BEYOND STATIC REFERENCES: AUDIO-VISUAL REFERENCE CONDITIONING FOR HUMAN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-conditioned human video generation has advanced rapidly, yet existing datasets still predominantly represent human subjects using one or a few static reference images. Such references provide only sparse observations of subject appearance while discarding temporal visual cues and voice characteristics naturally available in videos. Large-scale training data with richer audio-visual references for human video generation therefore remain limited. We introduce **VidVoice-Ref**, a large-scale dataset for human video generation with audio-visual subject references. Instead of reducing a subject to isolated frames, VidVoice-Ref preserves complete reference videos together with synchronized audio, capturing multi-view appearance, pose and expression variations, motion patterns, and voice timbre. We construct identity-consistent cross-video samples by identifying target subjects, retrieving temporally separated reference clips, and selecting visually diverse references across viewpoints, poses, expressions, motions, and scenes. Audio references are further refined through speaker diarization, speech validity filtering, intra-clip speaker consistency, voice quality filtering, and cross-video voice matching. VidVoice-Ref contains approximately 1M reference–target samples and over 1K hours of video, covering both single- and multi-subject scenarios. VidVoice-Ref provides a scalable data foundation for studying richer human subject representations beyond static image references, particularly for identity preservation, subject binding, and audio-visual reference-conditioned human video generation.

1 INTRODUCTION

Recent advances in large-scale video generation models have substantially improved the visual fidelity, temporal coherence, and controllability of synthesized videos, driving growing interest in subject-conditioned video generation. A prominent direction is **reference-conditioned human video generation**, which aims to synthesize novel videos containing one or multiple referenced human subjects while preserving their identities across new scenes, viewpoints, poses, and interactions. Recent methods, including Phantom Liu et al. (2025), BindWeave Li et al. (2026c), MAGREF Deng et al. (2025), and Kaleido Zhang et al. (2025b), have substantially advanced single- and multi-subject generation through improved reference conditioning and subject binding. Meanwhile, OpenS2V-Nexus Yuan et al. (2026) introduces a large-scale dataset and benchmark for systematically evaluating subject consistency, naturalness, and text alignment.

Despite this progress, current reference-conditioned human video generation remains predominantly image-centric. Existing models typically represent a human subject using a single reference image or a small collection of static images. Large-scale resources such as Phantom-Data Chen et al. (2026d) and OpenS2V-5M Yuan et al. (2026) substantially expand cross-pair and subject-consistent training data, yet subject conditions are still largely represented through static visual observations. Static references capture only limited views of a subject and cannot fully represent variations in pose, expression, viewpoint, dynamic appearance, and motion over time. Consequently, current models must infer these missing temporal characteristics from incomplete evidence, making subject preservation challenging under large pose changes, novel viewpoints, complex actions, and diverse scenes.

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.

Videos, in contrast, provide a substantially richer source of subject information. A reference video naturally captures a subject across multiple viewpoints, poses, expressions, and temporal states, while preserving dynamic visual characteristics and motion patterns that cannot be fully represented by isolated images. Moreover, when accompanied by audio, videos provide another important subject-specific cue: voice timbre. Visual appearance and vocal characteristics therefore provide complementary information for representing a subject. This motivates extending reference conditioning beyond static visual inputs toward complete audio-visual subject references.

Recent studies have begun to explore video references for identity-preserving human generation. Slot-ID Lai et al. (2026) and Avatar V Liang et al. (2026) show that reference videos can provide richer temporal identity cues than static images. However, these methods primarily focus on single-person identity or avatar generation. Open-domain reference-conditioned human video generation requires handling substantially broader variations in subject appearance, scenes, actions, and interactions, potentially involving multiple subjects within the same video. Large-scale training resources that preserve complete reference videos together with subject-associated audio for such settings remain limited.

To address this gap, we introduce **VidVoice-Ref**, a large-scale dataset for audio-visual reference-conditioned human video generation. Instead of reducing a subject to one or several static observations, VidVoice-Ref preserves complete reference videos together with their synchronized audio, providing richer subject information across appearance, pose, expression, motion, and voice. The dataset is designed to support both single- and multi-subject generation, enabling richer subject conditioning beyond static visual references.

Our data construction follows a target-centric pipeline for building identity-consistent yet visually diverse reference–target pairs. We first identify the subjects appearing in each target video and retrieve reference clips containing the same identities from distinct video segments. The retrieved references are further selected to reduce visual redundancy while preserving variations in viewpoint, pose, expression, motion, and scene context, followed by video-level verification to ensure reliable subject observations. We then refine the associated audio by retaining references with valid speech, consistent speaker characteristics, and matched voice identity across reference and target videos. Finally, structured audio-visual annotations are generated to describe target content while maintaining explicit subject-reference correspondence. VidVoice-Ref consists of approximately 1M reference–target samples spanning over 1K hours of video.

Our contributions can be summarized as follows:

- We introduce VidVoice-Ref, a large-scale dataset for reference-conditioned human video generation with complete reference videos and synchronized audio, enabling richer subject conditioning beyond static image references and supporting both single- and multi-subject generation.
- We develop a target-centric curation pipeline for constructing identity-consistent and visually diverse reference–target pairs, with dedicated visual reference selection and audio refinement for reliable subject and voice consistency.
- VidVoice-Ref contains approximately **1M** reference–target samples spanning over **1K** hours of video, covering both single- and multi-subject settings, together with rich structured audio-visual annotations for subject appearance, actions, interactions, scenes, motion, and speech.

2 Related Work

2.1 Subject-to-Video Generation and Its Training Data

S2V methods differ in how they inject the reference into the video model (Liu et al., 2025; Deng et al., 2025; Chen et al., 2025; Li et al., 2026c; Zhang et al., 2025b; Chen et al., 2026c). All of them encode the subject as one or a few still images. A still image fixes one viewpoint and one moment, so identity has to be extrapolated to the poses and expressions the target video asks for (Song et al., 2026; Zhang et al., 2025a). Much recent work adds viewpoints, assembled from training frames or synthesized (Meng et al., 2026; Wang et al., 2026a; Song et al., 2026; Zhang et al., 2025a). The reference is also usually cropped from the target clip, so the model can copy its pose and

lighting (Deng et al., 2025; Chen et al., 2026d). Mv^{2}ID (Hu et al., 2026) notes that cross-paired data reduces this copy-paste but is expensive to collect, and that adding still views without it makes the artifact worse. Phantom-Data (Chen et al., 2026d), Vera (Xu et al., 2026), and MuSS (Zhang et al., 2026a) retrieve references from a context different from the target. OpenS2V-5M (Yuan et al., 2026) does so only for its 0.35M Nexus subset. In all four datasets the reference is an image rather than a video. The temporal information in the source video is discarded when the corpus is built, and the audio is dropped or left unprocessed (Yuan et al., 2026).

2.2 Video and Voice as Subject References

Recent works have explored richer subject references beyond static images along two directions. One line conditions generation on reference videos (Liang et al., 2026; Guo et al., 2026; Lai et al., 2026; Baptiste et al., 2026), but primarily focuses on single-person or avatar generation and does not provide large-scale training corpora. Another line jointly customizes visual identity and voice characteristics (Li et al., 2026a; Chen et al., 2026b; Dahan et al., 2026; Liu et al., 2026), typically using static images together with separately provided voice clips. This differs from audio-driven animation, where speech mainly controls lip motion and gestures while identity is specified visually (Lin et al., 2025; Gao et al., 2025; Kong et al., 2026; Chen et al., 2026a). Existing datasets in these settings remain limited in scale, availability, or reference formulation, with large-scale resources such as OmniCustom-1M and MAVINSet not publicly released. In contrast, VidVoice-Ref preserves complete reference videos together with their synchronized audio, directly associating visual identity with voice characteristics, and extends this formulation to open-domain single- and multi-subject reference-conditioned human video generation.

2.3 Audio-Visual Human Corpora

Existing audio-visual human corpora generally preserve speech but are not organized around same-subject reference and target pairs. OpenHumanVid (Li et al., 2025), SpeakerVid-5M (Zhang et al., 2026b), TalkVerse (Wang et al., 2026b), and AVSpeech (Ephrat et al., 2018) provide large-scale human videos with speech, but lack cross-video subject identity annotations suitable for reference-based generation. VoxCeleb2 (Chung et al., 2018) provides speaker identities, but is restricted to cropped face tracks, while ASID-1M (Li et al., 2026b) targets video understanding and LPM 1.0 (Zeng et al., 2026) does not release its curated corpus.

The scarcity of high-quality paired audio-visual data has also been noted in prior human-generation work (Chen et al., 2026a; Lin et al., 2025; Cui et al., 2025). Meanwhile, datasets such as Phantom-Data, Vera, and MuSS construct reference and target pairs without preserving paired audio (Chen et al., 2026d; Xu et al., 2026; Zhang et al., 2026a). In contrast, our setting requires reference and target videos to be consistent in both visual identity and voice characteristics.

3 Dataset Curation

The construction of VidVoice-Ref aims to obtain identity-consistent audio-visual reference–target pairs from large-scale real-world videos. Existing datasets for reference-conditioned human video generation predominantly represent human subjects using one or several isolated images. In contrast, we preserve complete reference clips together with their synchronized audio, providing richer observations of subject appearance, pose variation, motion, and voice characteristics. As illustrated in Fig. 1, our data curation pipeline consists of four major stages:

3.1 Source Data Collection

We construct the raw video pool from multiple public video sources, including MF^2 Zaranis et al. (2025), FineVideo Farré et al. (2024), the TRECVID Deep Video Understanding (DVU) movie collection Loc et al. (2022), and AVA Gu et al. (2018). These sources provide complementary video content, ranging from long-form movies and television-style narratives to diverse web videos and human-centered activity clips. Together, they cover a broad range of subjects, scenes, viewpoints, body compositions, actions, and human interactions, while providing repeated observations of the same individuals across different temporal and visual contexts. We retain videos with clearly visible

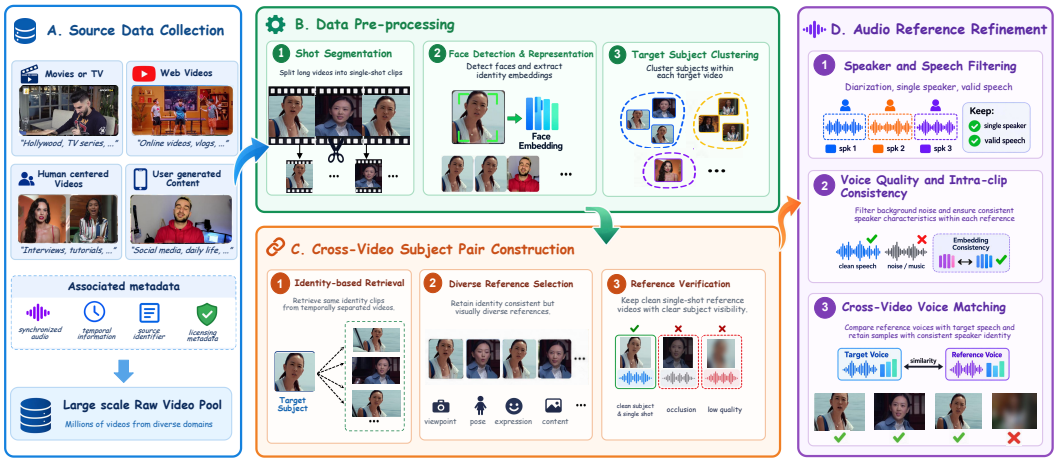


Figure 1: The VidVoice-Ref curation pipeline. The process consists of: (1) large-scale source data collection from diverse human-centered videos; (2) audio-visual pre-processing with shot segmentation, face representation, and target subject clustering; (3) cross-video subject pair construction through identity-based retrieval, diverse reference selection, and reference verification; and (4) audio reference refinement with speaker filtering, voice consistency checking, and cross-video voice matching.

human subjects that reappear across distinct video segments, making them suitable for constructing cross-video reference–target pairs. Whenever available, we preserve the original synchronized audio together with temporal information, source identifiers, and licensing metadata for subsequent audio-visual processing.

3.2 Audio-Visual Pre-processing

Shot segmentation. Long-form videos commonly contain frequent scene transitions, which introduce abrupt changes in subjects, backgrounds, and camera configurations. We therefore first divide each source video into individual shots using AdaptiveDetector. Each shot corresponds to a continuous video segment without internal scene cuts, while its temporal boundaries and synchronized audio are preserved. We use these single-shot clips as the basic units for subsequent subject analysis and reference retrieval. We additionally require the retained clips to exhibit consistent audio-visual content, such that speech is associated with visible on-screen subjects. Clips containing off-screen narration or voice-over speech without a corresponding visible speaker are discarded.

Face detection and representation. To identify human subjects appearing throughout the videos, we uniformly sample frames from the resulting clips and employ InsightFace Deng et al. (2019) for face detection and identity feature extraction. Each detected face is represented using an L_2-normalized identity embedding. We apply basic validity constraints to remove unreliable detections, such as low-confidence or excessively small faces. These constraints ensure that the extracted embeddings provide reliable identity representations for subsequent clustering and retrieval. More detailed filtering thresholds are provided in the supplementary material.

Target subject clustering. A target video may contain one or multiple human subjects. We therefore cluster detected facial embeddings within each target video according to cosine similarity, while leveraging temporal continuity between neighboring observations. Each resulting cluster represents a distinct target subject and is summarized by an identity prototype computed from its facial embeddings. The resulting subject prototypes are subsequently used for identity-based reference retrieval.

3.3 Cross-Video Subject Pair Construction

A central objective of VidVoice-Ref is to construct reference–target pairs in which the subject identity is preserved while the surrounding visual context is allowed to vary. Given a target video con-

taining N subjects,

$$\mathcal{S} = \{S_1, S_2, \ldots, S_N\}, \tag{1}$$

we independently retrieve reference clips for each subject from temporally separated video segments.

Identity-based reference retrieval. For each target subject S_i, we compute the cosine similarity between its identity prototype and facial embeddings extracted from candidate reference clips. Candidate clips with sufficiently high identity similarity are retained:

$$\mathcal{R}_i = \left\{ V_j^{\text{ref}} \mid \text{sim}\left(e(S_i), e(V_j^{\text{ref}})\right) > \tau_{\text{id}} \right\}, \tag{2}$$

where $e(\cdot)$ denotes the facial identity representation and τ_{id} is the identity similarity threshold.

Reference and target clips are temporally separated, such that they preserve the same subject identity while naturally allowing variations in viewpoint, pose, expression, motion, camera configuration, and scene context. This cross-video formulation provides identity-consistent reference–target pairs with diverse visual contexts.

Diverse reference selection. To reduce redundancy among retrieved references, we jointly consider identity similarity and visual diversity. For candidates from the same reference video, only the frame with the highest identity similarity is retained. The remaining candidates are ranked by identity similarity and sequentially filtered using DINO features Caron et al. (2021), removing candidates that are visually too similar to previously selected references. This yields identity-consistent references with greater diversity in viewpoint, pose, expression, and scene context.

Reference video verification. We further verify the retrieved references at the video level. Each candidate is restricted to a single-shot clip and is sampled at regular temporal intervals for face detection. If any valid sampled frame contains multiple faces, the candidate is discarded. We also remove clips for which no valid single-face observation can be detected across the sampled frames. The remaining references are deduplicated at the video level and retained as the final visual reference candidates for each target subject.

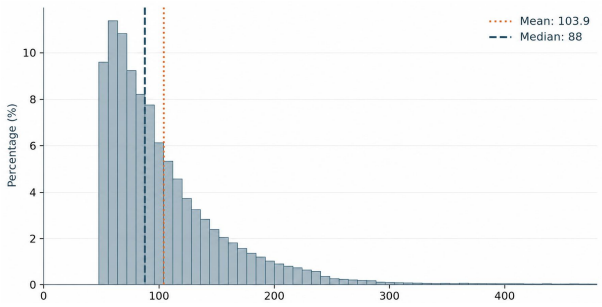
3.4 Audio Reference Refinement

For the visually verified reference videos, we further examine their associated audio to retain references with valid speech, consistent speaker characteristics, and matched voice identity across reference and target videos.

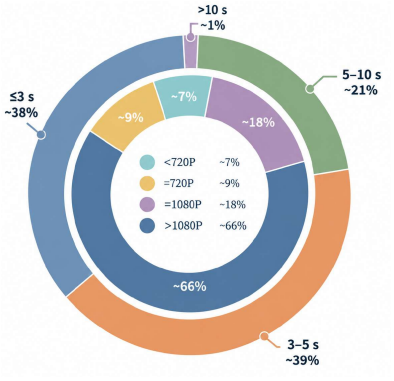
Speaker and speech filtering. We first verify the presence of an audio track and apply Pyannote Bredin (2023) for speaker diarization. The diarization results provide speaker identities together with their corresponding speech intervals. Overlapping intervals belonging to the same speaker are merged to compute the total effective speech duration. Only references containing exactly one detected speaker with more than one second of valid speech are retained for subsequent processing.

Voice quality and intra-clip consistency. The retained audio is resampled to 16 kHz and separated into vocal and non-vocal components using Demucs Rouard et al. (2023). Silero VAD is then applied to the separated vocals to identify active speech regions. The detected speech regions are concatenated and divided into non-overlapping one-second windows. We require at least two valid windows and extract a speaker embedding from each window using Resemblyzer. Pairwise cosine similarities are computed among these embeddings, and references whose minimum similarity falls below a predefined threshold are discarded. We further compute the RMS energy of the vocal and non-vocal components within active speech regions. References are retained only when the vocal energy is not lower than the background component.

Cross-video voice consistency. We further compare the voice characteristics of each reference clip with the speech segments in its corresponding target video. For the reference, we extract a speaker embedding from the separated vocal track. For the target video, Demucs and Silero VAD



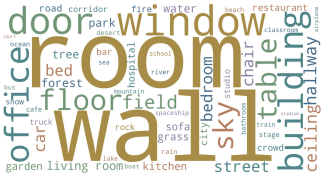
(a) Video Frames Distribution



(b) Duration and Resolution



(c) Appearance Word Cloud



(d) Scene Word Cloud



(e) Reference Word Cloud

Figure 2: Statistics of VidVoice-Ref from multiple perspectives, including video frame counts, duration and resolution distributions, and word clouds of subject appearance, scene context, and reference descriptions.

are similarly applied to obtain valid speech segments, from which speaker embeddings are extracted independently. Given a reference clip V_i^{ref} and a target video V^{tar}, we compute

$$s_{\text{voice}} = \max_{a_j \in \mathcal{A}^{\text{tar}}} \text{sim}\left(e_a(V_i^{\text{ref}}), e_a(a_j)\right), \tag{3}$$

where $e_a(\cdot)$ denotes the speaker representation and $\mathcal{A}^{\text{tar}}$ denotes the valid speech segments detected in the target video. We retain the reference if the maximum similarity exceeds τ_{voice} and the corresponding target speech segment has sufficient duration.

3.5 Data Annotation

Semantic annotation. We employ Qwen3-Omni Xu et al. (2025) to generate structured audio-visual annotations for both reference and target videos. Given a video clip together with its synchronized audio, the model describes the subjects, actions and interactions, scene context, camera behavior, motion dynamics, speech activity, and other salient audio-visual events. For target videos, we further convert the structured annotations into generation prompts. Referenced subjects are associated with explicit identity placeholders, e.g., `<ID_A>` and `<ID_B>`, to maintain consistent correspondence with their audio-visual references. The resulting prompts preserve target-side descriptions of subject appearance, actions and interactions, scene context, camera motion, visual style, and temporal dynamics. For samples containing speech, speech content and relevant audio context are additionally preserved in the structured annotations. For reference videos, we retain more detailed descriptions of subject appearance, viewpoint, pose, motion, and audio characteristics for dataset analysis and reference characterization. More details of the annotation schema and representative caption examples are provided in the supplementary material.

4 Dataset Statistics

VidVoice-Ref contains approximately 1 million reference–target samples and over 1,000 hours of video content. The dataset mainly consists of short video clips with an average duration of around

4 seconds, which aligns well with the common training setting of current video generation models. Each reference preserves both the complete video clip and its corresponding audio, providing synchronized visual and vocal information for subject conditioning.

Key characteristics of VidVoice-Ref include: (1) **Large-scale cross-video subject pairs.** The dataset contains approximately 1M reference–target samples and over 1K hours of video, covering both single- and multi-subject settings. (2) **High-resolution content.** High-resolution videos dominate the dataset, with the majority of content at 1080P or above across both landscape and portrait formats. (3) **Rich structured captions.** Each video is annotated with subject appearance, actions and interactions, scenes, camera behavior, motion, visual style, and speech-related information, while target prompts use identity placeholders to maintain explicit subject-reference correspondence while preserving generation-relevant target descriptions. (4) **Complete audio-visual references.** Each subject reference retains the complete video and synchronized audio, with additional filtering for speech validity, speaker consistency, audio quality, and cross-video voice similarity. (5) **Diverse visual content.** The dataset covers diverse subject appearances, clothing, hairstyles, viewpoints, poses, actions, and indoor/outdoor scenes, as illustrated by the appearance and scene distributions in Fig. 2.

5 EXPERIMENTS

5.1 EVALUATION BENCHMARK

We construct a diverse benchmark of 120 cases for reference-conditioned human video generation. It spans diverse subject attributes, scene contexts, and human behaviors, covering both single-subject and multi-subject settings. Each case is paired with a structured text prompt describing the target subjects, actions, and environment, together with reference videos and their synchronized audio.

5.2 EVALUATION METRICS

Specifically, we evaluate model performance along five dimensions.

(1) Basic Quality. We assess the perceptual quality of generated audio and video using UTMOS Saeki et al. (2022) and Naturalness, respectively. UTMOS measures speech quality, while Naturalness is evaluated by Gemini-3.1-Pro on frames uniformly sampled from each generated video, considering visual artifacts, human appearance, and overall scene realism. The final Naturalness score is averaged over all sampled frames. Higher scores indicate better perceptual quality. **(2) Reference Consistency.** We evaluate the preservation of identity and visual characteristics from the reference videos. Facial identity consistency is measured using CurricularFace Huang et al. (2020) (Cur) and ArcFace Deng et al. (2019) (Arc) embeddings. For multi-subject cases, one-to-one matching is performed between generated and reference identities. We further measure holistic appearance consistency using CLIP Radford et al. (2021) (R-CLIP) and DINOv2 Oquab et al. (2023) (DINO) features.

(3) Voice Consistency. We evaluate whether the generated speech preserves the vocal characteristics of the reference speaker. Speaker similarity is measured using ECAPA-TDNN Desplanques et al. (2020) (ECAPA) and the zero-shot speaker similarity metric SIM-o Le et al. (2023). Higher scores indicate stronger preservation of speaker identity and voice characteristics.

(4) Audio–Visual Synchronization. We evaluate the correspondence between generated speech and visible lip motion using SyncNet Chung & Zisserman (2016) confidence (Conf). Higher confidence indicates stronger audio–visual synchronization.

(5) Instruction Following. We measure prompt adherence using CLIP-based video–text similarity (P-CLIP) Radford et al. (2021). The target prompt is compared with uniformly sampled generated frames, and the frame-level similarities are averaged. For structured prompts, we retain generation-relevant semantics, including subject actions and interactions, scene context, and camera behavior.

	Basic Quality		Reference Consistency				Voice Consistency		Audio–Visual Sync	Instruction Following
Model	**UTMOS↑**	**Naturalness↑**	**Cur↑**	**Arc↑**	**R-CLIP↑**	**DINO↑**	**ECAPA↑**	**SIM-o↑**	**Conf↑**	**P-CLIP↑**
Kling 3.0	1.9912	3.8114	0.5824	0.5300	0.7193	0.3078	0.2243	0.2416	1.3981	**0.3450**
Seedance 2.5	1.9168	3.8686	**0.6450**	**0.5897**	0.7470	0.3225	0.4510	**0.5113**	4.3034	0.3379
Wan 3.0	2.0956	**3.9554**	0.6124	0.5656	**0.7596**	**0.3641**	0.4472	0.4864	**4.4959**	0.3340
MiniMax H3	**2.6198**	3.8529	0.6214	0.5768	0.7436	0.3145	**0.4655**	0.5029	4.4693	0.3440

Table 1: Quantitative comparison of different video generation models on the VidVoice-Ref benchmark.

5.3 Quantitative Comparisons

Table 1 presents the quantitative comparison of four representative models, including Wan 3.0 Alibaba Cloud (2026), MiniMax H3 MiniMax (2026), Seedance 2.5 ByteDance Seed (2026), and Kling 3.0 (Video 3.0 Omni) Kuaishou Technology (2026), on the VidVoice-Ref benchmark. Overall, the evaluated models exhibit complementary strengths across different dimensions, while no single system achieves consistently strong performance in perceptual quality, reference consistency, audio–visual consistency, and instruction following.

For basic quality, MiniMax H3 achieves the highest UTMOS score, whereas Wan 3.0 obtains the best Naturalness, suggesting that perceptual speech quality and visual naturalness remain distinct strengths across current models. In terms of reference consistency, Seedance 2.5 performs best on CurricularFace and ArcFace, demonstrating stronger facial identity preservation, while Wan 3.0 achieves the highest R-CLIP and DINO scores, indicating better preservation of broader visual characteristics from the reference videos. This discrepancy suggests that face-centric identity similarity and holistic reference appearance capture complementary aspects of subject preservation, and strong facial identity consistency does not necessarily imply equally strong preservation of overall subject appearance. A similar pattern can be observed across voice consistency and audio–visual synchronization. H3 achieves the highest ECAPA score, Seedance 2.5 performs best on SIM-o, and Wan 3.0 obtains the strongest synchronization confidence, indicating that speaker identity preservation and speech–lip synchronization remain only partially coupled capabilities. For instruction following, Kling 3.0 achieves the highest P-CLIP score, although the differences among the evaluated models are relatively modest, suggesting broadly comparable prompt-level semantic alignment.

Overall, these results reveal a fragmented capability profile among current audio-visual reference-conditioned human video generation systems. Strong performance in facial identity, holistic appearance, speaker identity, or audio–visual synchronization does not necessarily translate into consistently strong performance across the other dimensions, highlighting the importance of evaluating reference-conditioned human video generation from multiple complementary perspectives.

5.4 Qualitative Comparisons

As shown in Fig. 3, all evaluated models capture the main scene and coarse subject attributes, but differ in prompt alignment and reference consistency. For prompt alignment, MiniMax H3 MiniMax (2026), Seedance 2.5 ByteDance Seed (2026), and Kling 3.0 Kuaishou Technology (2026) more consistently depict the specified cooking action, including the saucepan and stirring behavior, whereas Wan 3.0 Alibaba Cloud (2026) gradually shifts toward a talking portrait and weakens the requested interaction with the cookware. For reference consistency, the models generally preserve salient characteristics of the reference subject, such as the elderly appearance, gray hair, beard, and glasses, but facial identity and hairstyle consistency vary across frames. These results suggest that, even with complete audio-visual references, there remains room for further improvement in maintaining subject identity consistency.

6 Conclusion

In this paper, we present VidVoice-Ref, a large-scale dataset for reference-conditioned human video generation with audio-visual subject references. The dataset preserves complete reference videos together with synchronized audio and constructs identity-consistent reference–target pairs across diverse visual contexts with reliable speaker characteristics. VidVoice-Ref contains approximately

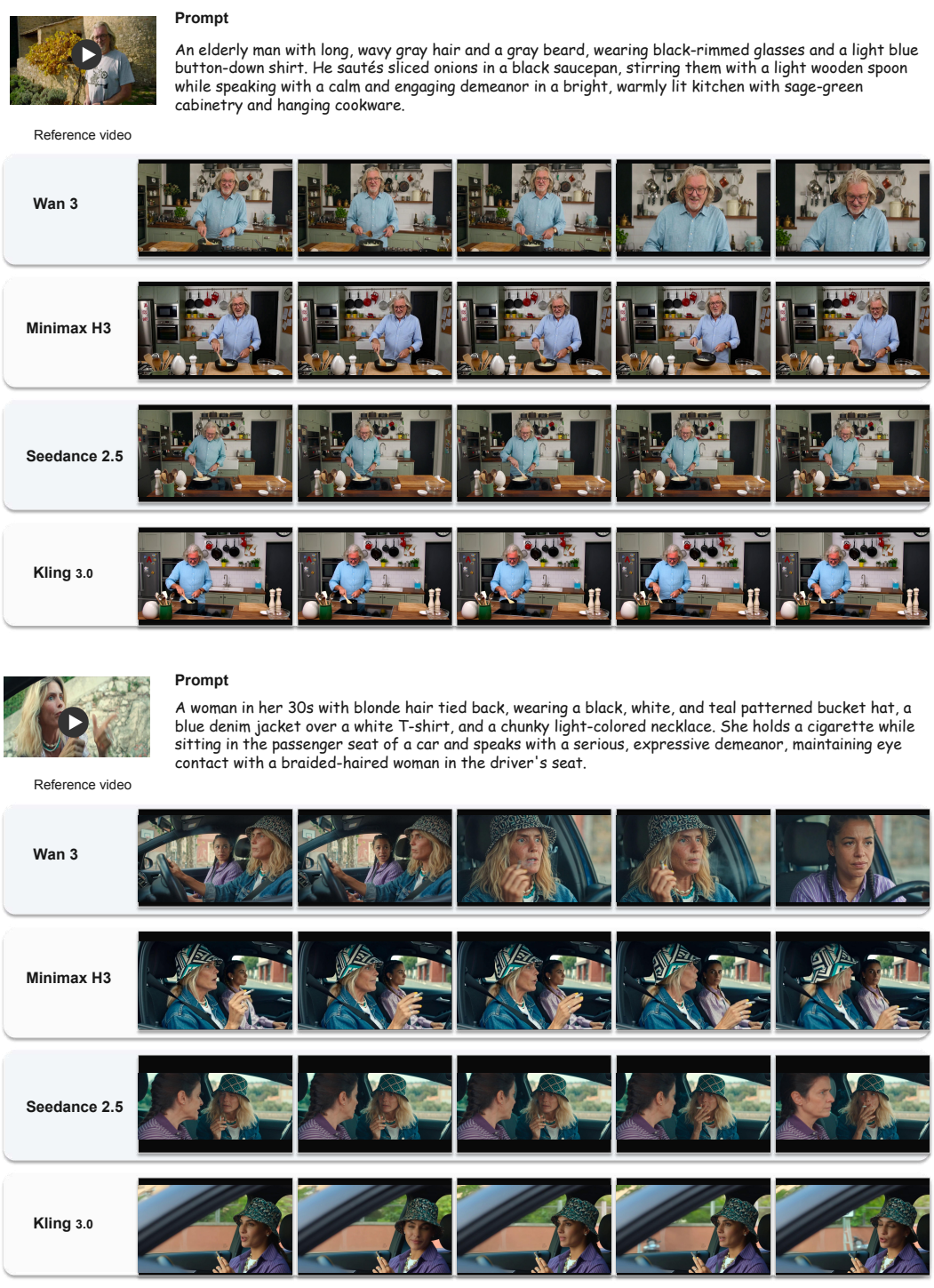


Figure 3: Qualitative Comparisons of different video generation models on the VidVoice-Ref benchmark.

1M samples spanning over 1K hours of video, covering both single- and multi-subject scenarios with structured audio-visual annotations. We hope VidVoice-Ref provides a practical data foundation for studying richer human subject representations beyond static image references and advancing audio-visual reference-conditioned human video generation.

REFERENCES

Alibaba Cloud. Wan 3.0: Leading ai video generation model. https://github.com/AlibabaCloud-Official/Wan3.0, 2026. Official model release.

Pokrzywa Baptiste, Nabyl Quignon, Yara Bahram, Muhammad Osama Zeeshan, Antitza Dantcheva, and Eric Granger. Beacon: Behavior and appearance control for subject-specific video generation. *arXiv preprint arXiv:2609.13264*, 2026.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1983–1987, 2023. doi: 10.21437/Interspeech.2023-105.

ByteDance Seed. One-take creation, flexible referencing: Introducing seedance 2.5. https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5, 2026. Official model release.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Human-centric video generation via collaborative multi-modal conditioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2939–2947, 2026a.

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*, 2026b.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6099–6110, 2025.

Yuheng Chen, Teng Hu, Yuji Wang, Qingdong He, Lizhuang Ma, and Jiangning Zhang. Spatial-temporal decoupled reference conditioning for identity-preserving text-to-video generation. *arXiv preprint arXiv:2606.02441*, 2026c.

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

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.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, pp. 1086–1090, 2018.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21086–21095, 2025.

Aviad Dahan, Moran Yanuka, Noa Kraicer, Lior Wolf, and Raja Giryes. ID-LoRA: Identity-driven audio-video personalization with in-context lora. *arXiv preprint arXiv:2603.10256*, 2026.

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

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation with subject disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024.

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*, 2025.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056. IEEE, 2018.

Zujin Guo, Zhenhui Ye, Yi Ren, Yuanming Li, Ce Chen, Zhibin Hong, and Chen Change Loy. Generate your talking avatar from video reference. *arXiv preprint arXiv:2604.27918*, 2026.

Bin Hu, Zipeng Qi, Guoxi Huang, Zunnan Xu, Ruicheng Zhang, Chongjie Ye, Jun Zhou, Xiu Li, and Jingdong Wang. Identity-consistent video generation under large facial-angle variations. *arXiv preprint arXiv:2603.21299*, 2026.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5900–5909. IEEE, 2020.

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.

Kuaishou Technology. Kling ai launches 3.0 model, ushering in an era where everyone can be a director. `https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be`, 2026. Official release of Kling Video 3.0 and Video 3.0 Omni.

Yixuan Lai, He Wang, Kun Zhou, and Tianjia Shao. Slot-id: Identity-preserving video generation from reference videos via slot-based temporal identity encoding. *arXiv preprint arXiv:2601.01352*, 2026.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, and Siyu Zhu. OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7752–7762, 2025.

Maomao Li, Zhen Li, Kaipeng Zhang, Guosheng Yin, Zhifeng Li, and Dong Xu. OmniCustom: Sync audio-video customization via joint audio-video generation model. *arXiv preprint arXiv:2602.12304*, 2026a.

Yunheng Li, Hengrui Zhang, Meng-Hao Guo, Wenzhao Gao, Shaoyong Jia, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Towards universal video MLLMs with attribute-structured and quality-verified instructions. *arXiv preprint arXiv:2602.13013*, 2026b.

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

Benjamin Liang, Ce Chen, Desmond Lin, Ivan Somov, Jiajun Zhao, Jiewei Yuan, Jingfeng Zhang, Junhao Huang, Nik Nolte, Pedram Haqiqi, et al. Avatar v: Scaling video-reference avatar video generation. *arXiv preprint arXiv:2606.13872*, 2026.

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 (ICCV)*, pp. 13847–13858, 2025.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, and Boxin Shi. MAVIN: Multi-shot audio-visual generation with customized narrative control. In *European Conference on Computer Vision (ECCV)*, 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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961. IEEE, 2025.

Erika Loc, Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. Development of a multimodal annotation framework and dataset for deep video understanding. In *Proceedings of the 2nd Workshop on People in Vision, Language, and the Mind*, pp. 12–16, 2022.

Zijie Meng, Jiwen Liu, Yufei Liu, Chengzhuo Tong, Xiaoqiang Liu, Yuanxing Zhang, Yulong Xu, and Pengfei Wan. ARGUS: Stacked multi-view identity mosaic injection for subject-preserving video generation. *arXiv preprint arXiv:2606.11670*, 2026.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. `https://www.minimax.io/blog/minimax-h3`, 2026. Official model release.

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. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Ziyang Song, Xinyu Gong, Bangya Liu, and Zelin Zhao. MV-S2V: Multi-view subject-consistent video generation. In *ACM SIGGRAPH 2026 Conference Papers*, 2026.

Cong Wang, Zhentao Yu, Hongmei Wang, Weicong Liang, Zixiang Zhou, Zilin Yang, Jiarong Ou, Rui Chen, Yuan Zhou, and Qinglin Lu. HarmoView: Harmonizing multi-view constraints for identity-consistent video generation. *arXiv preprint arXiv:2606.10839*, 2026a.

Zhenzhi Wang, Jian Wang, Ke Ma, Dahua Lin, and Bing Zhou. Talkverse: Democratizing minute-long audio-driven video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4527, 2026b.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Yulong Xu, Xinyue Liu, Shujuan Li, Huafeng Shi, Yan Zhou, Jiwen Liu, Xintao Wang, Yu-Shen Liu, and Huaibo Huang. Vera: Identity-faithful human subject-to-video generation. *arXiv preprint arXiv:2607.20247*, 2026.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Chongyang Ma, Jiebo Luo, Li Yuan, et al. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, et al. Movie facts and fibs (mf 2): A benchmark for long movie understanding. *arXiv preprint arXiv:2506.06275*, 2025.

Ailing Zeng, Casper Yang, Chauncey Ge, Eddie Zhang, Garvey Xu, Gavin Lin, Gilbert Gu, Jeremy Pi, Leo Li, Mingyi Shi, Shawn Wang, Sheng Bi, Steven Tang, Thorn Hang, Tobey Guo, Vincent Li, Xin Tong, Yikang Li, Yuchen Sun, Yue Zhao, Yuhan Lu, Yuwei Li, Zane Zhang, Zeshi Yang, and Zi Ye. LPM 1.0: Video-based character performance model. *arXiv preprint arXiv:2604.07823*, 2026.

Haojie Zhang, Di Wu, Bingyan Liu, Linjie Zhong, Yuancheng Wei, Xingsong Ye, Nanqing Liu, and Yaling Liang. MuSS: A large-scale dataset and cinematic narrative benchmark for multi-shot subject-to-video generation. *arXiv preprint arXiv:2604.23789*, 2026a.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, Xiu Li, et al. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. In *International Conference on Learning Representations*, volume 2026, pp. 117896–117926, 2026b.

Yunpeng Zhang, Qiang Wang, Fan Jiang, Yaqi Fan, Mu Xu, and Yonggang Qi. FantasyID: Face knowledge enhanced id-preserving video generation. *arXiv preprint arXiv:2502.13995*, 2025a.

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

Beyond Static References: Audio-Visual Reference Conditioning for Human Video Generation

Supplementary Material

In these supplementary materials, we present the following: the definition of audio-visual reference-conditioned human video generation (Sec. A), detailed dataset construction of VidVoice-Ref (Sec. B), visual reference construction and audio reference refinement (Secs. B.2 and B.3), semantic annotation and prompt construction (Sec. B.4), key filtering parameters (Sec. B.6), the structured captioning prompt and representative prompt examples (Sec. C), and representative dataset visualizations (Sec. D).

A Audio-Visual Reference-Conditioned Human Video Generation

Reference-conditioned human video generation aims to synthesize novel videos of one or multiple human subjects while preserving their identities under new scenes, viewpoints, poses, motions, and interactions. Existing approaches predominantly represent each referenced subject using a single image or a small set of static images. In this work, we extend such static subject references to an **audio-visual reference**, consisting of a reference video together with the corresponding speech audio of the same subject.

Given a textual prompt P and N subject references,

$$\mathcal{R} = \left\{ \left(V_i^{\mathrm{ref}}, A_i^{\mathrm{ref}} \right) \right\}_{i=1}^{N}, \tag{4}$$

the goal is to generate a target video V^{tar} that follows the textual description while preserving the identities of the referenced human subjects.

The two reference modalities provide complementary subject-specific information. The reference video V_i^{ref} preserves temporally varying visual observations of the subject, including appearance across different viewpoints, pose and expression variations, dynamic visual characteristics, and motion patterns. Unlike isolated reference images, these observations jointly characterize how identity-related visual cues evolve over time. The reference audio A_i^{ref} contains speech from the corresponding subject and provides complementary vocal identity cues, such as **voice timbre** and other speaker-specific acoustic characteristics. Accordingly, the audio is used as a subject reference rather than merely as background sound or semantic content. This formulation extends conventional static visual reference conditioning toward **audio-visual reference conditioning** for human video generation. For multi-subject generation, each subject is associated with an independent audio-visual reference pair, while the textual prompt specifies the target scene, actions, and interactions among the referenced subjects.

B Dataset Construction Details

This section further details the construction pipeline of VidVoice-Ref, including visual reference construction, audio reference refinement, semantic annotation, and final sample organization. Our pipeline is target-centric: for each target video, we first identify the subjects appearing in the target, retrieve visually consistent reference clips, and then progressively refine the associated audio to obtain reliable audio-visual subject references.

B.1 Overview and Design Rationale

The construction of VidVoice-Ref separates visual identity matching from voice identity matching. Visual references are first retrieved according to facial identity and visual diversity, while their associated audio is subsequently filtered according to speaker validity, voice consistency, and reference–target speaker similarity. This design allows each retained reference to preserve both temporally varying visual observations and subject-specific vocal characteristics.

For a target video containing N subjects, each subject is processed independently. The resulting reference set is represented as

$$\mathcal{R} = \left\{ \left(V_i^{\mathrm{ref}}, A_i^{\mathrm{ref}} \right) \right\}_{i=1}^{N}, \tag{5}$$

where V_i^{ref} and A_i^{ref} denote the reference video and its corresponding reference audio, respectively. This formulation naturally supports both single-subject and multi-subject reference-conditioned human video generation.

B.2 Visual Reference Construction

Frame sampling and identity representation. We uniformly sample frames from each candidate clip at approximately 1 FPS and use InsightFace for face detection and identity feature extraction. Only sufficiently reliable face observations are used for identity modeling. Specifically, we require a detection confidence of at least 0.7, a face region of at least 1024 pixels^2, and a face-to-frame area ratio of at least 0.007. Each valid face is represented by an L_2-normalized identity embedding.

Target subject clustering. Faces observed in each target video are grouped according to cosine distance in the identity embedding space. We use a distance threshold of 0.5 and additionally exploit temporal continuity between neighboring observations when assigning faces to identity clusters. Cluster centers are iteratively updated and nearby clusters are merged, yielding one identity prototype for each detected target subject.

Identity-based reference retrieval. For each target identity, we compare its prototype against face embeddings extracted from candidate reference clips. Candidates are ranked according to cosine similarity, and those with identity similarity within $[0.65, 0.90]$ are retained. When multiple frames originate from the same reference video, only the frame with the highest identity similarity is kept.

Diversity-aware reference selection. Identity similarity alone may retrieve highly redundant observations. We therefore further compare the retained candidates using DINO visual features. Starting from the highest-ranked candidate, a new reference is retained only if its cosine distance from all previously selected references is greater than 0.4. This removes near-duplicate observations while preserving variations in viewpoint, pose, expression, and scene context.

Video-level verification. The retrieved source videos are further checked at the temporal level. We sample reference videos at approximately 2 FPS and perform face detection on the sampled frames. A reference is rejected if any valid sampled frame contains multiple faces, while at least one sampled frame must contain a valid single-face observation. The remaining clips form the visual reference pool used for subsequent audio refinement.

B.3 Audio Reference Refinement

Speaker and speech filtering. For each visually verified reference, we first inspect its audio track and perform speaker diarization using Pyannote. We retain references containing exactly one detected speaker and require the accumulated valid speech duration to exceed one second.

Voice quality and intra-reference consistency. The audio is converted to 16 kHz and separated into vocal and non-vocal components using Demucs. Silero VAD is then applied to identify active speech intervals. The valid speech is divided into non-overlapping one-second windows, from which speaker embeddings are extracted using Resemblyzer. At least two valid windows are required, and the minimum pairwise cosine similarity among their speaker embeddings must be at least 0.5.

We further compare the RMS energy of the vocal and non-vocal components within active speech regions. A reference is retained only when the vocal energy is not lower than the background component, i.e.,

$$E_{\mathrm{voice}} - E_{\mathrm{bg}} \geq 0 \ \mathrm{dB}. \tag{6}$$

Reference–target voice matching. To associate the reference voice with the corresponding target content, we extract speaker embeddings from the valid speech segments of the target video and compare them with the reference speaker representation. For a reference V_i^{ref}, we compute

$$s_{\mathrm{voice}} = \max_{a_j \in \mathcal{A}^{\mathrm{tar}}} \mathrm{sim}\left(e_a(V_i^{\mathrm{ref}}), e_a(a_j)\right), \tag{7}$$

where $\mathcal{A}^{\mathrm{tar}}$ denotes the valid target speech segments. We require $s_{\mathrm{voice}} \geq 0.6$, and the best-matching target speech segment must be longer than one second.

The current pipeline uses speaker similarity as the audio matching criterion and does not explicitly enforce per-subject audio-visual association through lip-motion tracking.

Reference audio extraction. After audio filtering, the reference audio delivered with each sample is extracted directly from the original audio track of the retained reference video. The separated vocal track produced by Demucs is used only for quality assessment and speaker representation, rather than as the final reference audio.

B.4 Semantic Annotation and Prompt Construction

We use Qwen3-Omni to generate structured audio-visual annotations from each video together with its synchronized audio. The annotations describe subject appearance, actions and interactions, scene context, camera behavior, motion dynamics, visual style, speech activity, and other salient audio-visual events.

For target videos, the structured annotations are further converted into generation prompts. Subjects are represented using explicit placeholders such as `<ID_A>` and `<ID_B>` to maintain consistent correspondence with their associated audio-visual references. Explicit person names, when available in the annotations, are excluded from the final generation prompts. The resulting prompt preserves generation-relevant target descriptions, including subject appearance, actions, interactions, scene context, camera motion, and temporal dynamics.

For reference videos, we preserve richer descriptions of subject appearance, viewpoint, pose, motion, and audio characteristics.

B.5 Final Sample Organization

After visual and audio refinement, each training sample is represented as

$$\left(\mathcal{R}, P, V^{\mathrm{tar}}\right), \tag{8}$$

where P denotes the target generation prompt and $\mathcal{R}$ contains the audio-visual references associated with the subjects appearing in V^{tar}.

For each target subject, we retain up to three valid audio-visual reference candidates in the full dataset. For downstream settings requiring a single reference per subject, one reference-video–audio pair can be selected from this candidate set.

B.6 Key Filtering Parameters

Table 2 summarizes the major filtering parameters used during dataset construction.

C Prompt Used in Video Caption

C.1 Structured caption generation.

As shown in Fig. 4, we design a structured captioning prompt conditioned on the target video together with its reference videos and audio tracks. The reference video provides the subject's visual identity and appearance, while the reference audio provides voice timbre and other speaker-specific characteristics; target-side content, including actions, interactions, scenes, camera behavior, and

Criterion	Parameter	Value
Frame sampling	Sampling rate	$\sim$1 FPS
Face detection	Confidence	≥ 0.7
Face visibility	Face area	≥ 1024 px^2
Face visibility	Face/frame ratio	≥ 0.007
Subject clustering	Cosine distance	< 0.5
Identity retrieval	Cosine similarity	$[0.65, 0.90]$
Visual diversity	DINO cosine distance	> 0.4
Video verification	Sampling rate	$\sim$2 FPS
Speaker filtering	Number of speakers	$= 1$
Speech filtering	Valid speech duration	> 1.0 s
Voice consistency	Minimum similarity	≥ 0.5
Audio quality	Voice–background energy	≥ 0 dB
Voice matching	Ref.–target similarity	≥ 0.6
Voice matching	Target speech duration	> 1.0 s

Table 2: Major filtering criteria used in the VidVoice-Ref construction pipeline.

dialogue, is described solely from the target video. The generated caption is organized with structured tags such as <Event>, <Role>, <Background>, <style>, and <Scene>, where persistent identifiers such as <ID_A> and <ID_B> maintain consistent subject-reference correspondence across shots. Figures 5 and 6 show representative examples for single-reference and two-reference settings, respectively. Shot-level descriptions further preserve the temporal order of subject actions, camera changes, and synchronized audio events, yielding fine-grained and temporally coherent annotations.

D Dataset Visualization

Figures 7 and 8 visualize representative samples from VidVoice-Ref. The single-reference cases cover diverse subjects, actions, and scene transitions, where the same subject appears across distinct visual contexts between the reference and target videos. The multi-reference cases further involve multiple referenced subjects and their interactions in the target video, illustrating the dataset's support for multi-subject identity preservation and subject binding. Overall, these examples demonstrate the diversity of reference–target pairs in terms of subject appearance, actions, interactions, and scene context.

You are a professional multimodal video captioner. Given one target video and one or more ordered reference videos/audio tracks, produce a dense, factual caption for audiovisual video generation. First establish each reference asset's role, then summarize the global event, and finally describe the visible and audible process shot by shot.

Source boundaries:
• The target video is the sole authority for target clothing, props, actions, expressions, setting, composition, camera work, timing, dialogue, subtitles, music, and sound.
• <Video_N> provides the identity/appearance anchor for its matched target person. <Audio_N> provides only that same person's voice timbre and vocal characteristics unless reuse is explicitly requested. Never copy reference dialogue, actions, clothing, or background into the target caption.
• Match a reference person to a target person using face, body, and voice evidence. Keep index N aligned across each reference video and audio pair. Never force an uncertain match or invent an identity.

Analysis procedure:
1. Watch every reference asset, then the complete target video with audio.
2. Assign <ID_A>, <ID_B>, ... to persistent main people/objects by first appearance and reuse each ID consistently across shots.
3. Detect every real cut and record each shot's duration to 0.1 second. Do not create a new shot for camera motion alone.
4. Transcribe spoken words and visible text verbatim in their original language. Mark the correct speaker, including off-screen speech; never translate, paraphrase, or guess unclear content.

Return only one caption using the following tags and order. Do not add Markdown, explanations, or unsupported details.

<header>State FPS, playback speed, audio-quality score and level, volume, and mono/stereo channel layout.</header>

<Event>First state the reference relationship: which person comes from each <Video_N> and that <Audio_N> controls only the corresponding voice timbre. Then summarize, in 2–4 sentences, who does what, where it happens, the content type, and the overall production tone.</Event>

<Role>State the number of main objects. Define every <ID_X> once with observable age range, ethnicity only when visually supportable, build, face/hair, clothing, accessories, salient props, expression, and spoken accent/language. For each matched person, explicitly write “the person from <Video_N>” and “Its voice timbre is derived from <Audio_N>.” Include an IP/person name only when confidently recognizable; otherwise omit it.</Role>

<Background>Describe the target environment, layout, important foreground/mid-ground/background elements, spatial relations, lighting sources, time-of-day cues, and dominant colors. Do not import the reference background.</Background>

<Music>Include only when music is audible. Describe whether it is diegetic or non-diegetic, instrumentation, tempo, rhythm, intensity, and temporal development.</Music>

<Subtitle>Include only when subtitles or other meaningful text are visible. Preserve every string exactly and report its order, position, font style, color, and outline/background.</Subtitle>

<ChannelLogo>Include only when a logo/watermark is visible; describe its exact text, appearance, and position.</ChannelLogo>

<style>Describe medium and genre, realism, cinematography, lighting, palette, saturation, contrast, brightness, color temperature, and editing character.</style>

<Scene>The video contains N shots.
For every shot, output:
<shot_N>[D.D seconds]: <section>Describe framing and angle; camera movement or static state; depth of field; each <ID_X>'s position, appearance, actions, reactions, and state changes in temporal order; relevant background changes; and synchronized audible events. Write dialogue as “<ID_X> says, <speech>exact original words</speech>” or mark the speaker as off-screen. End the section after the shot's final visible/audible event.</section>
Continue until every shot is covered, preserving chronology and cross-shot continuity.</Scene>

Quality rules: Be specific, visual, chronological, and non-redundant. Prefer observable facts over interpretation. Keep reference identity stable while preserving target-video evidence. Ensure durations sum to the target duration; every ID, reference index, speaker, quotation, and shot number must be internally consistent.

Inputs:
Target video: {target_video}
Reference videos: {reference_videos}
Reference audio tracks: {reference_audio}
Optional AV metadata: {fps, duration, audio_channels}

Figure 4: Structured captioning prompt for VidVoice-Ref.

Example 1 — Single-reference prompt

Role: There is 1 main object in the video. A middle-aged East Asian woman with fair skin and dark hair pulled back into a low bun, with loose strands framing her face. She is wearing a beige, textured wool coat over a cream-colored turtleneck sweater. Her expression is serious and contemplative, with her gaze directed slightly off-camera to her right. She is seated, and her posture is upright but relaxed. She speaks in a standard Mandarin accent. [Identified IP: Li Xiaoran - Famous Chinese actress]

Event: The woman sits in an office, speaking seriously to someone off-camera. This looks like a dramatic scene from a television series or movie. It features high-quality cinematography with a serious, professional tone.

Background: The background is an indoor office setting, slightly out of focus to emphasize the woman. Behind the woman, there is a white cabinet with drawers. On top of the cabinet, there are several framed items, including what appears to be a certificate and a round decorative object. To the right, a tall, dark thermos is visible. The lighting is soft and diffused, creating a calm and professional atmosphere. The walls are a muted, neutral color. A portion of a computer monitor is visible on the far right edge of the frame.

Audio Guidance: Use the reference audio only as a reference for the subject's voice timbre and vocal characteristics.

Figure 5: Example of a single-reference generation prompt in VidVoice-Ref.

Example 2 — Two-reference prompt

Role: There are 2 main objects in the video. <ID_A>: A young East Asian man, appearing to be in his 20s, with short, dark, slightly wavy hair and bangs, who speaks with a standard Korean accent. He is wearing a long-sleeved, salmon-pink crewneck shirt and a thin black bracelet on his left wrist. He is positioned behind <ID_B>, leaning in close to her. [Identified IP: Lee Jong-suk - Famous South Korean actor] <ID_B>: A young East Asian woman, appearing to be in her 20s, with shoulder-length brown hair with bangs, who speaks with a standard Korean accent. She is wearing a plain white t-shirt. She is seated in front of <ID_A>, facing forward. [Identified IP: Han Hyo-joo - Famous South Korean actress]

Event: <ID_A> affectionately styles <ID_B>'s hair in a cozy indoor setting. This looks like a romantic scene from a television drama. It has a warm, intimate vibe with high-quality, cinematic lighting.

Background: The scene takes place indoors, likely in a cozy apartment setting. To the left, there is a dark-framed glass door, reflecting some light. Behind <ID_A>, a section of a red brick wall is visible, adorned with dark metal hooks. To the right of <ID_B>, there is a large window covered by sheer white curtains that diffuse the light coming from outside, creating a soft, bright atmosphere. The lighting is warm and intimate, suggesting an evening setting. The overall color palette is warm, dominated by the pink of <ID_A>'s shirt, the white of <ID_B>'s shirt and the curtains, and the dark tones of the background elements. Laughter can be heard.

Audio Guidance: Use each reference audio only as a reference for the corresponding subject's voice timbre and vocal characteristics.

Figure 6: Example of a two-reference generation prompt in VidVoice-Ref.

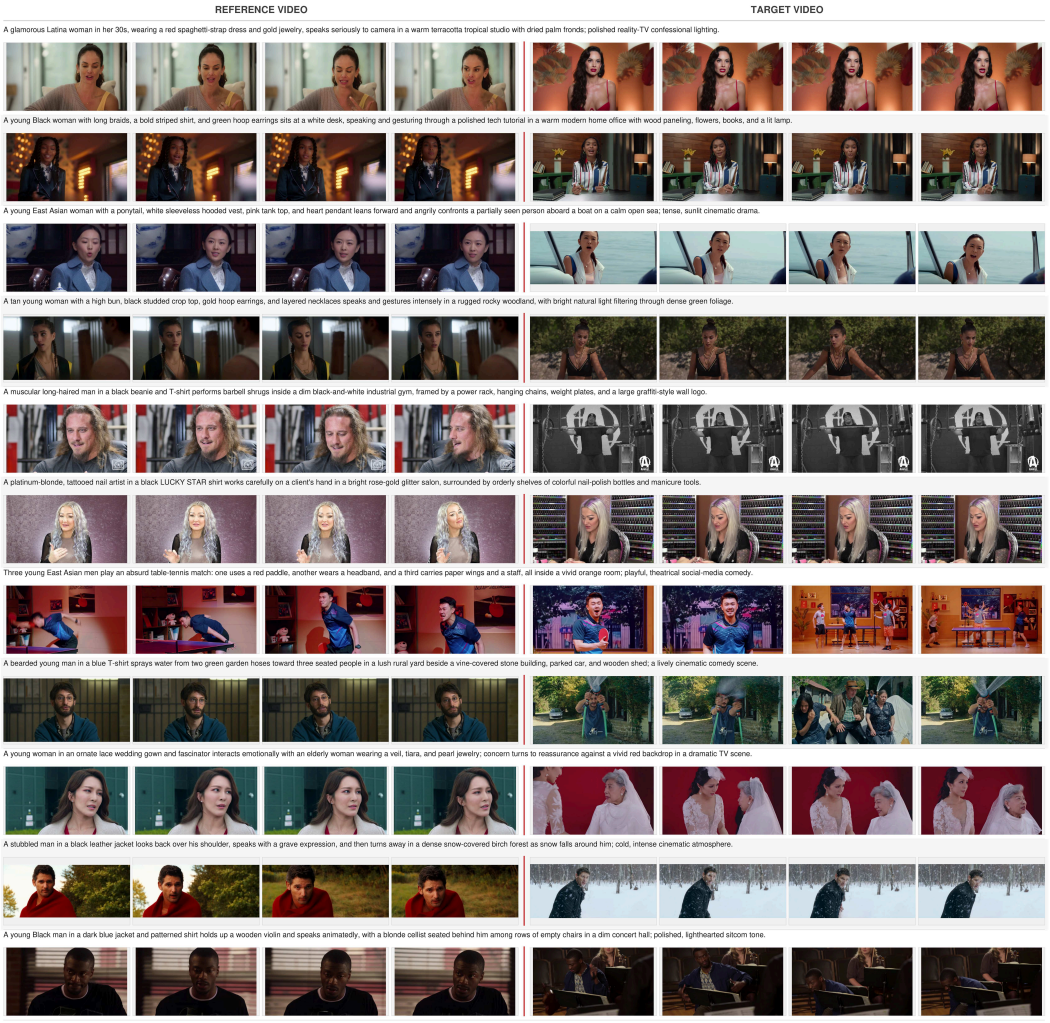


Figure 7: Representative single-reference samples from the VidVoice-Ref benchmark.

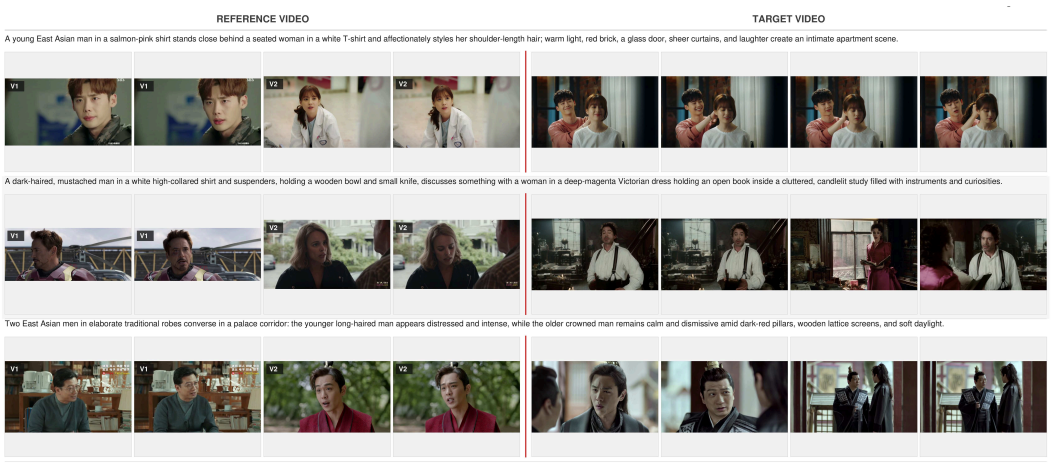


Figure 8: Representative multi-reference samples from the VidVoice-Ref benchmark.