VORCH-HUMAN: UNIFIED TASK REPRESENTATION FOR HUMAN-CENTRIC AUDIO–VISUAL GENERATION AND LONG-HORIZON CONTINUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-centric audio–visual generation remains fragmented across task-specific conditioning interfaces, limiting supervision sharing and consistent control across generation modes. To address this fragmentation, we present Vorch-Human, a unified framework built on a dual-stream audio–video diffusion transformer. Its unified task representation organizes target and condition tokens using role embeddings, temporal position types, and clean-token masks, bringing text- and image-conditioned generation, audio-driven human animation, and appearance–voice referenced joint generation into a common framework. Building on this representation, cross-task condition composition combines persistent appearance and voice references with rolling temporal context to extend both audio-driven and reference-conditioned generation to long-video continuation. We train the model in two stages: multi-task mixed training establishes shared generation capabilities, while subsequent context adaptation teaches the model to continue generation from historical audio and video. At inference, the temporal context is held fixed within each generation window and updated between windows, while persistent identity references maintain identity consistency throughout the sequence. Extensive experiments and human evaluations demonstrate competitive visual quality, identity preservation, voice similarity, and audio–visual synchronization, together with stable five-minute long-horizon video generation. Project page: https://anonymous.4open.science/w/project-vorchhuman-VRCH/.

1 INTRODUCTION

Human-centric audio–visual generation now covers audio driven human animation, reference-based audio–visual generation, and long-video continuation. Although these tasks concern the same audio–visual scene, existing systems often use task-specific interfaces: driving audio controls motion directly, a reference waveform or image specifies voice identity or appearance, and long-video continuation incorporates generated history through a dedicated temporal conditioning interface. A single model that accepts heterogeneous conditions and preserves their semantics would reduce this fragmentation and allow supervision to be shared across tasks.

Recent models have expanded audio–visual generation capabilities. UniAVGen (Zhang et al., 2025a) supports audio driven synthesis, voice conditioned generation, and continuation. DreamID-Omni (Guo et al., 2026) unifies reference generation, editing, and animation, while Identity-as-Presence (Chen et al., 2026) explores paired appearance and voice conditioning. Building on these advances, we seek a single model for first-frame audio driven animation and appearance–voice referenced generation through a unified target–condition representation. We further explore how this representation incorporates historical audio and video as conditions to extend these generation modes to long video continuation with persistent identity references and rolling temporal context.

We present Vorch-Human, a unified human-centric generation framework that uses a dual-stream diffusion transformer to jointly model audio and video. We augment the audio and video target streams with clean acoustic and visual condition groups and describe every group using three complementary annotations. Role IDs encode semantic function and subject association. Temporal position types distinguish independent references, initial observations, targets, and history. Clean-token

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.

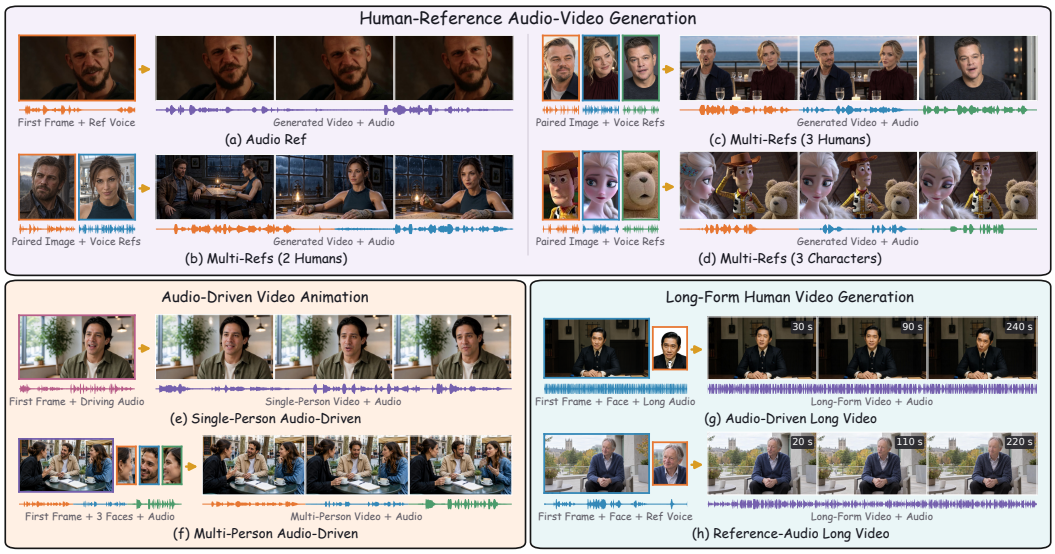


Figure 1: **Human-centric generation with Vorch-Human.** Vorch-Human uses a unified target–condition representation for text and image conditioned generation, audio driven human animation, reference conditioned audio–visual generation, and long video continuation.

masks separate fixed evidence from tokens to be denoised. These annotations provide a common interface for text and image conditioned generation, audio driven animation, and appearance–voice referenced generation. All configurations use the same target-only masked flow-matching objective.

The resulting representation composes conditions across generation modes. Appearance references can be paired with an observed driving waveform to predict video, or with voice references to jointly synthesize new speech and video from text. We further find that this condition composition mechanism can be extended to long video continuation by treating generated history as additional fixed conditions. Audio driven and timbre-referenced continuation retain video and audio–video context, respectively, to generate subsequent video or audio–video segments. In both cases, identity references persist while recurrent context is updated, allowing the same conditional interface to serve short clips and chunk-wise continuation without a separate long-video backbone.

We train Vorch-Human in two stages. Multi-task mixed training first learns shared generation capabilities under the unified representation. Context adaptation then presents historical video and audio latents as fixed conditions and trains the model to predict subsequent segments. At inference, each window holds its incoming context fixed during denoising, appends the new segment, and updates the context for the next window. Persistent appearance and voice references provide identity evidence throughout this recurrence, helping reduce accumulated long-horizon drift.

We evaluate short audio driven animation, paired appearance–voice referenced joint generation, and five-minute audio driven continuation using task-appropriate automatic metrics and human judgments, with representative generation modes illustrated in Figure 1. The results show competitive visual quality, identity preservation, voice similarity, and audio–visual synchronization, as well as stable identity and appearance over five-minute generation.

In conclusion, our main contributions are summarized as follows: (1) **Unified task representation.** We introduce a unified target–condition representation for text and image conditioned generation, audio driven human animation, reference conditioned audio–visual generation, and long video continuation. (2) **Cross-task condition composition for long-video continuation.** We combine persistent appearance references with driving audio or voice references and rolling context to support identity-conditioned continuation. (3) **Two-stage progressive training strategy.** We combine multi-task training with audio and video context adaptation for long-video continuation. (4) **Extensive experimental validation.** Quantitative comparisons and human evaluations demonstrate competitive generation quality, identity preservation, and audio–visual synchronization, with stable five-minute long video generation.

2 Related Work

Joint audio–visual generation. LaVie (Wang et al., 2024) and SEINE (Chen et al., 2024) establish video diffusion foundations, while Latte (Ma et al., 2025), CogVideoX (Yang et al., 2024), Hunyuan-Video (Tencent Hunyuan Team, 2024), Wan (Wan et al., 2025), and Seedance (Gao et al., 2025b) advance visual synthesis. Joint audio–visual models extend generation to synchronized sound and video. LTX-2 (HaCohen et al., 2026) uses a dual-stream transformer, Ovi (Low et al., 2025) couples diffusion backbones, and ALIVE (ByteDance ALIVE Team, 2026) and MagiHuman (Chern et al., 2026) explore joint scene and human generation. Our focus is a common conditioning interface that distinguishes driving speech, voice identity, appearance, and temporal context.

Conditional human generation. Human animation methods such as Leo (Wang et al., 2025) study human video synthesis. HunyuanVideo-Avatar (Chen et al., 2025b), SkyReels-Audio (SkyReels Team, 2025), Wan-S2V (Gao et al., 2025a), and InfiniteTalk (Yang et al., 2025) develop audio driven control. MultiTalk (Kong et al., 2026), AnyTalker (Zhong et al., 2025), and OmniHuman-1.5 (Jiang et al., 2025) address conversational or expressive generation. For reference control, OmniShow (Zhou et al., 2026) and HuMo (Chen et al., 2025a) combine visual and acoustic inputs, Phantom (Liu et al., 2025) supports subject images, and DreamID-Omni (Guo et al., 2026) unifies reference generation, editing, and animation. SpeakerVid-5M (Zhang et al., 2025b) and FunCineForge (Liu et al., 2026) provide conversation data and audiovisual annotation. Our model combines appearance references with driving audio or voice references through distinct roles.

Long video generation. LongCat-Avatar1.5 (Meituan LongCat Team et al., 2026), KlingAvatar 2.0 (Kling Team, 2025), LiveAvatar (Huang et al., 2025), StreamChar (Tian et al., 2026), Hallo-Live (Li et al., 2026), and LPM 1.0 (Zeng et al., 2026) extend human generation through streaming, causal, or windowed designs. Maintaining identity and temporal continuity across windows remains central to long video synthesis. We combine context adaptation with persistent identity references and recurrent latent context, keeping context fixed while generating each new segment. Visual or audiovisual history supports audio driven or voice-referenced joint continuation, respectively.

3 Method

We first formulate the generation tasks through a unified representation, then describe condition composition for long-video continuation, followed by two-stage progressive training. Figure 2 summarizes the architecture and the reuse of reference conditions across windows.

3.1 Unified Task Representation

Problem formulation and token groups. Given a text prompt p and optional visual and acoustic conditions, the model predicts video or joint audio–visual outputs. Driving audio specifies speech content and rhythm, voice references specify timbre, appearance references specify visual identity, and a first frame constrains the initial scene. Modality-specific VAEs encode these observations and training targets. After patchification, we organize the latent tokens into video targets X_v, visual conditions C_v, audio targets X_a, and acoustic conditions C_a:

$$\mathcal{X} = \{X_v, C_v, X_a, C_a\}, \qquad \tau = (\mathcal{G}_\tau, \mathcal{R}_\tau, \mathcal{Q}_\tau, \mathcal{B}_\tau). \tag{1}$$

The task specification τ determines the active groups $\mathcal{G}_\tau$, role IDs $\mathcal{R}_\tau$, temporal position types $\mathcal{Q}_\tau$, and clean-token masks $\mathcal{B}_\tau$. Missing groups are omitted. Let $\mathcal{T}_\tau$ denote the target modalities and $\mathcal{C}_\tau$ the provided conditions. The model represents

$$p_\theta(\{X_m\}_{m\in\mathcal{T}_\tau} \mid p, \mathcal{C}_\tau; \tau), \qquad \mathcal{T}_\tau \subseteq \{v, a\}. \tag{2}$$

Role embeddings and subject association. For token z_i^m in modality m, Proj_m maps its latent features to the transformer hidden dimension. We then add the role embedding $E_m(r_i)$ selected by its group's role ID r_i:

$$\widetilde{h}_i^m = \mathrm{Proj}_m(z_i^m) + \mathbf{1}[r_i \neq 0] E_m(r_i), \qquad m \in \{v, a\}. \tag{3}$$

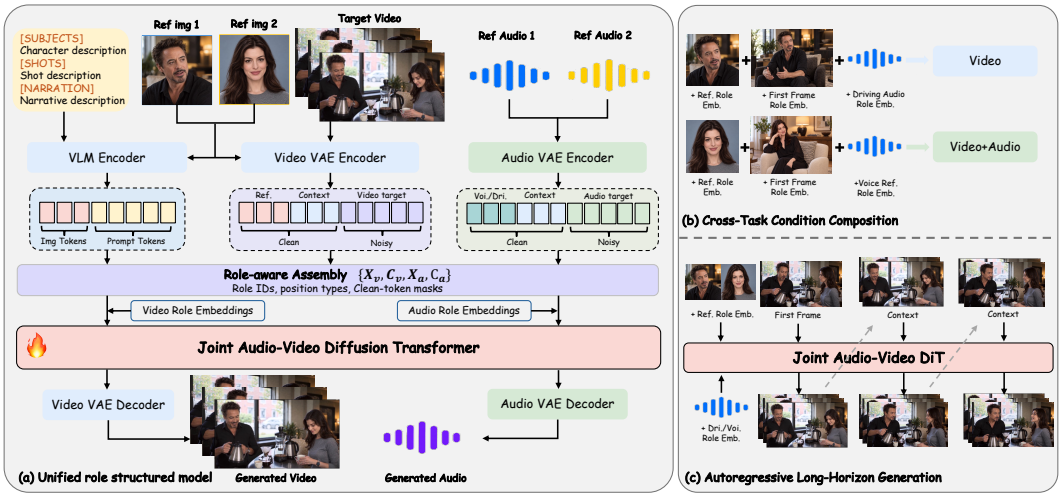


Figure 2: **Architecture and condition composition for long video continuation.** (a) A multimodal prompt encoder and modality-specific VAEs produce prompt, visual, and acoustic representations. Role structured assembly organizes target/condition groups, temporal positions, and clean-token masks for the shared audio–video diffusion transformer. (b) Cross-task condition composition supports two long video continuation modes: appearance references are combined with driving audio for video generation or with voice references for joint audio–video generation. (c) Autoregressive inference retains identity references and updates generated context across windows.

Table 1: Task configurations under the unified representation. T2AV and TI2AV correspond to base capabilities inherited from LTX-2.3. p denotes the text prompt. Driving audio is an observed target utterance. Voice-reference audio supplies identity evidence for newly generated speech. The semantic superscripts are independent of implementation-level role IDs.

Task	Predicted tokens	Conditioning tokens
Text-to-Audio-Video (T2AV)	X_v, X_a	p
Image-conditioned Text-to-Audio-Video (TI2AV)	X_v, X_a	p, first frame C_v^{first}
Image-conditioned Audio driven Video (TIA2V)	X_v	p, C_v^{first}, driving audio C_a^{drive}
Timbre-Referenced Text-to-Audio-Video (TRA2AV)	X_v, X_a	p, timbre reference C_a^{voice}
Image- and Timbre-Referenced Text-to-Audio-Video (TIRA2AV)	X_v, X_a	p, C_v^{first}, C_a^{voice}
Appearance–Voice Referenced Text-to-Audio-Video (TR2AV)	X_v, X_a	p, paired references $\{C_{v,k}^{\text{id}}, C_{a,k}^{\text{voice}}\}_{k=1}^{K}$

Video and audio use separate embedding tables. Driving audio and standalone voice references receive distinct roles. Paired appearance and voice references use corresponding subject indices, also used in the structured prompt to associate descriptions and dialogue with each subject. Appendix C describes the role conventions.

Temporal positions and token masks. Each group receives a position type $q \in \{-1, 0, 1\}$, indicating an independent reference, pre-target context, or target. Independent groups retain their own coordinates, while temporal groups are placed on the generation timeline. These coordinates encode temporal placement, and role embeddings specify semantic function. Independently encoded driving audio can still align with target video through physical-time coordinates. The clean-token mask B_i^m is 1 for fixed conditions and 0 for tokens to be generated. In both modalities, this mask specifies prediction status rather than attention visibility.

The unified representation accommodates six task configurations, as summarized in Table 1. T2AV and TI2AV correspond to the text and image conditioned generation capabilities inherited from the LTX-2.3 backbone, based on LTX-2 (HaCohen et al., 2026). T2AV predicts both video and audio from text, while TI2AV additionally fixes the first frame as a visual condition. TIA2V adds observed driving audio to the text and first frame, leaving only video as the prediction target. TRA2AV and TIRA2AV instead use audio as a voice reference, without and with a first frame, respectively, and generate both modalities. TR2AV uses text and a variable number of appearance–voice reference pairs for joint generation, with role indices binding each pair to a subject.

3.2 Cross-Task Composition for Long Video Continuation

Appearance reference composition with existing tasks. We augment the first-frame-conditioned TIA2V and TIRA2AV tasks with appearance reference groups from TR2AV, retaining their role embeddings and independent temporal positions. Let $\mathcal{C}_{\text{app}} = \{C^{\text{id}}_{v,k}\}_{k=1}^{K}$ denote these appearance references and $\mathcal{C}_{\text{voice}}$ the voice references in TIRA2AV. TIA2V generates video from a first frame and driving audio, while TIRA2AV generates both video and audio from a first frame, voice references, and text. The added appearance references persist across continuation windows to help preserve subject identity and appearance over long sequences.

Persistent references and rolling history. For generation window j, the conditions are

$$\mathcal{C}_j = \mathcal{C}_{\text{id}} \cup \mathcal{H}_{j-1} \cup \mathcal{U}_j. \tag{4}$$

The fixed set $\mathcal{C}_{\text{id}}$ contains appearance references and, for joint generation, voice references. The rolling context $\mathcal{H}_{j-1}$ contains video latents $H_{v,j-1}$ from the preceding segment and, for joint generation, audio latents $H_{a,j-1}$. The current inputs $\mathcal{U}_j$ consist of the text prompt p_j and, for audio driving, the waveform segment $C^{\text{drive}}_{a,j}$. After initialization with the first frame, the two continuation modes are formulated as

$$X_{v,j} \sim p_\theta\big(X_{v,j} \mid H_{v,j-1}, \mathcal{C}_{\text{app}}, C^{\text{drive}}_{a,j}, p_j\big)\,, \tag{5}$$

$$(X_{v,j}, X_{a,j}) \sim p_\theta(X_{v,j}, X_{a,j} \mid H_{v,j-1}, H_{a,j-1}, \mathcal{C}_{\text{app}}, \mathcal{C}_{\text{voice}}, p_j)\,. \tag{6}$$

The joint implementation carries forward both video and audio latent context, holds this context fixed while denoising each new segment, and removes repeated overlap tokens when assembling the long sequence. The resulting audio and video latent sequences are decoded with their respective decoders. Generation is autoregressive across chunks and diffusion-based within a chunk. This does not require token-by-token causal generation within each window. Persistent references provide stable identity cues alongside the evolving context to maintain identity consistency during long video continuation.

3.3 Training Strategy

Multi-task sampling and masked flow matching. Let $\pi(\tau)$ denote the training task mixture and $\mathcal{D}_\tau$ the corresponding records. A record supplies the target signals, text, reference assets, and role/position configuration. The data construction process provides speaker-indexed captions and cross-clip reference pairs. Its collection and annotation details are given in Appendix B. We use linear flow matching, interpolating each clean target token $x^m_{0,i}$ with Gaussian noise ϵ^m_i at flow time σ. The resulting input $\mathcal{X}_\sigma$ contains noisy targets and fixed conditions, and the target velocity is $u^m_i = \epsilon^m_i - x^m_{0,i}$. Condition tokens are held at their supplied values and receive timestep zero. Generated video and audio use the same sampled σ. The objective is

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{\tau\sim\pi,\, x\sim\mathcal{D}_\tau,\, \sigma,\, \epsilon}\left[\sum_{m\in\mathcal{T}_\tau} \mathcal{L}_m\right], \\ \mathcal{L}_m &= \frac{\sum_i (1 - B^m_i)\, \| f^m_\theta(\mathcal{X}_\sigma, p; \tau)_i - u^m_i \|_2^2}{d_m \sum_i (1 - B^m_i)}. \end{aligned} \tag{7}$$

Here $\mathcal{L}_m$ measures the squared error between the predicted velocity f^m_θ and target velocity u^m_i on tokens selected by $(1 - B^m_i)$. The denominator normalizes this error by the number of prediction tokens and latent feature dimension d_m. The implementation sums the normalized video and audio



Figure 3: **Qualitative comparison of audio driven generation.** Each column shows the first frame, driving audio, and an abbreviated prompt, followed by three frames from each method. The examples cover character animation and piano singing.

Table 2: Short-video single-person audio driven generation on VorchBench. Subj.-Cons. and BG-Cons. denote VBench subject and background consistency. Flicker and Motion denote its temporal-flickering and motion-smoothness scores.

	Identity		Sync		VBench					
Method	Face-Sim ↑	DINO-S ↑	Sync-C ↑	Sync-D ↓	Subj.-Cons. ↑	BG-Cons. ↑	AES ↑	IQA ↑	Flicker ↑	Motion ↑
InfiniteTalk (Yang et al., 2025)	<u>0.7881</u>	**0.9258**	5.1619	8.3840	**0.9754**	<u>0.9761</u>	<u>0.5581</u>	**0.6960**	<u>0.9896</u>	<u>0.9921</u>
LiveAvatar (Huang et al., 2025)	0.7513	<u>0.9235</u>	4.1553	9.1021	0.9691	0.9687	0.5473	0.6762	0.9851	0.9875
LongCat-Avatar1.5 (Meituan LongCat Team et al., 2026)	0.7455	0.8686	<u>5.1829</u>	**8.0563**	0.9509	0.9589	0.5405	0.6829	0.9854	0.9898
OmniAvatar (Gan et al., 2025)	0.6808	0.8843	4.9787	<u>8.3432</u>	0.9550	0.9657	0.5260	0.6676	0.9856	0.9911
StableAvatar (Tu et al., 2025)	0.6806	0.9037	2.4297	10.2793	0.9627	0.9618	0.5457	0.6605	0.9778	0.9856
Vorch-Human	**0.7887**	0.9206	**5.1885**	8.5502	<u>0.9737</u>	**0.9775**	**0.5663**	<u>0.6846</u>	**0.9907**	**0.9931**

losses. The expectation averages this sum over sampled tasks, training examples, flow times, and noise. Modalities with no prediction tokens are excluded, so audio driven generation uses only the video loss, while joint generation uses both modality losses.

Two-stage training. The first stage mixes the generation tasks to establish shared capabilities under the role structured interface. The second stage performs context adaptation: historical video and audio latents are supplied as known conditions, and the model learns to generate the subsequent segments. In joint generation, both video and audio context tokens have $B_i^m = 1$, while new video and audio tokens have $B_i^m = 0$. In audio driven generation, video context is retained and the driving waveform remains observed. Both stages use the masked flow-matching objective in Eq. 7. Context adaptation trains temporal continuation, while persistent appearance and voice references provide identity evidence during recurrent inference.

4 Experiments

We evaluate audio driven video and appearance–voice referenced joint generation, then examine condition composition and long-horizon behavior. Quantitative benchmarks characterize generation quality. The user study is provided in Appendix F.

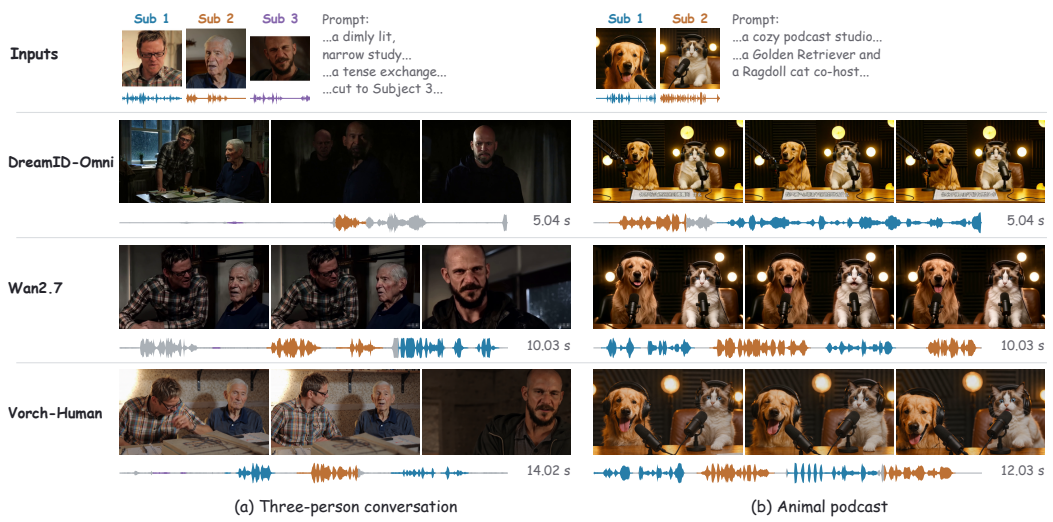


Figure 4: **Qualitative comparison of R2AV.** Inputs comprise paired image and voice references with an abbreviated prompt.

Table 3: Image–audio reference generation on IDBench-Omni. S/M denotes single-person and multi-person results.

	Video			Audio		Audio–Visual Consistency	
Method	AES ↑	ViCLIP ↑	ID-Sim (S/M) ↑	WER ↓	T-Sim (S/M) ↑	Sync-C ↑	Sync-D ↓
DreamID-Omni (Guo et al., 2026)	0.5931	**13.9792**	0.5816/0.5201	0.1331	0.5287/0.4017	2.9212	10.2381
Wan2.7 (Wan et al., 2025)	0.5843	13.7514	**0.7718/0.7261**	0.2117	**0.7431/0.6229**	2.5081	9.0367
Vorch-Human	**0.5952**	13.5815	0.6833/0.5814	**0.0234**	0.6395/0.5424	**5.4664**	**8.3790**

4.1 Experimental Setup

Audio driven evaluation. We construct VorchBench to evaluate audio driven video generation over short and long durations. It contains 50 short-video cases and 10 long-video cases, each providing a first frame, driving audio, and a text prompt. We evaluate visual quality, identity preservation, and audio–visual synchronization on both splits, and additionally examine temporal stability during five-minute continuation. Metric definitions and evaluation details are provided in Appendix E.

R2AV evaluation. We evaluate appearance–voice referenced joint generation on IDBench-Omni (Guo et al., 2026). Each case provides subject appearance images, voice references, and a text script. We compare with DreamID-Omni (Guo et al., 2026) and Wan2.7 (Wan et al., 2025), assessing video quality, identity preservation, speech accuracy, voice similarity, and audio–visual synchronization. Metric definitions and score aggregation are provided in Appendix E.

4.2 Implementation Details

We initialize Vorch-Human from LTX-2.3, based on LTX-2 (HaCohen et al., 2026), and fine-tune all diffusion transformer parameters with the two-stage strategy in Sec. 3.3, keeping the multimodal text encoder frozen. Training clips are at most 16 seconds, with per-frame pixel budgets of 409,920 for training and 921,600 for evaluation. Inference uses the distilled LTX-2.3 noise schedule with eight denoising steps. Further details are provided in Appendix C.

4.3 Comparison

Audio driven generation. We evaluate short audio driven generation using six VBench metrics and four additional metrics, with results reported in Table 2. Vorch-Human leads on six of the ten

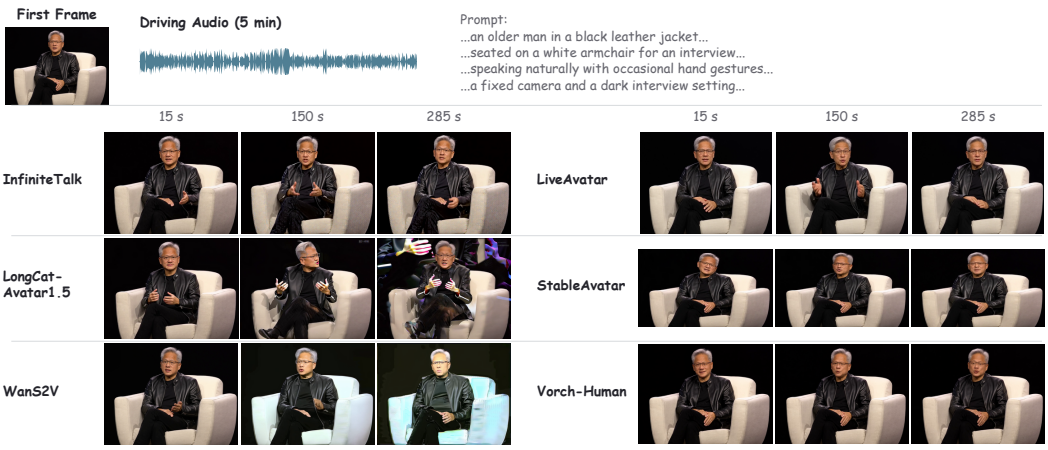


Figure 5: **Qualitative comparison of five-minute audio driven generation.** The top row shows the initial frame, driving audio, and an abbreviated prompt. Frames at 15, 150, and 285 seconds illustrate early, middle, and late generation.

Table 4: Generation capabilities and five-minute single-person audio driven results on VorchBench. Audio clone denotes voice-referenced joint audio–visual generation.

	Capabilities		Five-minute audio driven generation				
Method	Audio driven	Audio clone	Face-Sim ↑	DINO-S ↑	Sync-C ↑	AV-Sync ↓ (ms)	IQA ↑
InfiniteTalk (Yang et al., 2025)	✓	×	0.6502	0.9323	**7.8269**	**8.537**	<u>0.9691</u>
LiveAvatar (Huang et al., 2025)	✓	×	0.6605	0.9718	6.2207	<u>12.515</u>	0.9557
LongCat-Avatar1.5 (Meituan LongCat Team et al., 2026)	✓	×	0.5038	0.7563	<u>6.4904</u>	24.509	0.8942
StableAvatar (Tu et al., 2025)	✓	×	0.5840	<u>0.9772</u>	2.9876	51.650	0.9658
WanS2V (Gao et al., 2025a)	✓	×	<u>0.6965</u>	0.8718	5.5297	37.353	0.8489
Vorch-Human	✓	✓	**0.7814**	**0.9834**	6.4119	26.915	**0.9736**

metrics, with strong facial identity preservation, synchronization confidence, and visual and temporal quality. Qualitative comparisons of character animation and audio driven human generation are shown in Figure 3.

Appearance–voice referenced generation. On IDBench-Omni, Vorch-Human achieves the best AES, WER, Sync-C, and Sync-D, demonstrating competitive visual quality, speech accuracy, and audio–visual synchronization, as reported in Table 3. Wan2.7 achieves higher identity and timbre similarity, while DreamID-Omni (Guo et al., 2026) leads on video–text alignment. The examples in Figure 4 compare reference appearance preservation and speaking behavior across the three methods in a three-person conversation and an animal podcast. For single-clip generation, Vorch-Human supports a maximum duration of 15 seconds, compared with 5 seconds for DreamID-Omni (Guo et al., 2026) and 10 seconds for Wan2.7.

Long-video generation. For five-minute audio driven generation, Vorch-Human achieves the best Face-Sim, DINO-S, and IQA, while InfiniteTalk (Yang et al., 2025) obtains the strongest synchronization scores, as shown in Table 4. In Figure 5, Vorch-Human retains facial appearance, clothing, and scene composition at early, middle, and late time points, whereas some competing results exhibit color drift or structural artifacts. The temporal curves in Figure 10 (Appendix E) further show stable identity and appearance similarity throughout five minutes. By retaining both video and audio context, Vorch-Human can also extend the same interface to joint audio–video continuation.

4.4 Ablation Study

Unified task representation. We evaluate three variants modified during training on IDBench-Omni (Guo et al., 2026), keeping reference content, prompts, and clean-token masks unchanged, with results reported in Table 5. (1) **w/o RE:** Disabling role embeddings weakens identity preservation and speech accuracy. (2) **SI merged:** Assigning all subjects the same reference index causes

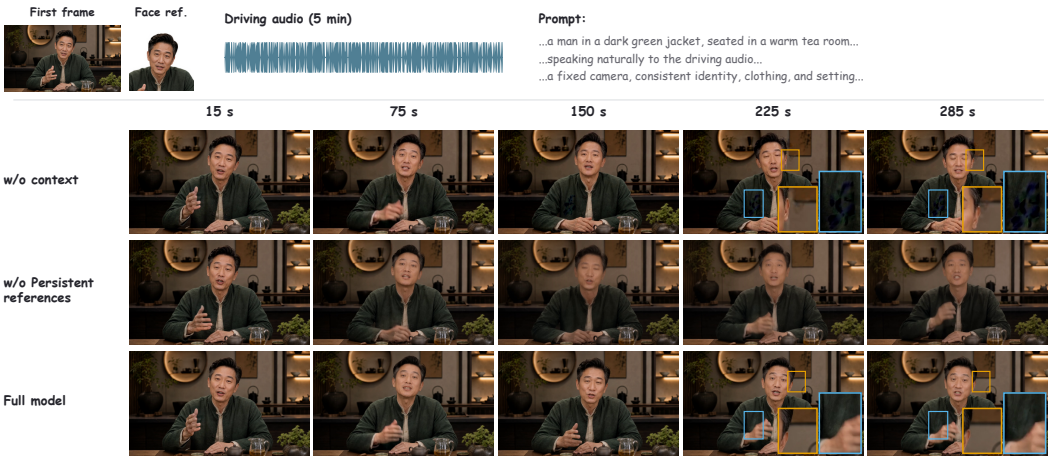


Figure 6: Qualitative ablation of long-video continuation.

Table 5: Ablation study on unified task representation.

Method	AES ↑	ViCLIP ↑	ID-Sim (S/M) ↑	WER ↓	T-Sim (S/M) ↑	Sync-C ↑
w/o RE	0.5921	**13.7129**	0.5762 / 0.3145	0.1064	**0.7052 / 0.5651**	4.6974
SI merged	0.5932	13.2305	0.6453 / 0.2951	0.1104	0.6289 / 0.5400	5.1822
w/o RP	**0.6004**	13.2373	0.6644 / 0.5085	0.1430	0.5961 / 0.5024	1.0086
Full	0.5952	13.5815	**0.6833 / 0.5814**	**0.0234**	0.6395 / 0.5424	**5.4664**

Table 6: Ablation study on long video continuation.

Method	Face-Sim ↑	DINO-S ↑	IQA ↑	Sync-C ↑
w/o CA	0.6716	0.9416	0.8919	6.2014
w/o PAR	0.5223	0.8227	0.6632	5.8024
Full	**0.7814**	**0.9834**	**0.9736**	**6.4119**

the largest decline in multi-person identity similarity, indicating the importance of distinct subject indices for appearance control. (3) **w/o RP:** Replacing independent reference positioning with pre-target placement yields the weakest synchronization and speech accuracy. Together, these results support the roles of semantic annotations, subject association, and reference positioning in conditional generation.

Context adaptation and persistent references. We evaluate two variants for five-minute audio driven generation on VorchBench, with results reported in Table 6. (1) **w/o CA:** Using the first-stage checkpoint without context-adaptation training, while retaining the same context inputs and references, reduces identity preservation and visual quality. (2) **w/o PAR:** Removing persistent appearance references after the initial window while retaining rolling context causes larger declines in identity, appearance consistency, and visual quality. Both variants weaken synchronization. As illustrated in Figure 6, the model without context-adaptation training exhibits local distortions around the head and clothing, along with abrupt lighting changes between chunks. Removing persistent references causes increasing facial distortion and loss of detail, while the full model better preserves appearance and continuity. Additional ablation comparisons are provided in Appendix G.

5 Conclusion

We presented Vorch-Human, a unified framework for human-centric audio–visual generation and long-horizon continuation. Its unified target–condition representation uses role embeddings, temporal position types, and clean-token masks to support heterogeneous generation tasks within a shared audio–video diffusion transformer. Cross-task condition composition combines persistent appearance references with driving audio or voice references and rolling context, supporting long video continuation in both audio driven and jointly generated audio–video modes through the same interface. A two-stage training strategy first develops shared capabilities through multi-task mixed training and then adapts the model to fixed video and audio context for continuation. At inference, rolling latent context supports temporal continuity, while persistent references provide identity evidence across windows. Experiments demonstrate competitive generation quality, identity preservation, and audio–visual synchronization, while supporting stable long video continuation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and language polishing, retrieve and discover relevant literature, and draft sections of the paper. These tools supported translation, improvements to clarity and readability, and the organization of manuscript drafts. We take responsibility for reviewing AI-assisted content, verifying retrieved references against original sources, and ensuring that descriptions and interpretations remain consistent with our methods and experimental evidence. The authors retain responsibility for the accuracy, originality, and final content of this work, including text and claims produced with the aid of generative AI.

ETHICS STATEMENT

Vorch-Human supports human-centric audio–visual generation with appearance and voice references, creating potential risks of impersonation, misleading synthetic media, and non-consensual identity replication. Its use should respect individuals' consent, privacy, and rights over their likeness and voice. We advocate clear disclosure of generated media and appropriate safeguards, including content provenance and watermarking, when deploying such systems. Developers should assess demographic biases in visual and speech generation and guard against harmful or discriminatory use. Human evaluations should protect participant privacy and ensure voluntary participation.

REFERENCES

ByteDance ALIVE Team. ALIVE: Animate your world with lifelike audio-video generation. *arXiv preprint arXiv:2602.08682*, 2026.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025a.

Qin Chen, Yingjie Chen, Shilun Lin, Cai Xing, Binxin Yang, Long Zhou, Qixin Yan, Wenjing Wang, Dingming Liu, Hao Liu, Chen Li, and Jing Lyu. Identity as Presence: Towards Appearance and Voice Personalized Joint Audio-Video Generation. *arXiv preprint arXiv:2603.17889*, 2026. URL https://arxiv.org/abs/2603.17889.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *International Conference on Learning Representations*, 2024.

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

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

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

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.

Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. OmniAvatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-S2V: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025a.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025b.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, et al. Live Avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5325–5329, 2024.

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.

Kling Team. Klingavatar 2.0 technical report. *Technical report*, 2025.

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. *Advances in Neural Information Processing Systems*, 38:70990–71013, 2026.

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.

Jiaxuan Liu, Yang Xiang, Han Zhao, Xiangang Li, and Zhenhua Ling. Funcineforge: A unified dataset toolkit and model for zero-shot movie dubbing in diverse cinematic scenes. *arXiv preprint arXiv:2601.14777*, 2026.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025.

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

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

Meituan LongCat Team, Xunliang Cai, Meng Cheng, Feng Gao, Zhe Kong, Jiamu Li, Le Li, Weiheng Li, Hongyu Liu, Shuai Tan, Xiaoming Wei, Tianyu Yang, and Yong Zhang. LongCat-Video-Avatar 1.5 technical report. *arXiv preprint arXiv:2605.26486*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

SkyReels Team. Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers. *arXiv preprint arXiv:2506.00830*, 2025.

Tencent Hunyuan Team. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Linrui Tian, Qi Wang, and Bang Zhang. Streamchar: Long-horizon streaming character audio-video generation with decoupled orchestration. *arXiv preprint arXiv:2605.25659*, 2026.

Shuyuan Tu, Yueming Pan, Yinming Huang, Xintong Han, Zhen Xing, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. StableAvatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pp. 1–20, 2024.

Yaohui Wang, Xin Ma, Xinyuan Chen, Cunjian Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *International Journal of Computer Vision*, 133(3):1277–1289, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-Align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54015–54029, 2024.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Ailing Zeng, Casper Yang, Chauncey Ge, Eddie Zhang, Garvey Xu, Gavin Lin, Gilbert Gu, et al. LPM 1.0: Video-based character performance model. *arXiv preprint arXiv:2604.07823*, 2026.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. UniAVGen: Unified Audio and Video Generation with Asymmetric Cross-Modal Interactions. *arXiv preprint arXiv:2511.03334*, 2025a. URL `https://arxiv.org/abs/2511.03334`.

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. *arXiv preprint arXiv:2507.09862*, 2025b.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, Ying Qin, Huan Li, Shuiyang Mao, Wei Liu, and Wenhan Luo. AnyTalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

Donghao Zhou, Guisheng Liu, Hao Yang, Jiatong Li, Jingyu Lin, Xiaohu Huang, Yichen Liu, Xin Gao, Cunjian Chen, Shilei Wen, Chi-Wing Fu, and Pheng-Ann Heng. Omnishow: Unifying multimodal conditions for human-object interaction video generation. *arXiv preprint arXiv:2604.11804*, 2026.

A Appendix

The appendix contains the following sections:

- Data Construction and Annotation (Section B).
- Implementation Details (Section C).
- Prompt Templates and Examples (Section D).
- Evaluation Details (Section E).
- User Study (Section F).
- Additional Ablation Results (Section G).
- Additional Qualitative Results (Section H).

B Data Construction and Annotation

B.1 Training-data Construction Overview

Our data pipeline comprises clip annotation and cross-clip reference mining, as illustrated in Figure 7. The clip-annotation stage operates within each clip and recovers the transcript, active speaker, appearance, voice characteristics, and stable subject identity. These records directly support audio driven human animation and first-image plus reference-audio training. The cross-clip mining stage searches other clips from the same source video for the same subject, producing cross-clip appearance references for image–audio reference generation. Before these two stages, inexpensive human, pose, and visual-quality filters remove samples on which the expensive audio–visual and language-model stages are unlikely to be reliable.

Collection and pre-filtering. We combine curated and open-source human videos. Each input is transcoded to a constant frame rate, its audio is extracted at 16 kHz, and shot boundaries are detected. Pose tracks and face/body visibility remove invalid or weakly visible subjects. Technical and perceptual filters then reject blur, poor motion, low-quality face and hand regions, poor aesthetics, and subtitle occlusion. This high-recall-to-high-precision ordering follows the practical principle of running costly speaker analysis only after a clip passes basic checks.

B.2 Structured Caption Format

Rather than training a separate captioner, we combine specialist measurements with Gemini- and Qwen-family multimodal models. Specialist models provide the verbatim transcript, timestamps, tracks, and audio–visual evidence. The MLLMs correct semantic identity assignments and organize those measurements without replacing them. The resulting caption has four blocks:

$$\mathcal{C} = \{\texttt{SUBJECTS}, \texttt{BACKGROUND_AUDIO}, \texttt{SHOTS}, \texttt{NARRATION}\}. \tag{8}$$

For each `Subject_k`, `SUBJECTS` stores a stable subject ID, appearance description, and timbre description. `SHOTS` describes composition and visible actions. `NARRATION` interleaves actions with verbatim speech spans and preserves their speaker IDs. This representation is both a training prompt and a machine-readable bridge to the reference assets.

B.3 Clip-level Speaker Separation and Annotation

The clip-annotation stage begins with audio decomposition and ASR. Vocal separation isolates speech from accompaniment and background sound, while ASR yields sentence and word timestamps. In parallel, face detection, temporal face tracking, aligned face recognition, and within-clip identity clustering provide persistent visual tracks. Active Speaker Detection (ASD) associates speech activity with face tracks, and SyncNet (Chung & Zisserman, 2016) supplies an additional lip–audio synchrony signal. Audio speaker embeddings are clustered independently. Audio, visual, and joint clusters are then reconciled to assign each sentence to a visible subject.

Signal-based diarization can still fail under off-screen speech, rapid cuts, and fragmented tracks. We therefore provide the MLLM with the transcript, sampled frames, face-track evidence, and cluster

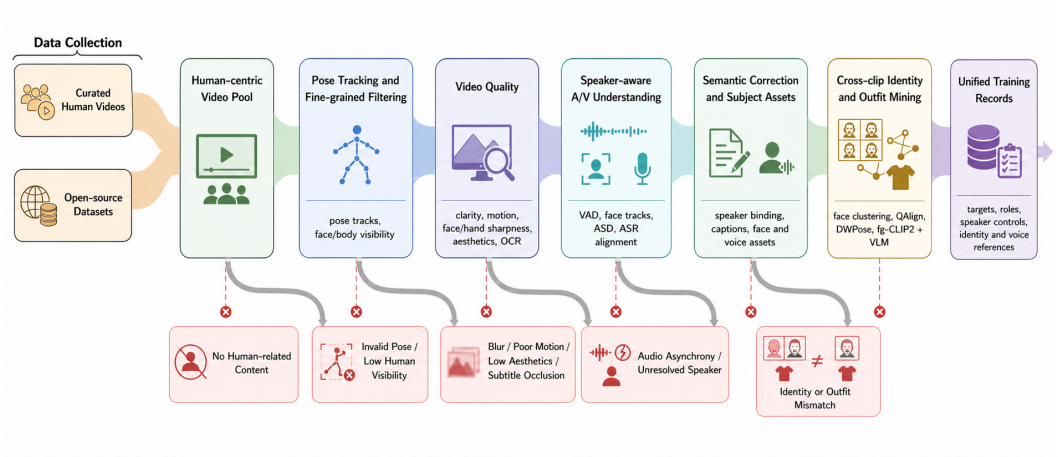


Figure 7: Human-centric data processing pipeline. The clip-annotation stage resolves speakers and produces a structured clip caption and per-subject appearance and voice assets. The cross-clip mining stage links the same identity across clips from one source video and validates reference-image quality and outfit consistency. The final records support audio driven, audio-clone, and paired image–audio reference training.

assignments to correct speaker–subject associations and map recurring visual identities to stable `Subject_k` labels. For every resolved subject, the stage outputs (i) speech text and its time interval, (ii) an appearance description, (iii) a timbre description, and (iv) the subject ID. It also extracts face references and clean speech segments. Image IQA and foreground processing are independent of voice extraction, so a subject can retain a usable voice asset even when all candidate images are rejected.

These aligned clip-level records are sufficient for audio driven animation, including single- and multi-person cases, and for first-image plus reference-audio training. They also provide the identities that the cross-clip mining stage uses as anchors.

B.4 Cross-clip Reference Mining

The cross-clip mining stage groups clips by their original source video and samples each clip at one frame per second. Human and face detectors locate candidate subjects. Landmark alignment produces face embeddings and head-pose estimates. We apply agglomerative clustering across all clips in the source group and merge nearby centroids, thereby turning clip-local IDs into a source-level identity index. This constraint is important: reference candidates come from other clips of the same source video rather than unrelated web images.

Before reference selection, we match each face to a unique containing body box, filter image quality with QAlign (Wu et al., 2024), and check torso visibility with DWPose. For retained same-identity candidates, appearance comparison checks clothing and hairstyle consistency. Clear pairs are accepted or rejected by embedding similarity, while ambiguous pairs are adjudicated by a vision-language model. Connected components of valid pairwise links form the final identity–outfit groups. We thus select same-subject reference images from other clips, introducing pose and background variation while preserving identity and outfit consistency.

B.5 Unified Training Records

The final record contains target video/audio, the structured caption, task name, condition image/audio lists, role IDs, position types, and optional speaker time intervals and boxes. Reference images are allowed to differ from the target in pose, framing, and background so the model must learn identity rather than copy pixels. The same schema represents driving-audio examples, audio-clone examples, and one- or multi-subject paired image–audio references.

B.6 Data Splits and Reference Independence

We select cross-clip references of the same subject with different poses and backgrounds. Our evaluation records document source-video and subject-identity overlap with the training data, as well as speech-content overlap between reference and target utterances, to distinguish identity transfer from reconstruction and potential content leakage.

Using this pipeline, we construct one million training records for appearance–voice referenced audio–visual generation.

C Implementation Details

C.1 Role and Position Conventions

We assign role 0 to ordinary targets and first-frame conditions. Their masks and position types distinguish them. Driving audio and standalone voice references use distinct role indices, denoted by r_{drive} and r_{voice}, respectively. In the paired appearance–voice task, subject images and subject voices both use indices $1, \ldots, K$ in their respective modality-specific embedding tables. Thus the numerical voice-role convention depends on the task layout. Position types $-1, 0, 1$ indicate independent references, pre-target context, and targets, respectively. The clean-token mask is 1 for known conditions and 0 for prediction targets.

C.2 Training and Checkpoint Specification

We initialize Vorch-Human from an LTX-2.3 checkpoint, using the dual-stream diffusion architecture of LTX-2 (HaCohen et al., 2026). We train on 64 NVIDIA H200 GPUs with the AdamW optimizer, a learning rate of 1×10^{-5}, a global batch size of 128, and two gradient accumulation steps. Training uses BF16 precision and an exponential moving average (EMA) decay of 0.999. We drop semantic conditioning with probability 0.1 for classifier-free guidance (CFG).

The training implementation holds condition tokens at timestep zero, uses a shared sampled flow time for generated audio and video, and sums normalized target-only losses. In the first stage, we sample training data for T2AV, TI2AV, TIA2V, TRA2AV, TIRA2AV, and TR2AV at a ratio of $3 : 3 : 3 : 2 : 2 : 2$, respectively. In the second stage, we use 200,000 training samples for context adaptation to long-video continuation, with video and audio contexts each spanning 0.8 seconds.

C.3 Semantic Encoding and Inference Settings

Training clips are grouped into resolution buckets under the specified per-frame pixel budget. At evaluation, spatial dimensions follow the aspect ratio of the input condition or the benchmark specification and are rounded to multiples compatible with the VAE. For audio driven generation, the driving audio determines the output duration, and the frame count is rounded to the nearest value supported by the VAE's temporal compression. All task families share the same inference procedure, with task configurations specifying the target and condition token groups, role identifiers, temporal position types, and VLM prompt templates.

C.4 Long-Video Continuation Inference

Long-video continuation uses 16-second generation windows with a 15.2-second stride, resulting in a 0.8-second overlap between adjacent windows. The first window uses the first frame to establish the initial scene. Subsequent windows omit the temporal first-frame input and retain persistent appearance references and, for joint audio–video generation, voice references.

Each subsequent window receives latent context from the preceding segment, which remains fixed while the new segment is denoised. Audio driven continuation carries forward video context and uses the corresponding segment of the supplied driving audio, while joint continuation carries forward both video and audio context. After generation, duplicate overlap tokens are removed from subsequent chunks, and the retained latents are concatenated, trimmed to the requested duration, and decoded. Video decoding uses temporal and spatial tiling.

D PROMPT TEMPLATES AND EXAMPLES

D.1 TASK-SPECIFIC VLM PROMPT TEMPLATES

The tasks in Table 1 use three VLM prompt layouts, illustrated in Figure 8. Panel (b) gives a normalized, shared presentation for TIA2V and TIRA2AV, including their long-form variants: the paired alternatives distinguish driving audio from timbre-reference audio and video prediction from joint audio–video prediction. The image-condition and user-instruction sections retain the same structure. The placeholder {instruction} is replaced with the sample caption, or the current segment's instruction during continuation. Long-form generation additionally receives recurrent latent context through the model's condition groups.

In panel (b), {image_condition_content} denotes the ordered image descriptions and image tokens, identifying the first-frame condition and any additional appearance references. For SUBJECT_REFERENCE, {ref_content} expands to entries such as The Image-1 <start_of_image>. The image counts are replaced with the number of included images, up to ten. Panel (a) presents the basic prompt layout for T2AV, TI2AV, and TRA2AV; panel (b) presents the shared layout for TIA2V, TIRA2AV, and their long-form extensions; panel (c) presents the subject-reference layout for TR2AV. The default template contains no image placeholders, so its first-frame inputs enter through the visual VAE condition branch rather than the VLM. Audio conditions enter through the acoustic branch.

Structured TR2AV input prompt. A two-subject example illustrates the structured input prompt in Figure 9. The four fields specify subject attributes, background audio, shot composition, and narration. Each subject is associated with an appearance image and a voice reference through matching indices, and the same subject identifiers link these references to actions and speaking turns.

(a) DEFAULT

T2AV, TI2AV, TRA2AV

```
[Task]
This is a video generation task. Follow user instruction to generate video.

[User Instruction]
{instruction}

[Output]
Output the target video based on user instructions clearly and faithfully.
```

(b) IMAGE_AUDIO_CONDITIONED

TIA2V / TIRA2AV: short-clip and long-form generation

```
[Task]
Use the first-frame condition image, {driving audio / reference audio}, and
optional subject reference images for {audio-driven / timbre-referenced}
generation.

[Conditions]
There are {num_images} condition images.
{image_condition_content}

[User Instruction]
{instruction}

[Output]
Output the target {video / audio-video} based on user instructions, the
first-frame condition, {driving audio / reference audio}, and subject visual
references clearly and faithfully.
```

(c) SUBJECT_REFERENCE

TR2AV

```
[Task]
This is a subject reference task. Use reference images of subject and follow
user instruction to generate video with consistent subject appearance.

[Conditions]
There are {num_refs} reference images of the subject.
{ref_content}

[User Instruction]
{instruction}

[Output]
Output the target video based on user instructions and visual conditions
clearly and faithfully.
```

Figure 8: VLM prompt layouts. Panels (a) and (c) reproduce implementation text. Panel (b) gives a normalized shared layout for TIA2V and TIRA2AV and their long-form extensions.

TR2AV: input prompt and references



Subject_1

Image-1 + Audio-1



Subject_2

Image-2 + Audio-2



[SUBJECTS]

Subject_1: [Visual] A rugged late-19th-century outlaw with swept-back medium-brown hair, a short rough beard and mustache, weathered fair skin, strong facial features, and subtle scars. He wears a faded blue work shirt, dark neckerchief, worn brown leather coat, dark trousers, and a weathered gun belt. His facial identity, hairstyle, clothing, and defining visual characteristics should closely follow Image-1. (Reference images: Image-1) [Audio] A low, weathered masculine voice with a restrained, deliberate delivery and subtle dry humor. (Reference audio: Audio-1)

Subject_2: [Visual] An athletic female archaeologist-adventurer with light olive skin, expressive brown eyes, subtle freckles, and dark-brown hair tied in a practical loose ponytail. She wears a dark teal expedition top, rugged brown trousers, fingerless gloves, and practical exploration gear. Her facial identity, hairstyle, clothing, and defining visual characteristics should closely follow Image-2. (Reference images: Image-2) [Audio] A clear, confident feminine voice with an educated British tone, measured pacing, and an adventurous but thoughtful delivery. (Reference audio: Audio-2)

[BACKGROUND_AUDIO]

A nearly empty frontier saloon during a heavy nighttime rainstorm. Rain strikes the windows, distant thunder rolls softly, and wind moves around the old wooden building. A fireplace crackles in the background while the floorboards occasionally creak. A metal lantern produces a quiet flame sound on the table. The environmental audio remains subtle and never overpowers the dialogue.

[SHOTS]

Shot_1: A static medium two-shot of Subject_1 and Subject_2 sitting across from each other at a small wooden table inside a dim frontier saloon. Subject_1 sits on the left side of the frame. He leans back slightly in his chair, with one forearm resting near the edge of the table. His posture appears guarded but relaxed. Subject_2 sits on the right side of the frame. She leans slightly toward an old treasure map that has been unfolded between them. One gloved hand rests beside an antique brass compass. A metal lantern stands near the center of the table. Its warm light illuminates their faces while cool blue storm light enters through a rain-covered window behind them. Both subjects are framed from approximately the waist up. Their faces, natural gestures, and eye contact remain clearly visible.

Shot_2: A separate medium close-up featuring only Subject_2. Subject_2 is framed from the chest upward. The camera is positioned slightly to the side of Subject_1's implied position, allowing Subject_2 to look naturally toward him without looking into the camera. The treasure map remains visible along the lower edge of the frame. The rain-covered window and warm lantern light remain consistent with Shot_1. Subject_1 must not appear anywhere in Shot_2, including reflections, mirrors, windows, photographs, shadows, or background silhouettes.

[NARRATION]

In Shot_1, Subject_2 slowly unfolds the weathered treasure map across the table. She studies a faded symbol near the center and places the brass compass beside it. Subject_1 watches her movements in silence. He glances toward the map, then raises his eyes toward Subject_2 with a skeptical expression. After a brief pause, Subject_1 says, <W>Every treasure map I have ever seen ends with somebody bleeding.</W> Subject_2 stops moving the compass and looks toward Subject_1. She listens without interrupting, her expression calm and attentive. Her lips remain closed while he is speaking. Subject_1 finishes his sentence and leans back slightly. The lantern flame flickers as distant thunder sounds outside. Cut cleanly to Shot_2. Subject_2 looks down at the map for a moment, considering his warning. She then raises her eyes toward Subject_1's implied off-camera position. A restrained, confident smile appears as she replies, <W>Then perhaps we should learn from the map instead of repeating its ending.</W> After speaking, Subject_2 gently turns the compass.

Figure 9: TR2AV input prompt with paired image and voice references, subject descriptions, background audio, shots, and narration.

E Evaluation Details

E.1 Audio Driven Evaluation on VorchBench

VorchBench comprises 60 out-of-distribution evaluation cases: 50 short-video cases and 10 long-video cases. Each case includes a first frame, driving audio, and a text prompt. First-frame inputs include real photographs, synthetic images, and animation images. Driving audio includes both real recordings and voice-cloned speech, covering diverse speakers, rhythms, and acoustic characteristics. The short-video protocol uses 15-second clips, and the long-video protocol uses five-minute driving audio. Tables 2 and 4 report the short- and long-video comparisons, respectively.

Face-Sim measures whether the generated face preserves the identity in the reference image using ArcFace (Deng et al., 2019) recognition embeddings, whereas DINO-S measures broader reference-image similarity, including appearance and visual structure, using DINOv2 (Oquab et al., 2024) features. Sync-C and Sync-D evaluate how well lip motion aligns with the driving speech using SyncNet (Chung & Zisserman, 2016): Sync-C is the synchronization confidence, while Sync-D is the audio–visual embedding distance. For the short-video single-person benchmark, we further report six VBench (Huang et al., 2024) dimensions. Subject consistency (Subj.-Cons.) measures whether the depicted person remains visually consistent across frames, and background consistency (BG-Cons.) measures the stability of the surrounding scene. Aesthetic quality (AES) captures the overall visual appeal, whereas imaging quality (IQA) reflects low-level fidelity such as sharpness, clarity, and visible artifacts. The temporal-flickering score (Flicker) measures resistance to frame-to-frame flicker, and motion smoothness (Motion) measures whether motion in the video remains continuous and smooth without abrupt temporal jumps. All metrics are higher-is-better except Sync-D. For the long-video benchmark, perceptual video IQA is measured using Q-Align (Wu et al., 2024), while AV-Sync is the temporal alignment error in milliseconds measured by SyncFormer (Iashin et al., 2024).

We evaluate identity and appearance stability over five-minute generation, with temporal trajectories shown in Figure 10. Vorch-Human maintains high Face-Sim with only a modest decline and nearly constant DINO-S, indicating sustained identity and appearance preservation. LongCat-Avatar1.5 and Wan-S2V show larger declines in both metrics as generation progresses. StableAvatar and LiveAvatar maintain stable DINO-S but lower Face-Sim, highlighting the distinction between overall visual consistency and facial identity preservation.

E.2 R2AV Evaluation on IDBench-Omni

For image–audio reference generation on IDBench-Omni (Guo et al., 2026), AES measures video aesthetics, ViCLIP measures video–text semantic alignment, and ID-Sim measures subject identity preservation. WER evaluates the intelligibility and correctness of generated speech, while T-Sim measures speaker-timbre similarity between the generated speech and the audio reference. Sync-C and Sync-D are the standard SyncNet (Chung & Zisserman, 2016) confidence and distance. The notation S/M denotes single-person and multi-person evaluation, respectively. AES, ViCLIP, ID-Sim, T-Sim, and Sync-C are higher-is-better. WER and Sync-D are lower-is-better.

Score aggregation. All three methods are evaluated against the original IDBench-Omni references. WER is averaged per video, while ID-Sim is averaged over successfully matched references.

F User Study

Study protocol. We complement automatic metrics with a Good–Same–Bad (GSB) pairwise study, for which we invited 100 evaluators. For each prompt, annotators view two anonymized videos in randomized left–right order. Here G indicates a preference for Vorch-Human, B indicates a preference for the compared method, and S indicates no perceptible difference. The comparison with LongCat-Avatar1.5 (Meituan LongCat Team et al., 2026) includes 42 single-person and 12 multi-person cases, with ten annotators per comparison, and evaluates visual quality, motion/pose/expression naturalness, and lip accuracy. The comparison with Wan2.7 evaluates reference-conditioned and audio driven generation using task-specific criteria. We aggregate judgments by dimension and present the preference distributions separately for the two comparisons.

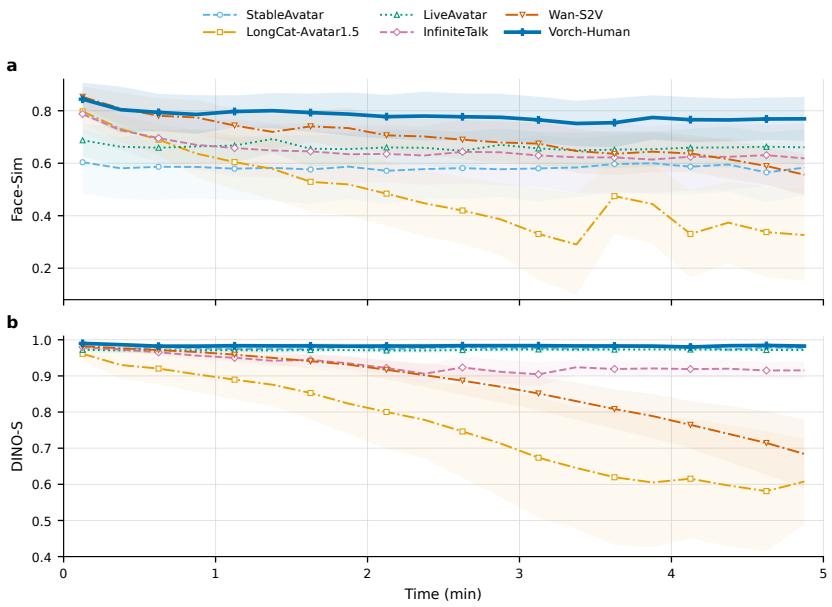
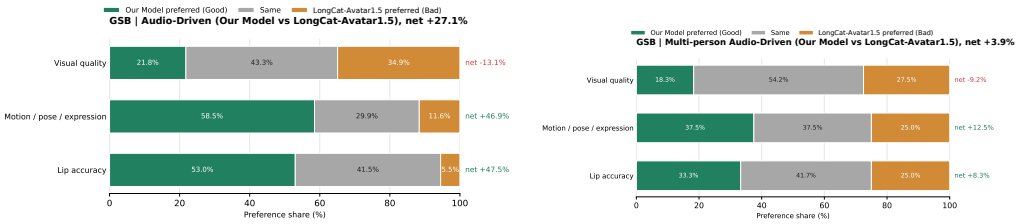
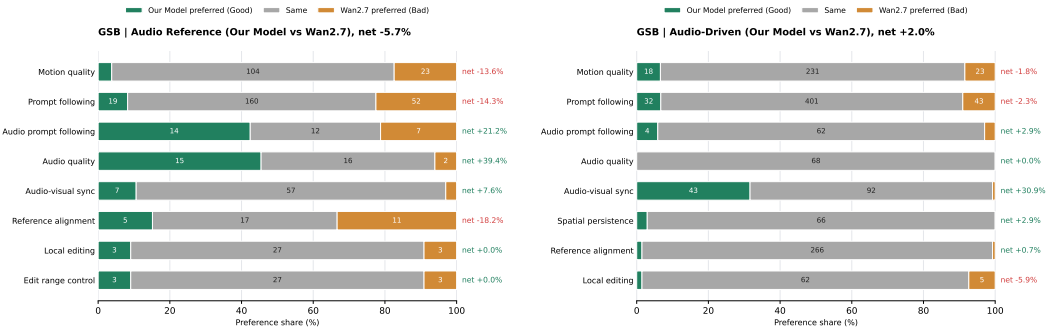


Figure 10: Temporal stability over five-minute generation. Each metric occupies one row: (a) Face-Sim and (b) DINO-S. Curves show window means and shaded regions show 95% confidence intervals.



(a) Single-person audio driven (b) Multi-person audio driven

Figure 11: GSB user study comparing Vorch-Human with LongCat-Avatar1.5 (Meituan LongCat Team et al., 2026). Bars show preference percentages, and the values on the right indicate net preference (Good minus Bad).



(a) Audio reference (b) Audio driven

Figure 12: GSB user study comparing Vorch-Human with Wan2.7. Bars show normalized preference distributions; labels indicate judgment counts and net scores.

Comparison with LongCat-Avatar1.5. Vorch-Human is preferred for motion naturalness and lip accuracy in both single- and multi-person audio driven generation, with a more pronounced

advantage in single-person cases, as shown in Figure 11. LongCat-Avatar1.5 is preferred for visual quality. These results highlight the strengths of Vorch-Human in natural human motion and lip synchronization, while indicating room for improvement in visual quality.

Comparison with Wan2.7. Audio driven generation is comparable overall, while annotators favor Vorch-Human for audio–visual synchronization, as shown in Figure 12. In reference-conditioned generation, Vorch-Human is preferred for audio quality, audio prompt following, and synchronization, whereas Wan2.7 is preferred for motion quality, visual prompt following, and reference alignment. The results indicate complementary strengths across acoustic and visual criteria.

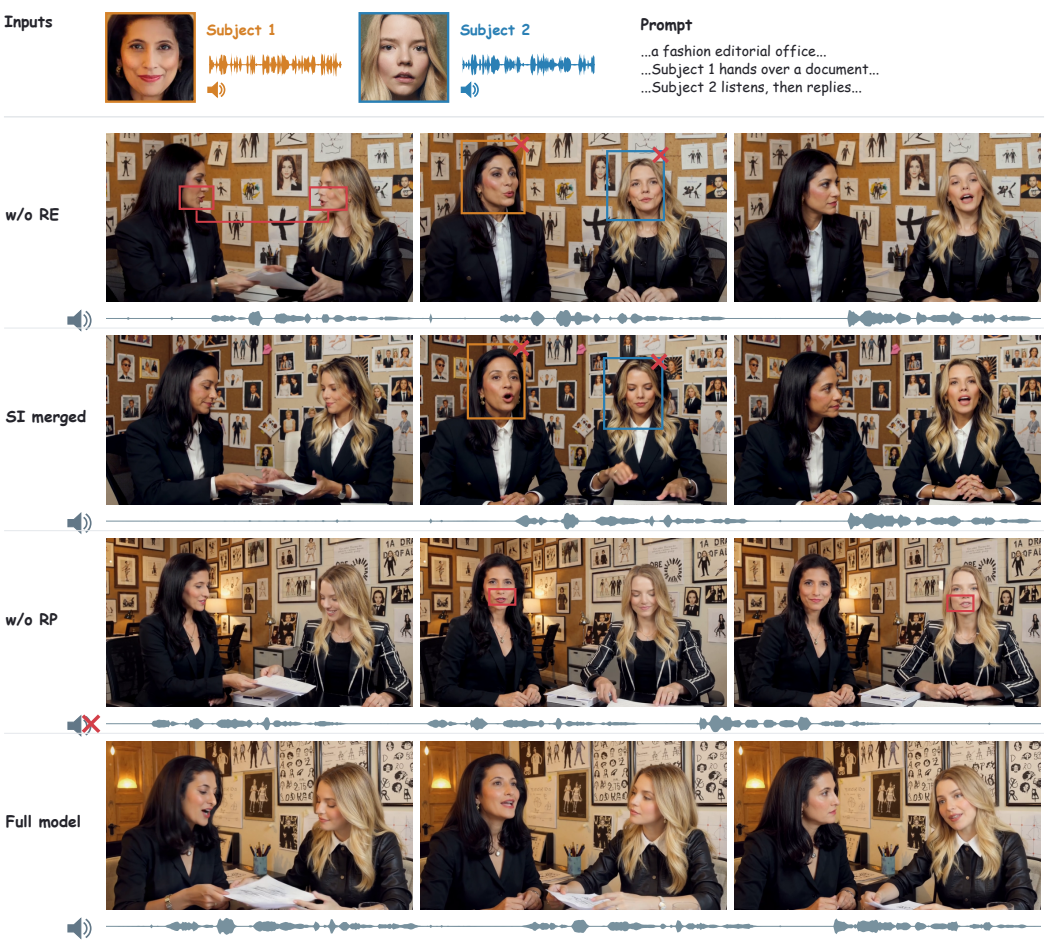


Figure 13: Qualitative comparison of representation ablations on TR2AV. Colored face boxes correspond to the input subjects, with crosses marking appearance deviations. Connected red mouth boxes highlight simultaneous mouth activity, and the cross beside the audio marks observed audio–visual desynchronization.

Analysis. In this example, w/o RE, trained without role embeddings, confuses the reference appearances and produces simultaneous speaking behavior, making the intended speaker difficult to distinguish. SI merged, trained with a shared subject index, also mixes appearance characteristics, indicating that retaining a common role embedding does not preserve the distinction between subjects. With w/o RP, using pre-target placement instead of independent reference positioning during training disrupts audio–visual synchronization. The full model better preserves the two reference identities and separates their speaking turns. These observations illustrate the complementary roles of role embeddings, subject indices, and reference positioning.

G Additional Ablation Results

G.1 Unified Task Representation

We compare the full model with three variants trained with modified task representations on the same TR2AV example to examine appearance preservation and speaking behavior, as illustrated in Figure 13.

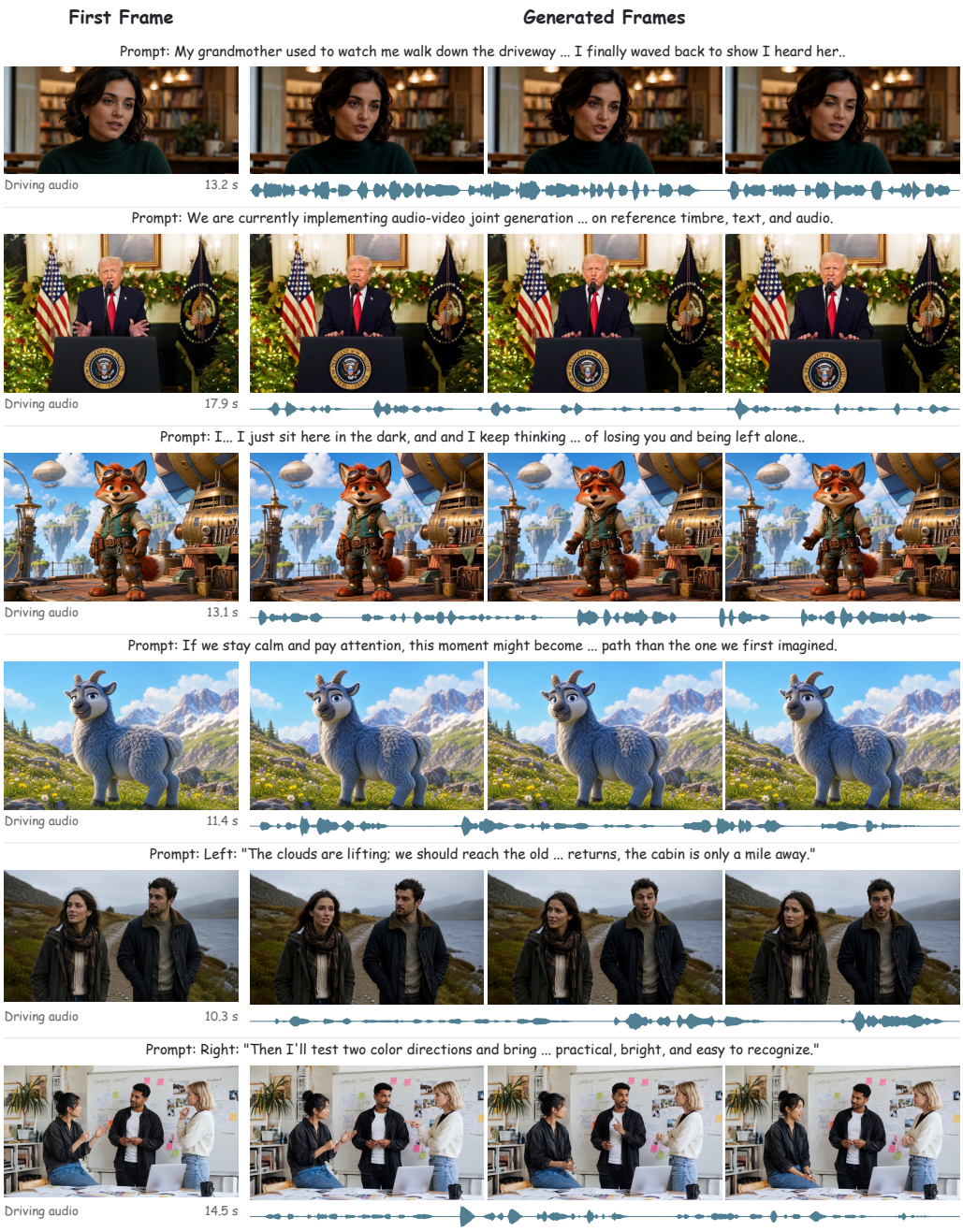


Figure 14: More qualitative results on TIA2V.

H Additional Qualitative Results

We provide additional TIA2V, TIRA2AV, and TR2AV examples in Figures 14, 15, and 16, respectively. Long-video examples cover singing, talking, and voice-referenced joint generation, as illustrated in Figure 17.

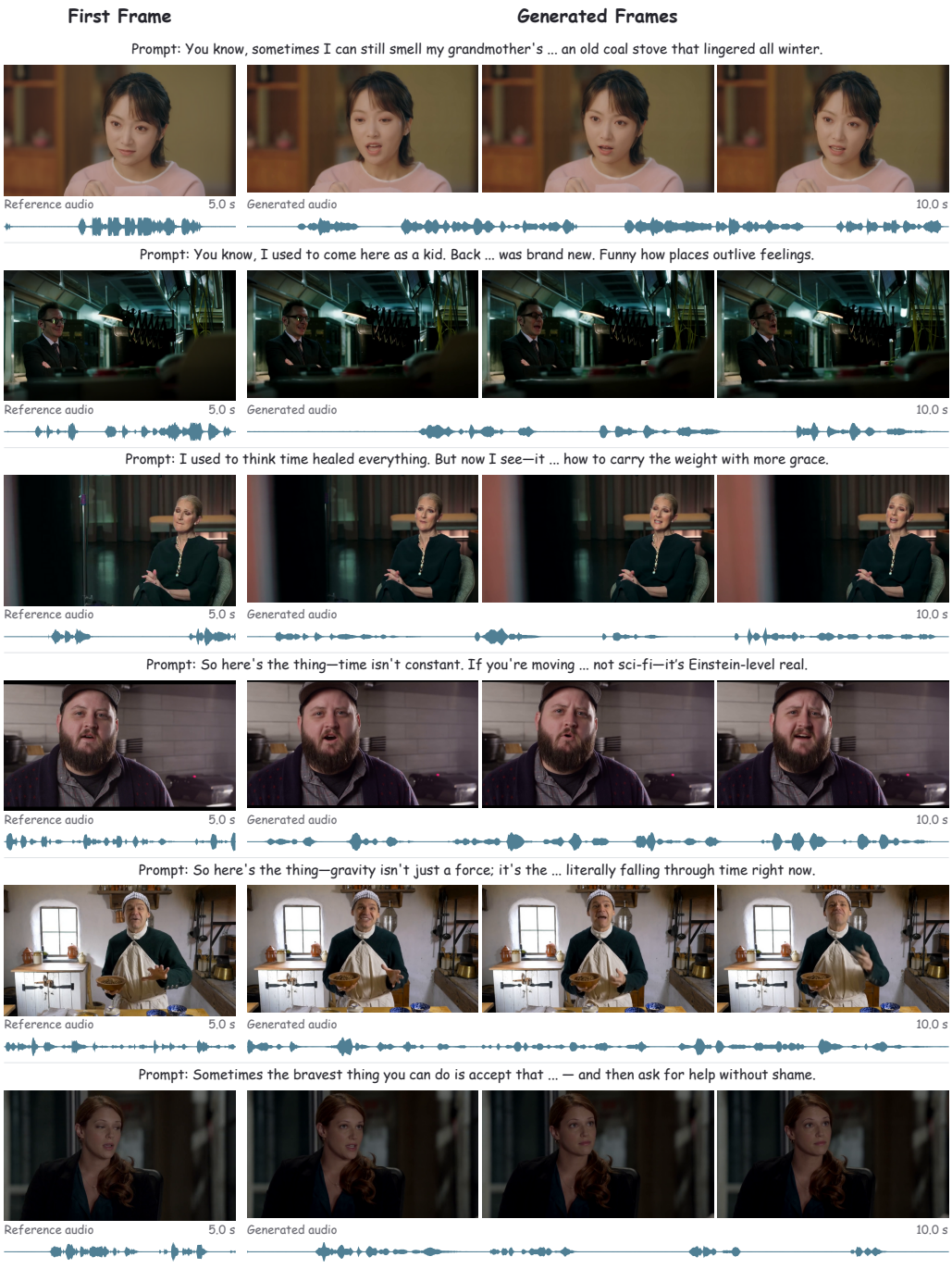


Figure 15: More qualitative results on TIRA2AV.

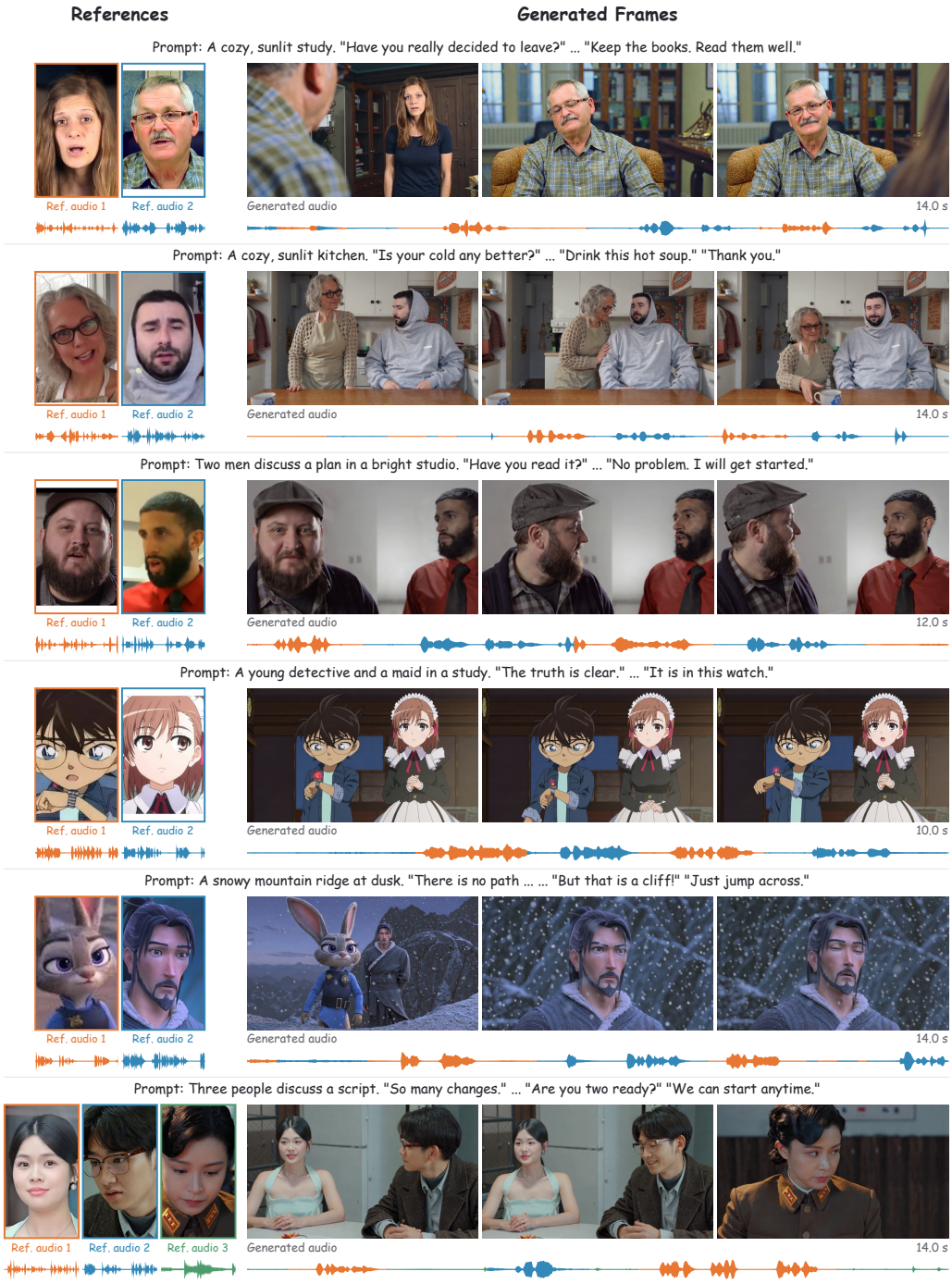


Figure 16: More qualitative results on TR2AV.

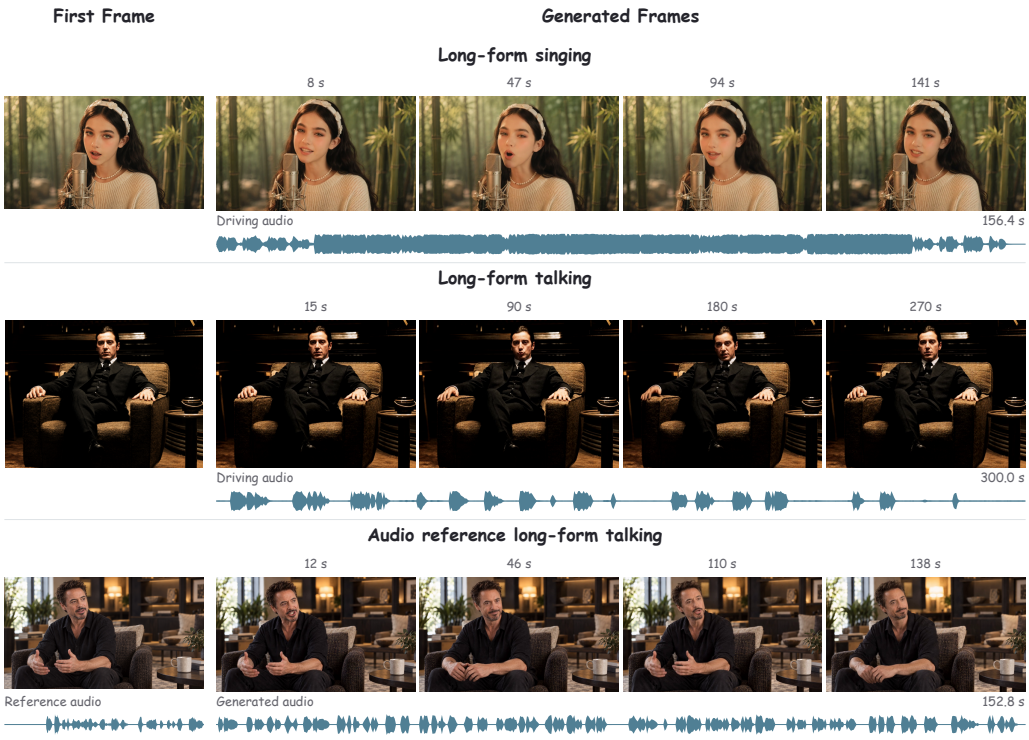


Figure 17: More qualitative results on long-video generation.