OmniHuman: A Large-scale Dataset and Benchmark for Human-Centric Video Generation

Lei Zhu[1‡*], Xing Cai[2*], Yingjie Chen[2], Yiheng Li[6], Binxin Yang[2], Hao Liu[2†], Jie Chen[5,3,1,4✉], Chen Li[2✉], and Jing LYu[2]

[1] School of Electronic and Computer Engineering, Peking University, Shenzhen,China
[2] WeChat Lab, Tencent, China
[3] Pengcheng Laboratory, Shenzhen, China
[4] AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China
[5] School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen, China.
[6] Chinese Academy of Sciences, Beijing, China

Abstract. Recent advancements in audio-video joint generation models have demonstrated impressive capabilities in content creation. However, generating high-fidelity human-centric videos in complex, real-world physical scenes remains a significant challenge. We identify that the root cause lies in the structural deficiencies of existing datasets across three dimensions: limited global scene and camera diversity, sparse interaction modeling, and insufficient individual attribute alignment. To bridge these gaps, we present **OmniHuman**, a large-scale, multi-scene dataset designed for fine-grained human modeling. OmniHuman provides a hierarchical annotation covering video-level scenes, frame-level interactions, and individual-level attributes. To facilitate this, we develop a fully automated pipeline for high-quality data collection and multi-modal annotation. Complementary to the dataset, we establish the OmniHuman Benchmark (**OHBench**), a three-level evaluation system that provides a scientific diagnosis for human-centric audio-video synthesis. Crucially, OHBench introduces metrics that are highly consistent with human perception, filling the gaps in existing benchmarks by providing a comprehensive diagnosis across global scenes, relational interactions, and individual attributes. Experiments show that fine-tuning on only 20% of OmniHuman significantly boosts performance, validating its effectiveness in advancing complex scenario modeling. The dataset is available at `https://huggingface.co/datasets/julia527/OmniHuman`

Keywords: Audio-visual dataset · Human-centric video generation

1 Introduction

The new generation of video-audio joint generation models, represented by Sora 2 [30], ovi [29] and Veo 3.1 [14], are gaining significant attention and are widely

[*] Equal contribution. [†]Project leader. ✉Corresponding authors
[‡] This work was done during Lei Zhu's internship at WeChat.

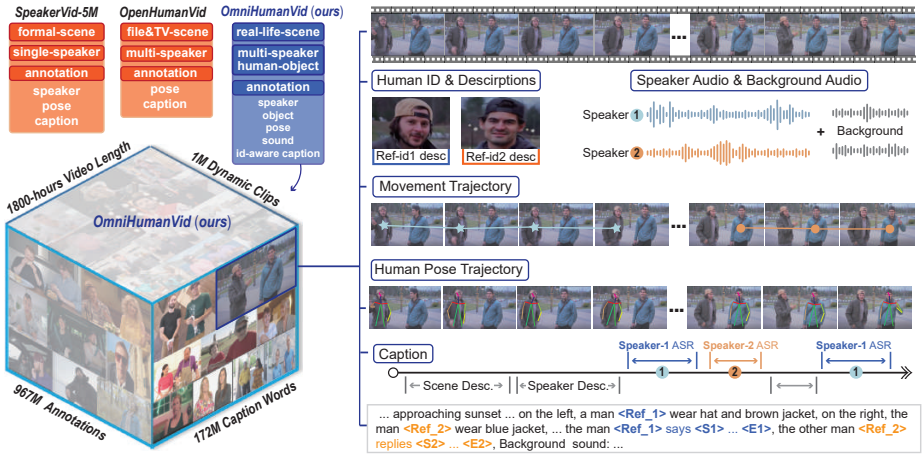

Fig. 1: OmniHuman: a 1M-video, 1800 hours, 80K-identity dataset with hierarchical annotations covering diverse natural scenes and social interactions.

applied in downstream tasks such as film production, social media, and interactive entertainment. Human behavior, serving as the visual focus and narrative core, carries extremely high semantic density and presents unprecedented challenges for the fine-grained control of generative models. Although existing models have achieved promising results in generating single-person close-ups or isolated actions, their ability to generalize to complex, real-world physical scenes remains severely hindered. We argue that the root cause lies in the systematic and structural deficiencies of current human-centric datasets across core dimensions. Specifically, these deficiencies are manifested in the following three levels.

(i) Global: Scene and Camera Diversity. Existing human-centric video datasets are mostly limited to controlled environments (e.g., formal or semi-formal scenes), causing poor generalization to complex real-world scenes. Most also lack ambient audio and environment annotations, hindering models from learning the coupling between human actions and environmental sounds, leading to semantic audio-visual mismatches in generation. Limited diversity in shot scales and camera movements further induces artifacts like feature instability and background distortion in specific shot types. ***(ii) Interaction: Person-Person and Person-Object.*** Current datasets focus predominantly on single-person videos, with multi-person interactions accounting for less than 3% [26]. This imbalance results in three key deficiencies in multi-person scenarios: relational distortion (lack of authentic social dynamics), audio-visual misassignment (failure to map speech to the correct speaker), and identity drift (facial feature blending or misalignment). Moreover, videos with person-object interactions such as precise hand-tool contact are scarce, limiting the modeling of physical interactions and narrowing the range of generated behaviors.

To address these challenges, we built an automatic pipeline to collect and annotate high-quality video and audio data. The annotations cover three lev-

els: video-level scene information, frame-level interactions, and individual-level attributes. This led to the creation of **OmniHuman**, a multi-scene dataset for detailed human modeling. A comparison with other datasets is shown in Table 1. Our dataset introduces new features at three levels: ***(i)Global Level***: OmniHuman includes most common real-life scenes. We provide clear labels for scene types and background sounds. This helps models learn to generate videos where people and their surroundings match in both look and sound. The videos also include different camera angles and both static and moving shots. ***(ii)Interaction Level***: We built a large collection of videos showing two people interacting—talking, working together, or competing. We label each person's actions and identity, so models can learn who does what in group scenes, avoiding issues like mixed-up identities or unnatural interactions. The dataset also includes videos where people interact with objects, expanding from social to physical interactions. ***(iii)Individual Level***: We selected high-quality videos with good lip-sync and clear visuals. For each person, we provide detailed labels: video-side labels include person ID, body movements, and face/hand clarity; audio-side labels include speech text, speaker ID, emotion, and lip-sync quality.

Based on the above hierarchical analysis and dataset construction, we establish a corresponding benchmark, OmniHuman Benchmark (**OHBench**). Existing benchmarks [17,18] have two key limitations: metrics misaligned with human perception and missing critical dimensions. At the scene level, they lack evaluation of background plausibility across shot types and semantic audio-visual consistency. At the interaction level, they miss assessments of interaction naturalness, audio-visual assignment accuracy, and human-object interaction plausibility. At the individual level, subject consistency and speech quality are overlooked. These gaps hinder accurate model evaluation in complex real-world scenarios. To address this, we select diverse samples from our dataset with domain gaps relative to the training data and construct a three-level evaluation system for comprehensive model diagnosis. Through this hierarchical construction, from global scenes to interaction relationships and then to individual attributes, the OmniHuman dataset and its corresponding benchmark form a closed loop. Specifically, the main contributions of this work are threefold:

- We construct **OmniHuman**, a large-scale, multi-scene audio-visual dataset. Its effectiveness is underscored by substantial performance gains achieved by fine-tuning an open-source model on a mere 20% of the data.
- We develop a fully automated hierarchical fine-grained annotation pipeline that caters to the diverse needs of downstream tasks across video, frame and individual levels.
- We establish **OHBench**, a comprehensive benchmark aligned with human perception for human-centric multimodal video generation, enabling comprehensive diagnosis of model performance in real-world scenarios.

Table 1: Comparison of different mainstream datasets. Columns are color-coded by levels: Global, Interactional, and Individual.

Dataset	Multi-scene	Shot-type	Camera	Sound	person-person	person-object	Speaker-anno	Pose-anno	Object-anno
ActivityNet [3]	✓	✓	✗	✗	✗	✗	✗	✗	✗
TikTok-v4 [11]	✓	✓	✗	✗	✗	✗	✗	✓	✗
Openhumanvid [25]	✓	✓	✓	✗	✓	✗	✗	✓	✗
CelebV-HQ [59]	✗	✗	✗	✓	✓	✗	✗	✗	✗
VoxCeleb2 [7]	✓	✗	✗	✗	✗	✗	✗	✗	✗
HDTF [57]	✗	✗	✗	✗	✗	✗	✗	✗	✗
SpeakerVid-5M [55]	✗	✓	✓	✗	✗	✗	✓	✓	✗
OmniHuman (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 Related Work

2.1 Human-Centric Video Generation

Audio-visual synchronized video generation has evolved from cascaded systems to a native bimodal paradigm [24, 27, 28, 31, 37, 45], jointly denoising visual and audio streams in a unified architecture. Early work like MM-LDM [36] explored multimodal latent diffusion for joint synthesis, while JavisDiT [27] enhanced alignment via hierarchical spatio-temporal prior estimation, though it struggles in complex human-centric scenes. Meanwhile, proprietary models such as Sora2, Veo3.1, and Wan2.5 [42] set industry benchmarks in physical simulation and high-resolution one-pass generation. In the open-source human-centric domain [19, 44, 54], Ovi uses symmetric twin-tower fusion, MOVA [6] employs mixture-of-experts, and LTX-2 [15] adopts asymmetric dual-stream for efficiency and precision. For speech-to-video (S2V), recent breakthroughs include OmniAvatar [13] with multi-hierarchical audio embeddings, MultiTalk [38] using Label Rotary Position Embedding for multi-character audio binding, and InfiniteTalk [52] enabling infinite-length generation via sparse-frame dubbing.

2.2 Evolution of Human-Centric Datasets and BenchMarks

Human-centric video datasets and evaluation benchmarks have evolved in synergy, transitioning from early constrained portraits (e.g., VoxCeleb [7], HDTF [57]) toward interactive scenarios (e.g., OpenHumanVid [25], SpeakerVid-5M [55]). Despite this progress, systematic gaps persist in scene diversity, holistic environment-aware acoustics, complex person-object physical interactions, and fine-grained attribute persistence. Correspondingly, while assessment paradigms have shifted from visual aesthetics (VBench [20]) toward multi-dimensional semantic and physical diagnostics (VABench [18], SVBench [53], MTAVG-Bench [58], PhyAVBench [49]), current benchmarks remain insufficient for evaluating dyadic audio-visual assignment, background plausibility, and precise physical contact, often exhibiting discrepancies with human perception. To bridge these dual gaps, we propose OmniHuman, a hierarchically annotated dataset, and its companion OHBench, a three-tier evaluation system that enables a comprehensive diagnosis of model deficiencies in complex dynamic scenarios through strictly aligned, fine-grained evaluation dimensions.

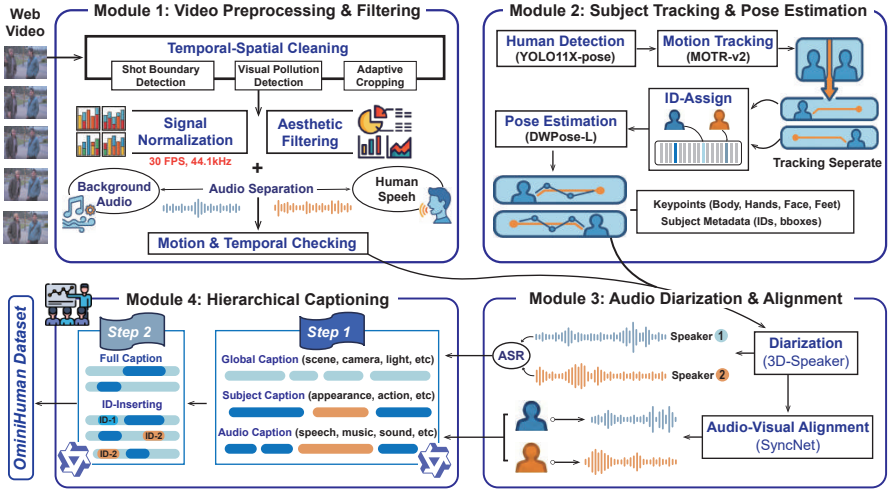

Fig. 2: OmniHuman employs a fully automated pipeline for high-quality data collection and fine-grained annotation, with each module applying progressive filtering to ensure both video quality and annotation accuracy.

3 OmniHuman

The construction process of OmniHuman consists of several decoupled modules: video preprocessing & filtering, subject tracking and pose estimation, audio-visual alignment and synchronization verification, and hierarchical multi-modal caption generation. Each module applies progressive filtering to ensure both video quality and annotation accuracy. Through modular deployment, the full data pipeline operates at scale in a streaming manner. The complete construction pipeline is illustrated in Fig. 2.

3.1 Video Preprocessing & Filtering.

This module forms the foundation of the data pipeline and aims to transform raw videos from open, complex scenarios into spatiotemporally consistent, high-availability segments with clean signals. This module is decomposed into four hierarchical stages for progressive purification of multi-source video data.

Temporal–spatial cleaning. Temporally, we obtain semantically coherent short clips through shot segmentation using the TransNetV2 [35] algorithm. Spatially, we use OCR and logo detection algorithms to estimate text-contaminated regions, then crop frames to retain the main subject area, thereby removing visual pollutants such as subtitles and logos. This process establishes foundational visual purification, improves frame quality, and ensures controllable clip duration.

Audio governance and signal standardization. This layer ensures audio usability and consistency by removing samples with missing tracks, abnormal

duration, or low quality (high silence ratio and low volume). All clips are standardized to 30 FPS and 44.1 kHz to eliminate potential source clock offsets. Demucs [9] is then applied for four-source separation, extracting vocals as the target track and mixing the remaining tracks as background. This design provides clean audio for subsequent audio-content processing tasks, including ASR and lip-audio alignment.

Aesthetic filtering. This layer adopts a lightweight-to-heavy filtering strategy. First, frame-level CLIP [32] aesthetic scoring rapidly removes low-quality clips. Next, OCR-based text-density analysis and watermark detection estimate the density and confidence of residual text, subtitles, and watermarks, and filter clips accordingly. Furthermore, DOVER [46] evaluates composition and clarity at the video level and filters clips based on the resulting scores. This process further enhances visual cleanliness and video quality.

Motion filtering and temporal consistency. Finally, we perform motion and temporal consistency checks. First, we use UniMatch [51] flow to analyze motion quality, retaining valid motion while filtering static, shaky, or abnormal segments. Then, we enforce temporal continuity by combining perceptual hashing with SigLIP [41] cosine similarity to filter semantically inconsistent frames. The four-tier pipeline progressively optimizes spatial-temporal usability, audio purity, visual fidelity, and temporal semantic coherence.

3.2 High-Fidelity Subject Tracking and Pose Estimation.

Spatio-temporal Multi-Object Tracking. To capture human motion in interactive scenarios, we build a highly robust multi-object tracking pipeline. We use YOLOv11 [16] to detect human instances, and feed NMS-refined detections into MOTRv2 [56] to model cross-frame associations via query propagation. We allow a maximum loss span of 5 frames to handle short-term occlusion. The pipeline outputs per-identity metadata as $B_i = \{(box_m, score_m)\}_{m=t_{start}}^{t_{end}}$, where box_m and $score_m$ denote the detection box and confidence score at frame m.

Fine-grained Whole-body Pose Representation. Based on identity-coherent tracking, we perform fine-grained 2D whole-body pose estimation for each instance using DWPose-L for body, face, and feet, with dedicated hand detection and optimization. The pipeline performs frame-wise detection of 134 full-skeletal keypoints for each tracked instance, denoted as $\mathcal{K} = \{k_j\}_{j=0}^{133}$, where each keypoint $k_j = (x_j, y_j, c_j)$ consists of 2D coordinates and confidence $c_j \in [0, 1]$. These keypoints cover body core, foot support, facial features, and both hands. To ensure facial fidelity, we compute a clarity score $C_s = \mathrm{Var}(\Delta R)$ for face region R based on Laplacian variance, and remove videos with persistently sub-threshold facial clarity. For identity assignment, we extract face embeddings(ArcFace [10]) from the frame with the highest average confidence of face keypoints in each track. We assign the same ID to frames with embedding similarity above 0.55, yielding a unique person ID per video.

3.3 Audio-Visual Alignment and Synchronization Verification.

In human-centric scenarios, precise synchronization between audio and visual subjects is essential for a high-quality dataset. This module first uses the vocal signal A_{vocals} extracted in Section 3.1 and performs speaker diarization with 3Dspeaker [4]. It identifies M active speech intervals $\{[t_{start}, t_{end}]_m\}_{m=1}^{M}$ with unique speaker indices. To match speaking intervals with visual identity trajectories while suppressing voice-over interference, we further introduce SyncNet [8] to resolve audio-visual attribution. SyncNet takes face bounding-box sequences from Section 3.2 as visual input and computes cross-correlation between face regions and each speech interval in a joint embedding space, producing a synchronization score S_{sync}. The system then applies greedy matching to assign each audio segment to the visual ID trajectory with the highest response. A sample is retained only when all detected subjects satisfy S_{sync} above a preset threshold and the temporal offset is within 3 frames. Unmatched audio segments are preserved as background audio. We then apply ASR (FunASR-Nano [2]) to each retained audio segment to obtain detailed speech transcripts. For each subject i, the matched synchronization metadata is defined as $\mathcal{S}_i = \{[t_{start}, t_{end}]_m\}_{m \in \mathcal{M}_i}$, where $\mathcal{M}_i$ denotes the segment subset associated with identity i. The pipeline outputs a structured subject set $\mathcal{P} = \{(\mathrm{ID}_i, B_i, \mathcal{K}_i, \mathcal{S}_i)\}_{i=1}^{N}$ as comprehensive spatiotemporal constraints. This high-precision alignment provides a reliable physical basis for training audio-visual joint generation models and supplies clean cross-modal supervision for downstream speech-to-video generation tasks.

3.4 Hierarchical Multi-modal Caption Generation

Hierarchical Feature Deconstruction and Semantic Fusion. We employ Qwen3-Omni [50] as the inference core to implement a two-stage generation strategy that covers global audio-visual context and detailed subject-level attributes. In stage one, the MLLM captures global video context—video type, shot scale, camera motion, background, lighting and extracts fine-grained attribute sets $\mathcal{X}_i$ (appearance, motion, expression) for each subject ID_i. Audio analysis covers speech, music, and sound effects: **1)** speech transcription w_m with associated subject ID, emotion, and timestamps, each speech metadata is represented as $\mathcal{A}_i = \{([t_{start}, t_{end}]_m, w_m, \mathrm{ID}_m, \mathrm{emotion}_m)\}_{m \in \mathcal{M}_i}$; **2)** music attributes (type, mood, relative volume); **3)** sound effect categories with textual descriptions. We enforce strict mutual exclusivity between speech and music to avoid overlapping supervision signals and maintain disentangled audio representations. Additionally, we apply rigorous post-processing to deduplicate repeated speech and lyric content.

In stage two, the model synthesizes fragmented attributes into a coherent long-form narrative caption $\mathcal{C}_i$. To reduce hallucination, we introduce a placeholder mechanism. Instead of generating exact dialogue or lyrics directly, the model inserts fixed-position anchors such as $[speech\_m]$ and $[lyrics\_m]$. These anchors are later replaced with the corresponding transcribed content from stage one, preserving narrative fluency while ensuring strict consistency with

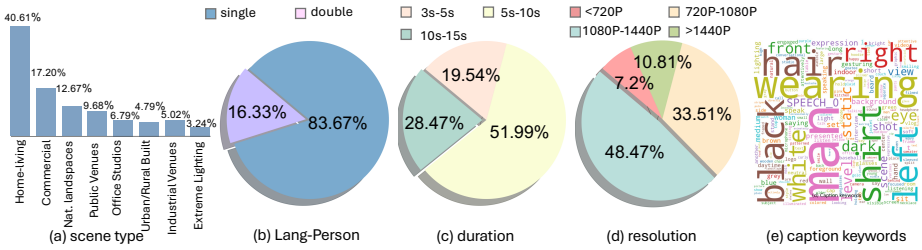

Fig. 3: Statistical analysis of the OmniHuman dataset composition.

the source video on critical information such as dialogue and lyrics. To mitigate MLLM hallucinations and labeling errors, we introduce multiple validation checks. The number of structured subject labels must match the tracking module's output (Sec. 3.2), and the speaker count and transcribed content must remain within an acceptable edit-distance margin from upstream ASR transcripts. A video is retained only if it passes all checks.

Identity-Aware Multi-subject Alignment and Insertion. To resolve identity ambiguity in multi-person scenes, we introduce an identity alignment mechanism based on reference face images. Based on whole-body tracking results in Section 3.2, we select, for each subject, the frame with the highest average confidence over facial keypoints as the reference face. During stage one, the system processes the video together with fixed reference faces $\{\mathrm{REF}_1, \mathrm{REF}_2\}$ and enforces explicit identity verification. Each subject i is assigned an identity label $\mathrm{ID}_i \in \{\mathrm{REF}_1, \mathrm{REF}_2\}$, establishing robust feature-level correspondence. In stage two semantic fusion, we adopt a two-step referential insertion strategy. The system first generates the raw caption $\mathcal{C}_i$. It then inserts identity anchors $\langle \mathrm{REF}_i \rangle$ via suffix tagging after each noun or pronoun referring to that subject. Ultimately, the pipeline outputs a structured subject set $\mathcal{P} = \{(\mathrm{ID}_i, B_i, \mathcal{K}_i, \mathcal{X}_i, \mathcal{A}_i, \mathcal{C}_i)\}_{i=1}^{N}$ as comprehensive spatiotemporal and semantic constraints.

3.5 OmniHuman Statistics and Analysis

As shown in Fig. 3, our dataset comprises 1 million videos totaling 1,800 hours, encompassing 80,000 distinct identities. The primary language is English, accounting for up to 80%. It covers eight typical shooting scenarios: home living spaces, business offices and creative studios, natural environments and landscapes, urban and rural built landscapes, public performance and sports venues, commercial and retail spaces, and industrial and work areas, with detailed distributions illustrated in Fig. 3(a). Additionally, the dataset includes eight video content types: talk shows and interviews, educational, food and cooking, sports and fitness, music and performance, gaming, product reviews, and movies and TV. Based on the number of visual subjects, our dataset is further categorized into two classes, single-person and two-person, with their proportions shown in Fig. 3(b). The pie charts in Fig. 3(c) and (d) illustrate the distributions of video

resolution and duration, respectively. Technically, all videos are of high-definition quality or higher, providing a robust visual foundation for model training.

4 OmniHuman Benchmark (OHBench)

Data Collection of OHBench. Our evaluation set samples from OmniHuman with domain gaps relative to the training set; its distribution is shown in Fig. 4. At the global level, we cover eight real-life scenarios across various shot scales (extreme long to extreme close-up). At the relational level, we include single and two-person videos at a fixed ratio, with a particular focus on challenging person-object interaction scenarios including tool use and physical manipulation. In total, the benchmark contains 331 single-person, 128 two-person, and 20 person-object interaction videos.
Supported Tasks. The OmniHuman Benchmark offers broad task applicability, fully supporting various human-centric video generation tasks, including audio-video joint generation, speech to video, video dubbing, controllable human video editing and downstream speech generation tasks.
Overview of Metrics. OHBench is organized into three progressive levels with seven dimensions. At the global level, we assess video quality, audio quality, and multi-modal alignment. At the relational level, we evaluate person-person and person-object interaction. At the individual level, we measure subject-video and subject-audio fidelity. These dimensions enable comprehensive diagnosis of fine-grained human-centric scene modeling. We find existing metrics (e.g., VBench's aesthetic score, identity consistency, and SyncFormer [21] for audio-visual synchronization) often misalign with human perception. To address this, we construct a new evaluation suite by curating perception-aligned existing metrics and designing novel ones for previously uncovered dimensions. We provide evidence in the supplementary material demonstrating that our metrics better align with human judgment. Our evaluation integrates expert model-based and MLLM-based assessments; all prompts are provided in the supplementary material.

4.1 Foundational Synthesis Quality

Video Quality. We evaluate video quality through two complementary aspects: foundational quality and background plausibility. For foundational quality, we adopt metrics from VBench aligned with human perception, focusing on imaging quality (IQ) and dynamic degree (DD). Specifically, we use MUSIQ [22] to predict per-frame imaging quality, capturing distortions like overexposure, noise, and blur, and RAFT [39] optical flow to assess motion intensity. For background plausibility, we leverage Gemini-3 to perform semantic analysis and physical plausibility judgment on video frames, assessing temporal realism—particularly temporal stability in static shots and parallax plausibility under dynamic camera movements. The model outputs a score from 1 to 10.
Audio Foundation Quality. We compute the aesthetic score (AbS) as the average of four Audiobox [40] dimensions: content enjoyment, content usefulness,

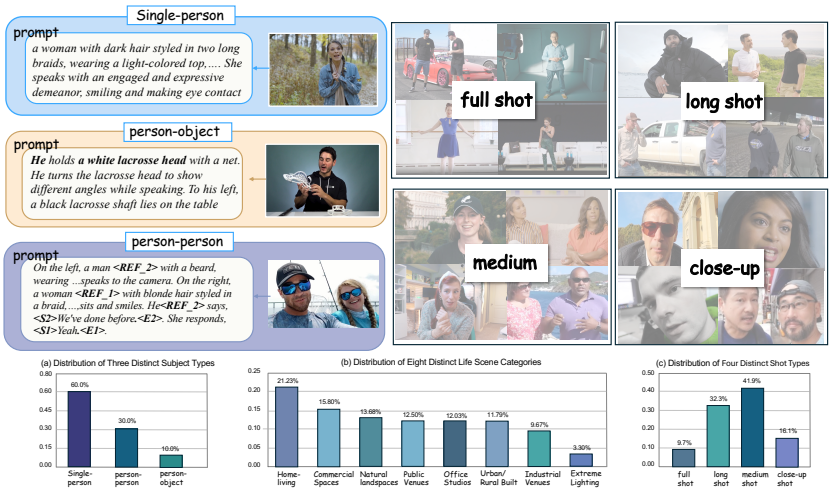

Fig. 4: Distribution of subject categories, scene types, and shot types in OHBench.

production complexity, and production quality. Additionally, we introduce KL divergence and Fréchet Distance (FD) to measure the distributional and perceptual similarity between generated and real audio, respectively.

Audio-Visual Synchronization and Semantic Consistency. To evaluate audio-visual synchronization (A-V) in complex environments, we use ImageBind [47] to measure temporal offsets between visual and auditory events, assessing synchronization accuracy between action sounds and background. Additionally, we employ CLAP [48] to compute audio-text semantic similarity (T-A), quantifying alignment between generated sound effects and corresponding text.

4.2 Social-Physical Interaction Realism

Person-Person Interaction Quality. For two-person interaction scenarios, we leverage the multi-modal large model Gemini-3-pro to evaluate interaction capabilities across three dimensions, each rated on a scale of 1 to 10:: interaction naturalness (IN) assesses social plausibility via eye contact, body gestures, and action responses (e.g., nodding, gaze following); listener realism (LR) evaluates both the accuracy of mapping audio signals to the corresponding speaker (multi-agent audio-visual assignment) and whether the silent listener maintains a natural expression without producing lip movements.; emotion similarity (ES) measures whether the emotional expressions of interacting parties are coordinated and align with semantic context and social expectations.

Dual-person ID Consistency. We employ ArcFace [10] to extract identity embeddings for both individuals and compute the cosine similarity between these embeddings and their ground-truth counterparts. The average similarity score (IC^*) quantifies identity drift in dual-person scenarios.

Table 2: Evaluation on I2AV task in global level metrics.

Models	Video Quality			Audio Quality			Multi-modal Align	
	IQ↑	DD↑	BP↑	AbS↑	KL↓	FD↓	T-A↑	V-A↑
Closed-source model								
Veo3	**0.737**	0.878	**8.74**	5.40	**0.79**	0.87	0.39	**0.37**
kling2.6	0.722	0.686	8.21	4.05	1.28	1.02	0.36	0.26
Wan2.5	0.707	0.721	8.45	4.42	1.04	0.87	0.33	0.21
Sora2	0.708	0.538	8.52	3.34	0.87	0.87	0.32	0.30
SeedDance1.5-pro	0.726	**0.897**	8.65	**5.63**	0.84	**0.83**	0.36	0.31
Open-source model								
Universe-1	0.681	0.390	7.19	2.51	1.18	1.17	0.16	0.19
Uniavgen	0.709	0.575	8.24	4.21	0.84	0.89	0.37	0.19
Ovi	0.701	0.691	8.58	3.75	0.97	0.90	**0.40**	0.23
MOVA	0.695	0.564	8.33	4.27	0.96	1.01	0.20	0.29
LTX-2	0.720	0.601	8.73	3.69	1.04	1.06	0.28	0.27
ours(LTX-ft)	0.721 (+0.1%)	0.665 (+10.7%)	8.78 (+0.5%)	4.13 (+11.9%)	0.77 (+25.9%)	0.93 (+12.3%)	0.35 (+25.0%)	0.30 (+11.1%)

Person-Object Interaction Quality. For person-object interaction, we evaluate two core dimensions using Gemini-3 (1–10, higher is better): object consistency (OC) measures the temporal stability of an object's appearance, position, and state, ensuring it remains identifiable and follows physical laws; contact naturalness (CN) assesses the physical plausibility of the human-object interface via spatial accuracy, temporal synchronization, and force interaction realism.

4.3 Subject-Centric Fine-grained Traits

Subject-Video Quality. We evaluate subject-level video attributes via three metrics: id fidelity, subject attribute consistency, and lip sync. For id fidelity, we employ ArcFace [10] to extract facial features and compute similarity between generated and reference faces, ensuring alignment with human perception. Subject attribute consistency leverages Gemini-3 to assess the subject's continuous presence and physical coherence over time, focusing on issues such as disappearance, missing or severed body parts, and abrupt facial or body mutations. Lip sync uses SyncNet [8] to measure synchronization accuracy between lip movements and speech, ensuring precise lip-audio alignment.

Subject-Audio Quality. We assess subject-level audio attributes via two metrics: pronunciation accuracy and speech realism. The former is measured by computing word error rate (WER) using SenseVoice [1] against ground-truth text, quantifying the accuracy and intelligibility of spoken content. speech quality is assessed using the OVRL (Overall Quality) score from DNSMOS [33], a robust non-intrusive metric that evaluates perceptual quality. It effectively captures both the suppression of background noise and the naturalness of the synthesized speech, providing an objective measure aligned with human perception.

Table 3: Evaluation on I2AV task in interaction and individual level metrics.

Models	Person-Person				Person-Object		Subject-Video			Subject-Audio	
	IC*↑	IN↑	LR↑	ES↑	OC↑	CN↑	IC↑	AC↑	Sync↑	SQ↑	WER↓
Closed-source model											
Veo3	0.580	**7.00**	8.84	7.31	7.05	8.12	0.675	**9.77**	**8.32**	2.96	0.24
kling2.6	0.660	5.50	8.13	5.97	**7.76**	**8.32**	0.724	9.06	7.91	2.71	**0.23**
Wan2.5	0.705	6.02	7.97	6.52	7.24	7.86	**0.802**	9.45	6.83	2.59	0.24
Sora2	0.579	6.15	7.91	6.29	7.03	7.24	0.690	9.74	7.47	2.40	0.30
SeedDance1.5-pro	**0.714**	6.59	8.54	6.97	6.56	6.82	0.758	9.62	7.93	**3.02**	0.29
Open-source model											
Universe-1	-	-	-	-	4.72	4.97	0.683	9.41	1.75	2.56	0.76
Uniavgen	-	-	-	-	5.24	5.36	0.599	9.43	4.45	3.01	0.30
Ovi	0.443	5.84	6.35	**7.67**	6.61	6.72	0.588	9.61	6.97	2.96	0.29
MOVA	0.561	6.09	7.81	6.36	5.98	5.37	0.682	9.24	6.18	2.99	0.26
LTX-2	0.562	6.47	**8.94**	6.63	6.95	6.98	0.656	9.51	6.91	2.89	0.29
ours(LTX-ft)	0.589 (+4.8%)	6.53 (+0.9%)	8.92 (-0.1%)	6.76 (+2.1%)	7.12 (+2.4%)	7.32 (+4.9%)	0.696 (+6.1%)	9.64 (+1.4%)	7.11 (+2.9%)	2.98 (+3.1%)	0.27 (+6.9%)

Table 4: Evaluation on S2V task: global quality, relational interaction, and individual subject performance.

Models	Global Level				Relational Level						Individual Level		
	Video Quality			AV-Align	Person-Person				Person-Object		Subject Video		
	IQ↑	DD↑	BP↑	V-A↑	IC*↑	IN↑	LR↑	ES↑	OC↑	CN↑	IC↑	AC↑	Sync↑
Closed-source model													
kling-avatar [12]	0.707	0.517	8.45	0.369	-	-	-	-	**6.52**	6.43	**0.844**	**9.81**	6.89
wan2.2-s2v	**0.710**	0.754	8.73	**0.374**	-	-	-	-	6.41	6.32	0.710	9.62	7.24
Omniavtar	0.703	0.352	8.66	0.369	-	-	-	-	6.17	**6.56**	0.655	9.73	6.88
hunyuanvideo-avatar [5]	0.681	**0.830**	8.40	0.351	-	-	-	-	5.01	5.24	0.642	9.29	5.52
Multitalk	0.700	0.433	8.86	0.362	0.578	5.90	7.64	6.26	6.32	6.49	0.704	9.80	7.59
InfiniteTalk	0.702	0.545	**8.87**	0.369	**0.679**	**6.06**	**7.94**	**6.31**	6.43	6.51	0.775	**9.81**	**7.63**

5 Experiment

5.1 Evaluation details

We evaluated recent open and closed source models on the I2AV task using the seven-dimensional OHBench. The closed-source models include Veo3.1, Wan2.5, Sora2, kling2.6 [23], and SeedDance-1.5-pro [34]; the open-source models include Universe-1 [43], UniAVGen [54], Ovi, LTX-2, and MOVA. For inference, we generated 5-second videos (or the closest available length for models not supporting 5 seconds). All models were evaluated at the resolution closest to 720P.

5.2 Main Results

Overall Performance Landscape. We show comprehensive OHBench results for all evaluated models in Tab. 2 and Tab. 3. Closed-source models generally outperform open-source ones across most metrics. Veo3 and SeedDance1.5-pro lead in overall video quality, interaction rationality, and individual attributes.

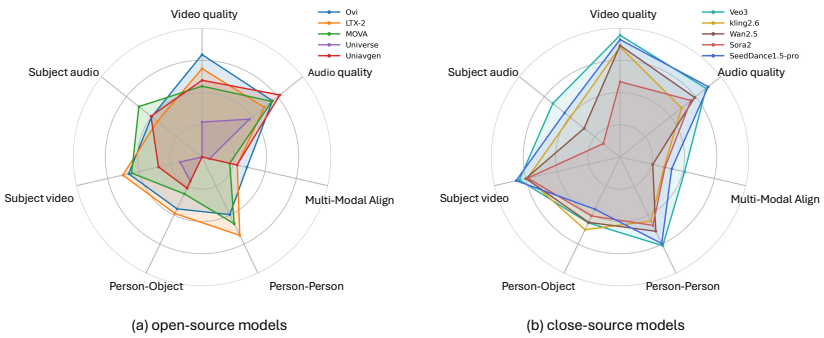

Fig. 5: Performance distribution of 10 models across seven dimensions on OHBench for audio-video joint generation task.

Closd-source Model Results. Veo3 achieves SOTA performance across nearly all metrics, excelling particularly in audio-visual alignment and lip-sync accuracy. SeedDance1.5-pro ranks second overall, with top performance in motion dynamics, audio aesthetics, perceptual similarity, and voice quality. kling2.6 shows a clear advantage in person-object interaction and speech accuracy. Wan2.5 achieves the best identity preservation but lags in speech quality, multi-modal alignment and lip-sync compared to other closed-source models. Sora2 performs relatively poorly among closed-source models, especially in speech quality.

Open-source Model Results. Regarding open-source models, Ovi rivals closed-source models in cross-modal semantic consistency and lip-sync. However, it underperforms in listener realism and occasionally exhibits audio-visual assignment errors (see supplementary material). LTX-2 emerges as the strongest open-source model overall, particularly in dyadic interaction and video quality; notably, its listener realism surpasses all closed-source models. MOVA shows balanced performance across most metrics but struggles with person-object interactions and individual attribute consistency, occasionally producing artifacts like subject disappearance or flickering. UniAVGen achieves competitive results in audio quality and subject attributes but lacks dyadic interaction capability due to insufficient multi-person and person-object training data. Universe-1, as an earlier baseline, lags behind across most evaluation dimensions.

Capabilities and Limitations Across Dimensions. To provide a more intuitive comparison between open-source and closed-source models, we summarize their comprehensive performance across seven dimensions. As illustrated in Fig. 5, existing models perform reasonably in global audio-visual quality, with closed-source models excelling in global video quality. However, individual-level speech and video quality lag behind overall metrics. Moreover, person-person, person-object interaction and multi-modal alignment, remain areas requiring significant improvement.

Closed-source models significantly outperform open-source ones, particularly in video quality and person-object interaction, while also maintaining a marginal

lead in person-person interaction, multi-modal alignment, and audio quality. This overall advantage suggests that closed-source models are not only stronger in low-level visual fidelity, but also more capable of modeling complex human-centric relationships and cross-modal dependencies. Such superiority largely stems from their massive training datasets, which provide sufficient coverage and support across scene-level, interaction-level, and individual-level tasks. Additionally, application-driven optimization encourages these models to incorporate post-training alignment with human preferences, further improving their robustness and perceptual quality in practical scenarios. Nevertheless, even for closed-source models, substantial room for improvement remains, especially in the quality of multi-modal alignment and person-object interaction. These aspects still require more fine-grained temporal understanding, stronger relational reasoning, and better consistency between visual actions, audio cues, and semantic conditions.

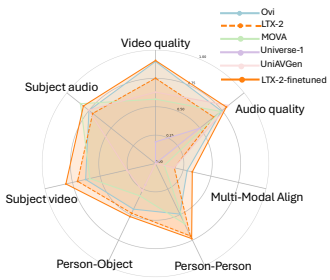

Fig. 6: Performance comparison of LTX-2 before and after finetuning on the OmniHuman data subset.

Generalists vs. Specialists: The Impact of Data Strategies and Backbone Designs on Model Quality. Notably, Fig. 5 reveals that open-source models exhibit more balanced performance, whereas closed-source models show uneven distributions, a phenomenon closely tied to the architectural trade-offs and data-centric strategies of open-source development. Specifically, Ovi demonstrates remarkable results in text-audio (T-A) synchronization within multi-modal alignment, even surpassing all closed-source models in the latter. This advantage is attributed to its large-scale audio pre-training. However, Ovi's dual-person identity consistency and listener realism (LR) are limited by scarce complex interaction samples in its training data. Conversely, LTX-2 demonstrates stunning capabilities in person-person interactions, especially in listner realism, slightly outperforming Veo3. It also maintains a significant lead in overall video quality (IQ). However, its audio quality suffers due to its audio branch's 5B parameters and highly compression Mel-spectrogram latent space, prioritizing speed over acoustic fidelity; training data quality is also be a limiting factor.

Our more Observations. We observe that model performance on human generation significantly degrades as the shot scale transitions from close-up to long shot. Both lip movements during speech and overall body motions become worse, with increased artifacts in the mouth region and facial distortions. This is likely attributable to the inherent difficulty of modeling humans in distant views and the imbalanced distribution of shot types. Additionally, we find that identity distortion rates are higher in two-person scenarios. Detailed case analyses are provided in the supplementary material.

More Application. For the task of speech-driven video generation, we selected appropriate evaluation metrics from OHBench and systematically assessed current state-of-the-art methods. The results are presented in Tab. 4. Overall, In-

finiteTalk achieves the best performance across most metrics. Due to space limitations, additional results and detailed analysis are provided in the supplementary material.

5.3 Fine-tuning with OmniHuman improves quality.

We randomly sampled 180,000 single-person and 20,000 two-person samples from OmniHuman dataset to finetune LTX-2, a representative open-source model. As shown in the last rows of Tab. 2 and Tab. 3, the finetuned model achieves improvements across all evaluation metrics. Notably, audio quality shows significant gains (KL+25.9%, FD+12.3%, AbS+11.9%), and multi-modal alignment is also substantially enhanced (T-A+25.0%, V-A+11.1%). Furthermore, clear improvements are observed in dynamic degree (DD+10.7%), identity consistency (IC+6.1%, IC*+4.8%), and contact naturalness (CN+4.9%).

Fig. 6 illustrates the performance gains from fine-tuning across seven dimensions. The results demonstrate that fine-tuning effectively compensates for LTX-2's original weaknesses in audio quality, multi-modal alignment, and subject-audio attributes, while further amplifying its advantages in subject Video, person object, and person-person dimensions. These results demonstrate that the high-quality audio-visual pairs from OmniHuman allows LTX-2 to transcend its original generation upper bound, establishing it as a leading open-source model.

6 Conclusion

We presented OmniHuman and OHBench to advance human-centric video-audio joint generation. By addressing structural data deficiencies, we introduced a hierarchical framework spanning global, relational, and individual realism. Our work offers a large-scale dataset with video/frame/individual-level annotations via an automated pipeline. Complementarily, OHBench establishes perception-consistent evaluation by filtering out misaligned metrics and incorporating missing dimensions. Experiments confirm OmniHuman boosts performance with minimal data, while our benchmark aligns with human judgment.

Acknowledgements

Chen Li and Jie Chen are the corresponding authors. This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122702), the Shenzhen Medical Research Funds in China (No. B2302037), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), and AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China, and the Shenzhen Loop Area Institute under grant FPF10120250014.

References

1. An, K., Chen, Q., Deng, C., Du, Z., Gao, C., Gao, Z., Gu, Y., He, T., Hu, H., Hu, K., et al.: Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. arXiv preprint arXiv:2407.04051 (2024)
2. An, K., Chen, Y., Chen, Z., Deng, C., Du, Z., Gao, C., Gao, Z., Gong, B., Li, X., Li, Y., Liu, Y., Lv, X., Ji, Y., Jiang, Y., Ma, B., Luo, H., Ni, C., Pan, Z., Peng, Y., Peng, Z., Wang, P., Wang, H., Wang, H., Wang, W., Wang, W., Wu, Y., Tian, B., Tan, Z., Yang, N., Yuan, B., Ye, J., Yu, J., Zhang, Q., Zou, K., Zhao, H., Zhao, S., Zhou, J., Zhu, Y.: Fun-asr technical report (2025), https://arxiv.org/abs/2509.12508
3. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
4. Chen, Y., Zheng, S., Wang, H., Cheng, L., Zhu, T., Huang, R., Deng, C., Chen, Q., Zhang, S., Wang, W., et al.: 3d-speaker-toolkit: An open-source toolkit for multimodal speaker verification and diarization. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
5. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025)
6. Chen, Z., Sun, J., Li, C., Nguyen, T.D., Yao, J., Yi, X., Xie, X., Tan, C., Xie, L.: Mova: Towards generalizable classification of human morals and values. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 33204–33248 (2025)
7. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
8. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian conference on computer vision. pp. 251–263. Springer (2016)
9. Défossez, A., Usunier, N., Bottou, L., Bach, F.: Demucs: Deep extractor for music sources with extra unlabeled data remixed. arXiv preprint arXiv:1909.01174 (2019)
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: 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)
11. Di Chang, Y.S., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Yang, X., Soleymani, M.: Magicdance: Realistic human dance video generation with motions & facial expressions transfer. arXiv preprint arXiv:2311.12052 **2**(3), 4 (2023)
12. Ding, Y., Liu, J., Zhang, W., Wang, Z., Hu, W., Cui, L., Lao, M., Shao, Y., Liu, H., Li, X., et al.: Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. arXiv preprint arXiv:2509.09595 (2025)
13. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
14. Google DeepMind: Veo 3 (5 2025), https://deepmind.google/models/veo/, accessed: 2026-02-17
15. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N.,

Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model (2026), https://arxiv.org/abs/2601.03233
16. Hidayatullah, P., Syakrani, N., Sholahuddin, M.R., Gelar, T., Tubagus, R.: Yolov8 to yolo11: A comprehensive architecture in-depth comparative review. arXiv preprint arXiv:2501.13400 (2025)
17. Hua, D., Wang, X., Zeng, B., Huang, X., Liang, H., Niu, J., Chen, X., Xu, Q., Zhang, W.: Vabench: A comprehensive benchmark for audio-video generation. arXiv preprint arXiv:2512.09299 (2025)
18. Hua, D., Wang, X., Zeng, B., Huang, X., Liang, H., Niu, J., Chen, X., Xu, Q., Zhang, W.: Vabench: A comprehensive benchmark for audio-video generation. arXiv preprint arXiv:2512.09299 (2025)
19. Huang, X., Zhou, H., Yang, Q., Wen, S., Han, K.: Jova: Unified multimodal learning for joint video-audio generation. arXiv preprint arXiv:2512.13677 (2025)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
21. Iashin, V., Xie, W., Rahtu, E., Zisserman, A.: Synchformer: Efficient synchronization from sparse cues. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5325–5329. IEEE (2024)
22. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
23. Kling: Kling. https://klingai.com (2025)
24. Li, H., Huang, J., Jin, P., Song, G., Wu, Q., Chen, J.: Weakly-supervised 3d spatial reasoning for text-based visual question answering. IEEE Transactions on Image Processing **32**, 3367–3382 (2023)
25. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7752–7762 (2025)
26. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision **132**(9), 3463–3483 (2024)
27. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
28. Liu, Y., Zhu, L., Lin, L., Zhu, Y., Zhang, A., Li, Y.: Teaser: Token enhanced spatial modeling for expressions reconstruction. arXiv preprint arXiv:2502.10982 (2025)
29. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation (2025), https://arxiv.org/abs/2510.01284
30. OpenAI: Sora 2: Video generation model (2025), https://openai.com/sora
31. Peng, Z., Liu, J., Zhang, H., Liu, X., Tang, S., Wan, P., Zhang, D., Liu, H., He, J.: Omnisync: Towards universal lip synchronization via diffusion transformers. Advances in Neural Information Processing Systems **38**, 11405–11424 (2026)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

33. Reddy, C.K., Gopal, V., Cutler, R.: Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6493–6497. IEEE (2021)
34. Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025)
35. Soucek, T., Lokoc, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11218–11221 (2024)
36. Sun, M., Wang, W., Qiao, Y., Sun, J., Qin, Z., Guo, L., Zhu, X., Liu, J.: Mm-ldm: Multi-modal latent diffusion model for sounding video generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10853–10861 (2024)
37. Sun, Z., Peng, Z., Ma, Y., Chen, Y., Zhou, Z., Zhou, Z., Zhang, G., Zhang, Y., Zhou, Y., Lu, Q., et al.: Streamavatar: Streaming diffusion models for real-time interactive human avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10887–10897 (2026)
38. Sung-Bin, K., Chae-Yeon, L., Son, G., Hyun-Bin, O., Ju, J., Nam, S., Oh, T.H.: Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. arXiv preprint arXiv:2406.14272 (2024)
39. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
40. Tjandra, A., Wu, Y.C., Guo, B., Hoffman, J., Ellis, B., Vyas, A., Shi, B., Chen, S., Le, M., Zacharov, N., et al.: Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. arXiv preprint arXiv:2502.05139 (2025)
41. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
42. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
43. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025)
44. Wang, J., Qiang, C., Guo, Y., Wang, Y., Zeng, X., Deng, F.: Apollo: Unified multitask audio-video joint generation. arXiv e-prints pp. arXiv–2601 (2026)
45. Wang, K., Deng, S., Shi, J., Hatzinakos, D., Tian, Y.: Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. arXiv preprint arXiv:2406.07686 (2024)
46. Wu, H., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Disentangling aesthetic and technical effects for video quality assessment of user generated content. arXiv preprint arXiv:2211.04894 **2**(5), 6 (2022)
47. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
48. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

49. Xie, T., Lei, W., Huang, G., Zhang, P., Jiang, K., Zhang, C., Ma, F., He, H., Zhang, H., He, J., et al.: Phyavbench: A challenging audio physics-sensitivity benchmark for physically grounded text-to-audio-video generation. arXiv preprint arXiv:2512.23994 (2025)
50. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., Lin, J.: Qwen3-omni technical report (2025), https://arxiv.org/abs/2509.17765
51. Yang, L., Zhao, Z., Zhao, H.: Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 3031–3048 (2025)
52. Yang, S., Kong, Z., Gao, F., Cheng, M., Liu, X., Zhang, Y., Kang, Z., Luo, W., Cai, X., He, R., et al.: Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. arXiv preprint arXiv:2508.14033 (2025)
53. Yang, Z., Hu, Y., Du, Z., Xue, D., Qian, S., Wu, J., Yang, F., Dong, W., Xu, C.: Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. arXiv preprint arXiv:2502.10810 (2025)
54. Zhang, G., Zhou, Z., Hu, T., Peng, Z., Zhang, Y., Chen, Y., Zhou, Y., Lu, Q., Wang, L.: Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. arXiv preprint arXiv:2511.03334 (2025)
55. Zhang, Y., Li, Z., Wang, D., Zhang, J., Zhou, D., Yin, Z., Dai, X., Yu, G., Li, X.: Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. arXiv preprint arXiv:2507.09862 (2025)
56. Zhang, Y., Wang, T., Zhang, X.: Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22056–22065 (2023)
57. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3661–3670 (2021)
58. Zhou, Y.H., Li, H., Lin, R., Huang, H., Zhou, J., Yuan, C., Lan, T., Zhou, Z., Li, Y., Xu, J., et al.: Mtavg-bench: A comprehensive benchmark for evaluating multi-talker dialogue-centric audio-video generation. arXiv preprint arXiv:2602.00607 (2026)
59. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: European conference on computer vision. pp. 650–667. Springer (2022)