as a whole: Self Forcing (Huang et al., 2025b) closes the train-test gap of autoregressive video diffusion, RAVEN (Lu et al., 2026) performs real-time autoregressive extrapolation under a consistency-model policy, and OmniForcing (Su et al., 2026) carries the formulation into real-time joint audio-video generation. In the audio-driven setting, SoulX-FlashTalk (Shen et al., 2025) uses self-correcting bidirectional distillation for streaming avatars, SoulX-FlashHead (Yu et al., 2026) develops oracle-guided real-time talking heads, and Hallo-Live (Li et al., 2026) combines asynchronous dual-stream generation with human-centric preference distillation.

A streaming Director must execute the global screenplay incrementally, preserve character IDs across chunks, and revise upcoming events when execution deviates from the plan. Other directions include longer scenes, larger casts, and overlapping speech, with CoVoMix2 (Zhang et al., 2025) providing a controllable speech-only formulation.