

KlingAvatar 2.0 Technical Report

Kling Team, Kuaishou Technology

Avatar video generation models have achieved remarkable progress in recent years. However, prior work exhibits limited efficiency in generating long duration high-resolution videos, suffering from temporal drifting, quality degradation, and a weak prompt follow-up as the video length increases. To address these challenges, we propose **KlingAvatar 2.0**, a spatio-temporal cascade framework that performs upscaling in both spatial resolution and temporal dimension by first generating low-resolution blueprint video keyframes that capture global semantics and motion, and then refining them into high-resolution, temporally coherent sub-clips using a first-last frame strategy, while retaining smooth temporal transitions in long-form videos. To enhance cross-modal instruction fusion and alignment in extended long videos, we introduce a Co-Reasoning Director composed of three modality-specific large language model (LLM) experts. These experts reason about modality priorities and infer the underlying user intent, converting inputs into detailed storylines through multi-turn dialogue. A negative director further refines negative prompts to improve instruction alignment. Building on these components, we extend the framework to support ID-specific multi-character control. Extensive experiments demonstrate that our model effectively addresses the challenges of efficient, multimodally aligned long-form high-resolution video generation, delivering enhanced visual clarity, realistic lip-teeth rendering with accurate lip synchronization, strong identity preservation, and coherent multimodal instruction following.

Date: December 15, 2025

Access: <https://app.klingai.com/global/ai-human/image/new/>

Model ID: Avatar 2.0

1 Introduction

Audio-driven avatar video synthesis aims to generate realistic and expressive human-centric videos, featuring synchronized facial and emotion expressions, coherent lip-teeth movements and body gestures, and natural character interactions with both the environment and other characters. This technology holds significant value across diverse domains, including education, personalized services, industrial training, entertainment, and advertising, where it enables more immersive and engaging experiences.

The field of audio-driven avatar video generation has evolved significantly through several stages of technological advancement. Early approaches focused primarily on lip-synchronized facial animation, leveraging audio-to-motion representations [8, 19, 21, 28, 57, 71] and direct audio-to-video synthesis for portrait animation [29, 41, 50, 55, 64, 65, 75]. Subsequent developments expanded the scope to semi-body video generation, incorporating hand gestures and upper body animation [34, 39, 54], followed by full-body video generation with background environments and character-environment interactions [9, 10, 13, 15–17, 30, 35, 60]. These advances in complex human motion modeling, realistic environment generation, and natural interactions have been largely enabled by pretrained DiT-based video diffusion models [36, 38, 56, 67, 74]. To address more complex real-world scenarios and improve controllability, recent works have explored human-object interactions [27] and multi-person conversational scenarios with per-character audio control [33, 62, 63]. Multimodal large language model (MLLM) driven storyline planning has emerged as a promising direction, enabling fine-grained expressions, vivid emotions, reasoned actions, and environmental interactions through shot-level guidance for long-duration generation [13, 30]. Despite these advances, existing methods remain inefficient for generating long-duration, high-resolution digital human videos, often suffering from visual degradation and limited coherence with



Figure 1 KlingAvatar 2.0 generates vivid, identity-preserving digital humans with accurate camera control, expressive emotions, high-quality motion, and precise facial-lip and audio synchronization. It achieves coherent alignment across audio, image, and text instructions, generalizes to diverse open-domain styles, and supports multi-character synthesis with identity-specific audio control. These capabilities are enabled by our multimodal instruction-following, omni-directed spatial-temporal cascade framework for high-resolution, long-duration video generation.

complex, long-horizon multimodal instructions.

Building upon the foundation of [13], we propose a unified framework that addresses the aforementioned challenges. To enable efficient long-form, high-resolution video generation, we introduce a temporal and spatial cascade framework that samples low-resolution blueprint videos for efficient generation and then gradually upsamples them to longer durations and higher resolutions. This approach enriches visual details while mitigating temporal drifting artifacts that plague long-form video generation. For long-duration video generation, effective multimodal fusion and adherence to user instructions are crucial. When handled poorly, modality conflicts can cause severe degradation. To address this, we develop a Co-Reasoning Director that improves upon existing MLLM reasoning capabilities, operating in a multi-turn dialogue manner to generate coherent shot-level storylines. This is complemented by a novel negative director that captures fine-grained negative prompts to enhance instruction-following accuracy. Moreover, complex digital human applications often involve multiple characters, where how to accurately drive each human becomes an essential problem. To this end, we leverage deep DiT block features and ID-aware attention to realize mask-controlled audio injection, enabling synchronized yet individually controlled character animations in complex conversational settings.

Together, these components form an integrated system that advances the state-of-the-art in audio-driven avatar video synthesis by simultaneously addressing efficiency, instruction alignment, and multi-character coordination.

To develop and train such a model, we curated an enhanced dataset that expands upon [13], featuring a substantially larger collection of high-quality, cinematic-level video data. Our dataset encompasses multilingual and multi-character conversational scenarios, with extensive filtering pipelines applied to ensure high visual fidelity and consistent audio-lip synchronization. We conduct extensive experiments to evaluate the performance of our proposed framework. Our evaluation demonstrates that our model achieves superior performance against leading competitors [1, 13, 30] across visual quality, camera movement, lip synchronization accuracy, fine-grained lip-teeth detail preservation, vivid and natural character animation, and audio-emotion alignment. In terms of generation efficiency, our spatial-temporal cascade framework enables long-duration and high-resolution video synthesis with improved computational efficiency compared to prior methods, while maintaining identity consistency and story continuity for videos up to 5 minutes. We highlight representative generation results in Figure 1. We summarize our contributions as follows:

- **Spatial-Temporal Cascade Framework:** We introduce a spatial-temporal cascade framework that enables efficient generation of long-duration, high-resolution videos through progressive and parallel temporal and spatial upsampling, effectively mitigating temporal drifting artifacts while enriching visual details.
- **Co-Reasoning Director:** We develop a Co-Reasoning Director that coordinates multiple MLLMs and LLMs in a multi-turn dialogue manner to capture details across modalities and resolve modality conflicts, generating coherent shot-level storylines. A complementary negative director further enhances fine-grained negative prompts to improve instruction-following accuracy and character emotion expression.
- **Multi-Character Multi-Audio Control:** We propose a multi-character, multi-audio control mechanism that exploits deep DiT features for character mask prediction, enabling synchronized yet individually controlled animations from multiple audio streams.
- **Strong performance and generalization:** KlingAvatar 2.0 achieves state-of-the-art performance across multiple dimensions, including visual quality, coherent and vivid character animations, natural camera movements, accurate lip synchronization, and precise audio-emotion alignment, while demonstrating strong generalization to diverse domains and scenarios.

2 Related Works

Video Generation. Visual content generation has achieved remarkable breakthroughs through diffusion models, from photorealistic image synthesis to video generation. Early video generation approaches extended pretrained U-Net [45] based image synthesis models [4, 6, 12, 22, 44, 49] by incorporating temporal dimensions or combining 1D temporal attention with 2D spatial attention blocks to capture inter-frame correspondences and reducing computational costs [2, 20, 23, 48, 59]. However, such designs that model all frames without temporal compression face limitations in scalability, along with issues of temporal drifting and visual artifacts. Recent DiT-based image synthesis methods [5, 14, 40] have advanced realistic image generation through scalable training paradigms, enabling high-fidelity appearance and improved instruction following. Building upon these progresses, video diffusion models have shifted focus toward DiT-based architectures [36, 38, 56, 67, 74]. These methods employ 3D convolutional VAEs to compress videos both temporally and spatially into compact tokens, and leverage large transformer models combined with growing training data and computational resources to capture temporal dynamics and visual details, establishing new state-of-the-art results for video generation. Recent extensions of video diffusion models further advance efficient and long-context generation [7, 18, 32, 53, 70, 73], unified multimodal conditioning with cascaded super-resolution for high-resolution synthesis [52], and world modeling [26, 51, 69]. Despite these advances, these methods are primarily designed for general video generation with text or image prompts, lacking audio conditioning, and remain inadequate for speech-driven digital human modeling.

Multimodal Avatar Synthesis. Multimodal avatar synthesis has achieved significant progress through rapid development. The field has evolved from early non-audio driven methods [24, 42, 46, 47, 61, 72] that transfer motion from reference videos or landmark sequences, as well as 3D-based digital human synthesis approaches [3, 11, 25], to modern audio-driven video generation systems. Among audio-driven methods, some

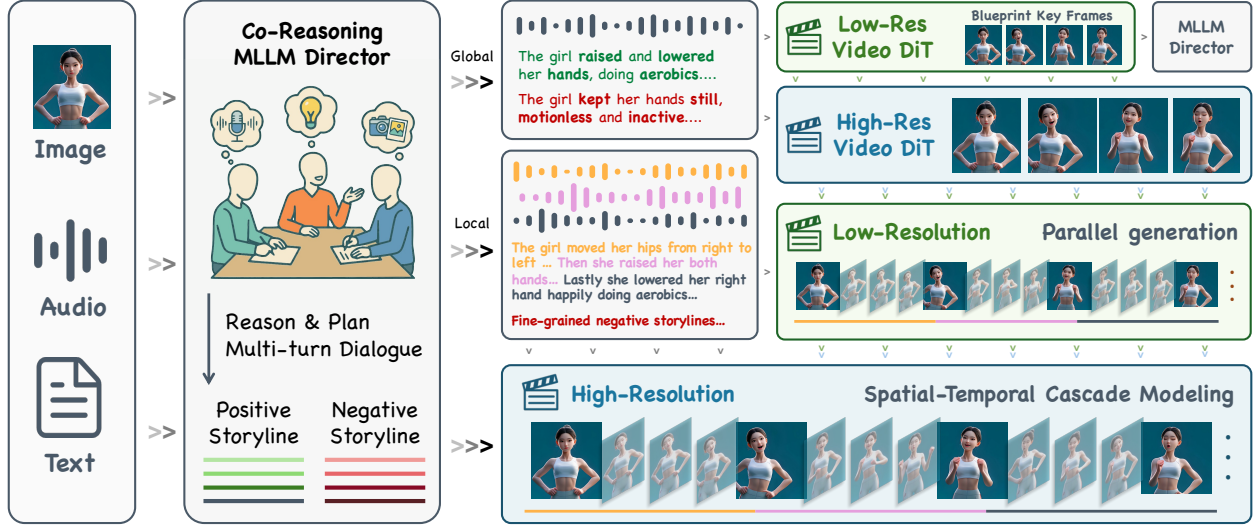


Figure 2 Overview of the KlingAvatar 2.0 framework. Given multimodal instructions, the Co-Reasoning Director reasons and plans hierarchical, fine-grained positive and negative storylines in a multi-turn dialogue manner, and the spatio-temporal cascade pipeline generates coherent, long-form, high-resolution avatar videos in parallel.

employ explicit facial landmarks derived from audio features for precise control [8, 28, 57], while others learn implicit motion representations from audio to enable more flexible animation [19, 21, 71]. More recently, transformer-based audio-driven approaches utilize cross-attention mechanisms to eliminate intermediate motion representations, directly generating talking avatars from audio using diffusion models with end-to-end synthesis [29, 41, 50, 55, 64, 65, 75]. Beyond facial lip synchronization, semi-body methods synchronize hand gestures and upper body movements with audio [34, 39, 54], enabling more expressive digital human generation. Recent advances leverage large-scale pretrained video diffusion models to achieve enhanced temporal coherence and visual fidelity [9, 10, 15–17, 35, 60], supporting the generation of complex background environments and full body interactions. Further developments include specialized extensions for human-object interactions [27] and multi-person conversational scenarios with per-character audio control [33, 62, 63], further increasing the applicability and realism of avatar synthesis. Most recently, MLLM-based multimodal planning enables fine-grained expression, vivid emotions, reasoned actions, and interaction with the environment through shot-level guidance for long-duration generation [13, 30].

3 Method

KlingAvatar 2.0 extends the Kling-Avatar [13] pipeline, as illustrated in Fig. 2, given the reference image, input audios, and textual instructions, the system efficiently generates high-fidelity, long-form digital human videos with accurate lip synchronization and fine-grained control over multiple speakers and roles. In the following, we detail the spatial-temporal cascade diffusion framework (Sec. 3.1), the co-reasoning multimodal storyline director (Sec. 3.2), the multi-character control module (Sec. 3.3), and the acceleration techniques (Sec. 3.4).

3.1 Spatial-Temporal Cascade Modeling

To support long-duration, high-resolution avatar synthesis with efficient computation, KlingAvatar 2.0 adopts a spatial-temporal cascade of audio-driven DiTs built on top of pretrained video diffusion models, as illustrated in Fig. 2. The pipeline comprises two nested cascades that jointly handle global storyline planning over long horizons and local spatio-temporal refinement. First, a low-resolution diffusion model generates a blueprint video that captures global dynamics, content, and layout; representative low-resolution keyframes are then upscaled by a high-resolution DiT, enriching fine details while preserving identity and scene composition under the same Co-Reasoning Director’s global prompts. Next, a low-resolution video diffusion model expands these high-resolution anchor keyframes into audio-synchronized sub-clips via first-last-frame conditioned

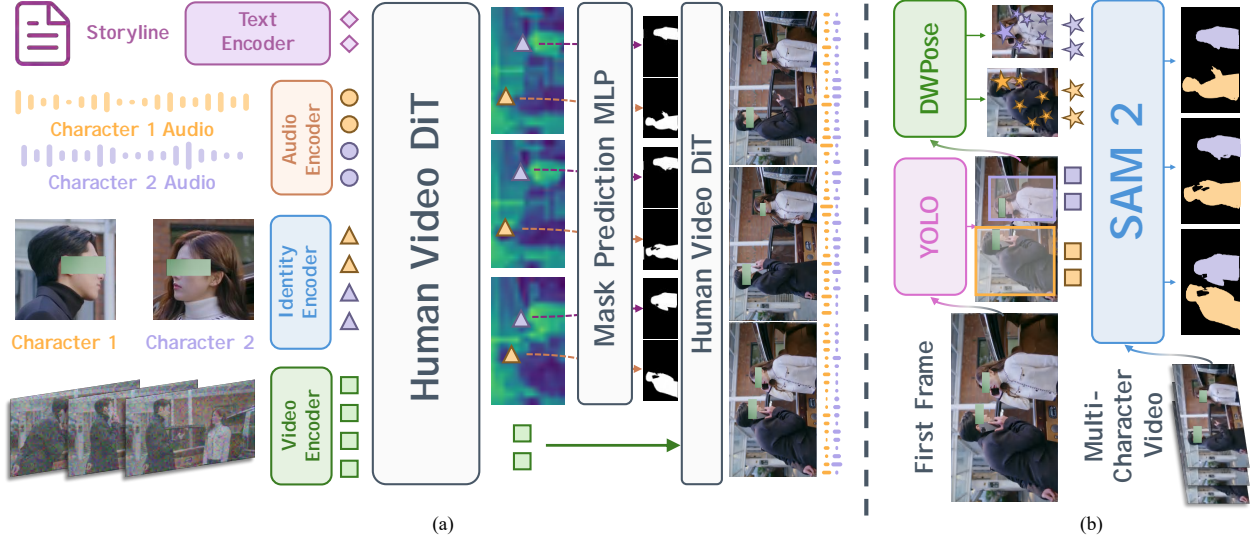


Figure 3 (a) Multi-character video generation pipeline with identity-specific audio control. A mask-prediction head is attached to deep DiT features, and the predicted masks gate ID-specific audio injection into corresponding regions. (b) Automated multi-character video annotation pipeline.

generation, where the prompts are augmented with the blueprint keyframes to refine fine-grained motion and expression. An audio-aware interpolation strategy synthesizes transition frames to enhance temporal connectivity, lip synchronization, and spatial consistency. Finally, a high-resolution video diffusion model performs super-resolution on the low-resolution sub-clips, producing high-fidelity, temporally coherent video segments.

3.2 Co-Reasoning Director

KlingAvatar 2.0 employs a Co-Reasoning Director that jointly reasons over audio, images, and text in a multi-turn dialogue manner, building on recent MLLM-based avatar planners [13, 30]. The Director is instantiated with three experts: (i) an audio-centric expert performs transcription and paralinguistic analysis (emotion, prosody, speaking intent); (ii) a visual expert summarizes appearance, layout, and scene context from reference images; and (iii) a textual expert interprets user instructions, incorporates conversational history from the other experts, and synthesizes a logically coherent storyline plan. These experts engage in several rounds of co-reasoning with chain-of-thought, exposing intermediate thoughts to resolve conflicts (e.g., an angry vocal tone paired with a neutral script) and to fill in underspecified details such as implied actions or camera movements. The director outputs a structured storyline that decomposes the video into a sequence of shots. Additionally, we also introduce a negative director, where positive prompts emphasize desired visual and behavioral attributes, and negative prompts explicitly down-weight implausible poses, artifacts, and fine-grained opposite emotions (e.g., sad vs. happy) or motion styles (e.g., overly fast vs. slow).

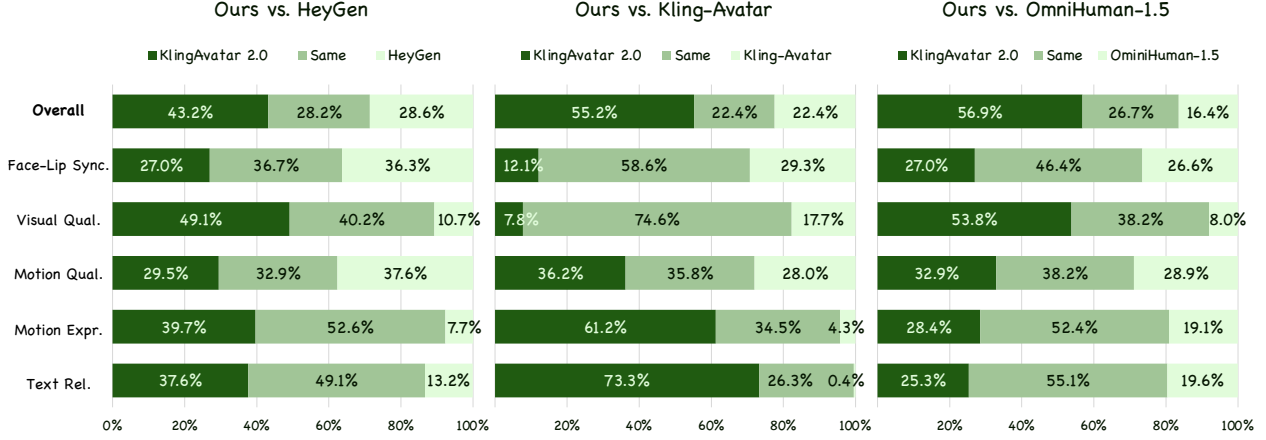
For long videos, the director further refines the global storyline into segment-level plans aligned with the audio timeline, which directly parameterize the keyframe cascade and clip-level refinement modules. This high-level multimodal planning converts loosely specified instructions into a coherent script that can be consistently followed by the diffusion backbone, substantially improving semantic alignment and temporal coherence.

3.3 Multi-Character Control

KlingAvatar 2.0 generalizes the single-speaker avatar setting to multi-character scenes and identity-specific audio control. Our design follows the character-aware audio-injection paradigm used in recent multi-person conversational avatars [33, 62, 63]. Empirically, we observe an important architectural property: hidden features at different depths of the DiT blocks exhibit distinct feature representations. In particular, latent representations in deep DiT layers are organized into semantically coherent spatial regions with reduced noise, and these regions align well with individual characters and other salient objects.

Table 1 Quantitative results of GSB metrics between our approach and other competitors across diverse criteria.

GSB	Overall	Face-Lip Sync.	Visual Qual.	Motion Qual.	Motion Expr.	Text Rel.
Ours vs. HeyGen	1.26	0.86	1.76	0.88	1.53	1.39
Ours vs. Kling-Avatar	1.73	0.80	0.89	1.13	2.47	3.73
Ours vs. OmniHuman-1.5	1.94	1.02	1.99	1.06	1.13	1.08


Figure 4 Visualization of GSB benchmark results comparing KlingAvatar 2.0 with HeyGen, Kling-Avatar, and OmniHuman-1.5 across various evaluation criteria.

Motivated by this observation, we attach a mask-prediction head to selected deep DiT blocks, as shown in Fig. 3(a). Concretely, given a specified character in the first frame, we encode the reference identity crops using the same patchification scheme without adding noise to reference tokens. We then compute cross-attention between deep video latent tokens and these reference tokens for each identity, and apply MLP modules to regress per-frame character masks. Ground-truth (GT) masks are downsampled to match the spatial and temporal resolution of the intermediate latent features. During training, the DiT video backbone is frozen and only the mask-prediction modules are optimized. During denoising, the predicted masks are used to gate the identity-specific audio stream injection to corresponding regions.

To facilitate curation of a large-scale multi-character dataset for training, we expand our data sources to include podcasts, interviews, multi-character television series and more. To collect GT character masks at scale, we developed an automated annotation pipeline that produces per-character video masks, as illustrated in Fig. 3(b). The pipeline leverages several expert models: YOLO [31] for person detection, DWPose [66] for keypoint estimation, and SAM2 [43] for segmentation and temporal tracking. Specifically, we detect all characters in the first frame with YOLO, estimate keypoints within each detection using DWPose, and use the resulting bounding boxes and keypoints as prompts for SAM2 to segment and track each person in subsequent frames. Finally, we validate the generated video masks against per-frame YOLO and DWPose estimation results and filter out misaligned or low-overlap segments to ensure high-quality annotations for training.

3.4 Accelerated Video Generation

To achieve accelerated inference efficiency, we explored distillation schemes based on trajectory-preserving distillation exemplified by PCM [58] and DCM [37], and distribution matching distillation exemplified by DMD [68]. Based on comprehensive evaluations of experimental cost, training stability, inference flexibility, and final generative performance metrics, we ultimately selected the trajectory-preserving distillation approach. To further enhance distillation efficiency, we developed customized time schedulers by analyzing the performance of the base model across different timesteps, thereby balancing the inference speedup ratio against model performance. Within our distillation algorithm, we introduced a multi-task distillation paradigm through a series of precisely designed configurations. This paradigm not only yields a synergistic effect ($1+1>2$), improving the distillation outcomes for each individual task.

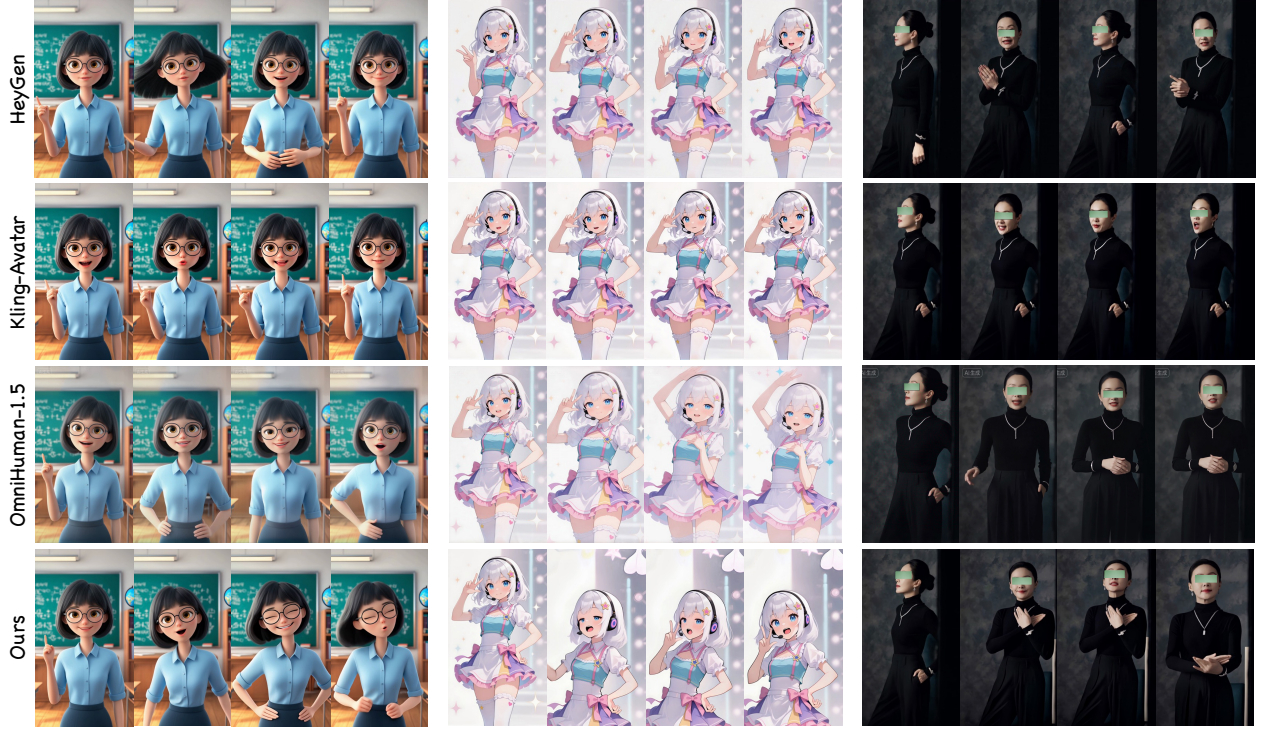


Figure 5 Qualitative comparison between KlingAvatar 2.0 and baseline methods. **Left:** Our method produces more natural hair dynamics and vivid facial expressions. **Middle:** Our results adhere more closely to the specified bottom-to-top camera motion. **Right:** Our generated video aligns better with the prompt “...turned to the front and folded her hands in front of her chest”.

4 Experiments

4.1 Experimental Setup

We follow the human preference-based subjective evaluation protocol in [13] to conduct comprehensive evaluations of KlingAvatar 2.0. We construct 300 high-quality test cases, each consisting of paired image, audio, and text prompts, including 100 Chinese speech, 100 English speech, and 100 singing samples. For each case, annotators perform Good/Same/Bad (GSB) pairwise comparisons between our results and those of baseline methods. We report $(G+S)/(B+S)$ as the main metric, where higher scores indicate stronger human preference. To capture fine-grained aspects of video quality and multimodal alignment, we further extend GSB to a richer set of detailed criteria, including:

- **Face-Lip Synchronization.** Measures temporal alignment between face and lip motions and speech, the naturalness and continuity of facial expressions, and the consistency between movements, facial dynamics, and phonetic content.
- **Visual Quality.** Evaluates overall aesthetics, sharpness, and fidelity of fine details such as hair, teeth, and skin texture, as well as temporal consistency of appearance and robustness to artifacts, flickering, and implausible lighting or color.
- **Motion Quality.** Assesses the plausibility, smoothness, and temporal coherence of body, head, and camera motion, avoiding geometric distortions, jitter, unnatural warping, body-part collapses, and unstable character tracking.
- **Motion Expressiveness.** Characterizes the richness, diversity, and vividness of lip, facial, and full-body movements, and their emotional match with the audio and text in terms of intensity, timing, and modulation of gestures and head poses.
- **Text Relevance.** Reviews the alignment and coherence of the generated storyline, camera trajectories,

and scene dynamics with the textual instructions.

4.2 Experimental Results

We compare KlingAvatar 2.0 against three strong baselines: HeyGen [1], Kling-Avatar [13], and OmniHuman-1.5 [30]. Quantitative GSB results across diverse dimensions are summarized in Table 1 and visualized in Fig. 4. Our method achieves strong overall performance, with especially notable improvements in motion expressiveness and text relevance. Qualitative comparisons are provided in Fig. 5.

Our model generates richer, more natural dynamic effects and follows multimodal instructions more faithfully, demonstrating an enhanced understanding of complex audiovisual intent and effectively addressing the key limitations of baseline methods. Across baselines, hair dynamics are either relatively rigid (Kling-Avatar, OmniHuman-1.5) or occasionally less physically grounded (HeyGen). Our method produces more temporally consistent and physically plausible hair motion and head poses, leading to improved perceived naturalness. For multimodal instruction following, HeyGen and Kling-Avatar generally generate stable but relatively simple camera trajectories, while OmniHuman-1.5 sometimes deviates from the specified camera instructions. KlingAvatar 2.0 produces camera motions and scene dynamics that are more closely aligned with the textual prompts, yielding detailed and coherent interpretations. Regarding emotional expression and fine-grained motion instructions, HeyGen and Kling-Avatar sometimes under-emphasize the target actions, whereas OmniHuman-1.5 incorrectly folds the hands at the waist instead of in front of the chest. In contrast, our approach more reliably captures the intended motion of folding the hands in front of the chest, produces movements synchronized with the audio and target emotion, and yields facial and body expressions that are both expressive and realistic.

Fig. 6 showcases results generated by our framework across diverse scenarios. Powered by the spatial-temporal cascade and the multimodal co-reasoning director, our approach accurately interprets and fuses image, audio, and text instructions, producing emotionally expressive characters, coherent full-body and camera motions, and precise, fine-grained lip synchronization. Beyond single-speaker talking scenarios, our method generalizes well to multi-person interactions with per-character audio conditioning. These results demonstrate the robustness and versatility of KlingAvatar 2.0 in open, complex settings.

As shown in Fig. 7, prior work typically uses a small, fixed set of generic negative prompts (e.g., “artifacts, bad quality, blur”) for an entire video, offering only coarse control over undesired content. In contrast, our negative director employs detailed, shot-specific negative prompts that track the evolving storyline and target emotion. This per-shot control discourages implausible expressions, unstable motion, and narrative-inconsistent artifacts, leading to more natural, emotionally faithful, and temporally stable results that better follow the text description.

5 Conclusion

In this paper, we present KlingAvatar 2.0, a unified framework that enables spatio-temporal cascade generation for high-resolution, long-duration, lifelike multi-person avatar videos with omni-directed co-reasoning directors. Our multimodal, multi-expert co-reasoning director thinks and plans over audio cues, visual contexts, and complex instructions through multi-turn dialogues to resolve ambiguities and conflicting signals, producing coherent global storylines to guide the long-form synthesis trajectory and detailed local prompts to refine sub-clip dynamics. The hierarchical storyline drives generation of low-resolution blueprint keyframes, and spatio-temporal upscaled high-resolution, audio-synchronized sub-clips, which are efficiently composed into long-form videos in parallel via first-last frame conditioning. We further extend the application scenarios to multi-character settings with identity-specific audio control and develop an automated annotation pipeline to curate large-scale multi-person video datasets. Experiments demonstrate that KlingAvatar 2.0 delivers leading performance in visual fidelity, identity preserving, lip-audio synchronization, instruction-following, long-duration coherence, and multi-character, multi-audio controllability. We believe our exploration of an omni-directed, multi-character, multi-audio, long-form, high-resolution avatar synthesis framework paves the way for future research and applications in digital human generation.

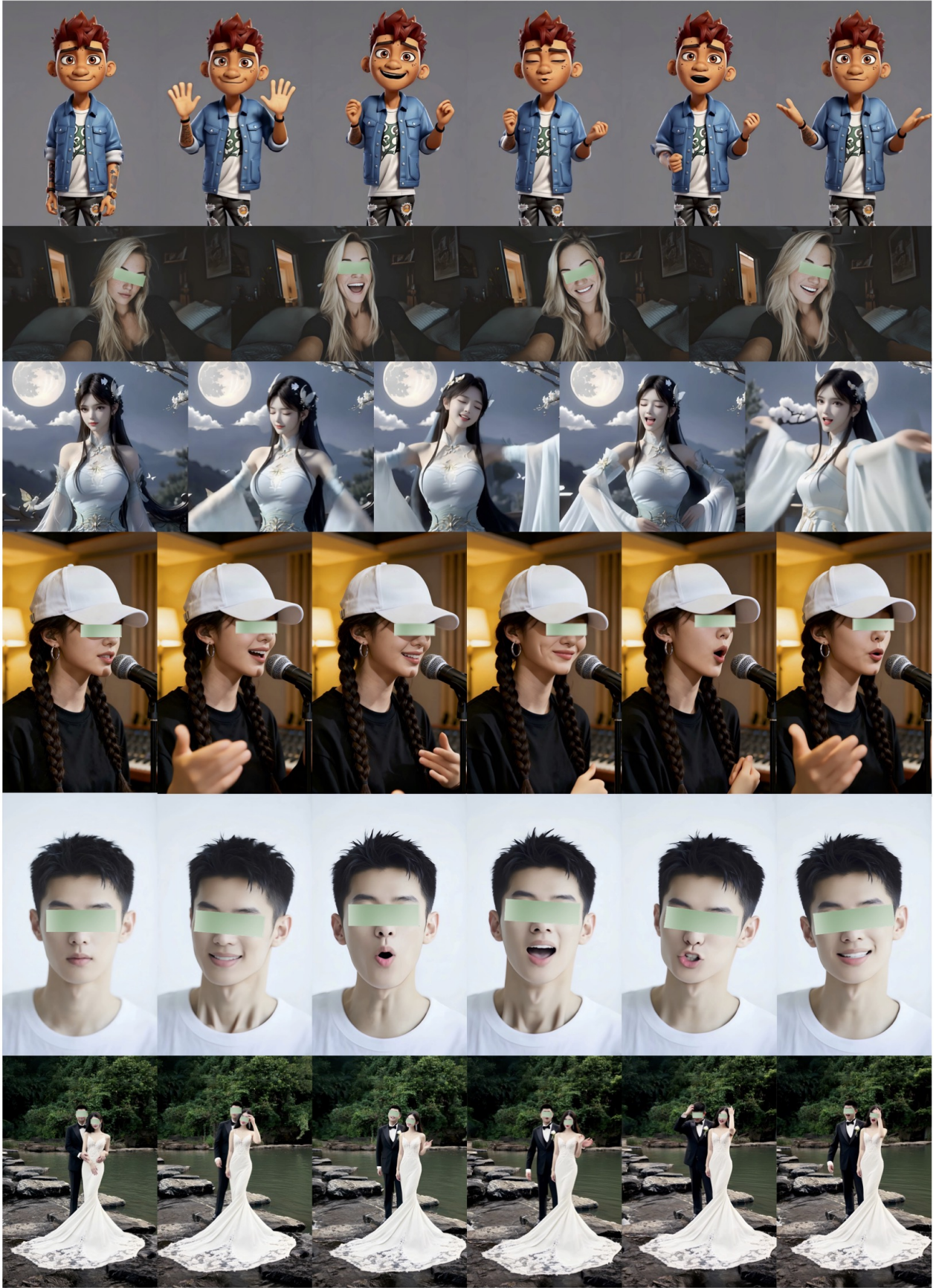


Figure 6 Representative qualitative results generated by our spatial-temporal cascade framework with the multimodal co-reasoning director.



Figure 7 Ablation study of the negative director on blueprint keyframes. The negative director enhances facial expressiveness, improves temporal stability and emotional controllability, and reduces lighting and exposure artifacts.

6 Contributors

All contributors are listed in alphabetical order by their last names.

Jialu Chen, Yikang Ding, Zhixue Fang, Kun Gai, Yuan Gao, Kang He, Jingyun Hua, Boyuan Jiang, Mingming Lao, Xiaohan Li, Hui Liu, Jiwen Liu, Xiaoqiang Liu*, Yuan Liu, Shun Lu, Yongsen Mao, Yingchao Shao, Huafeng Shi, Xiaoyu Shi, Peiqin Sun, Songlin Tang, Pengfei Wan, Chao Wang, Xuebo Wang, Haoxian Zhang, Yuanxing Zhang, Yan Zhou.

*Project Lead

References

- [1] HeyGen. <https://www.heygen.com/>.
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [3] Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Pengfei Wan, Di ZHANG, Shuai Li, et al. Cafe-talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\{\alpha\}$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.
- [5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\{\sigma\}$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.
- [6] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart- $\{\delta\}$: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024.
- [7] Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Xiaoqiang Liu, and Pengfei Wan. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation. *arXiv preprint arXiv:2508.19320*, 2025.
- [8] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *AAAI*, pages 2403–2410, 2025.
- [9] 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 diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024.
- [10] Jiahao Cui, Yan Chen, Mingwang Xu, Hanlin Shang, Yuxuan Chen, Yun Zhan, Zilong Dong, Yao Yao, Jingdong Wang, and Siyu Zhu. Hallo4: High-fidelity dynamic portrait animation via direct preference optimization and temporal motion modulation. *arXiv preprint arXiv:2505.23525*, 2025.
- [11] Liyuan Cui, Xiaogang Xu, Wenqi Dong, Zesong Yang, Hujun Bao, and Zhaopeng Cui. Cfsynthesis: Controllable and free-view 3d human video synthesis. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 135–144, 2025.
- [12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021.
- [13] Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, Ming Chen, Xiaoqiang Liu, Yu-shen Liu, and Wan Pengfei. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025.
- [14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.
- [15] Zhengcong Fei, Hao Jiang, Di Qiu, Baoxuan Gu, Youqiang Zhang, Jiahua Wang, Jialin Bai, Debang Li, Mingyuan Fan, Guibin Chen, et al. Skyreels-audio: Omni audio-conditioned talking portraits in video diffusion transformers. *arXiv preprint arXiv:2506.00830*, 2025.
- [16] Qijun Gan, Ruizhi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniaavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025.
- [17] Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.
- [18] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

- [19] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.
- [20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.
- [21] Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023.
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020.
- [23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, volume 35, pages 8633–8646, 2022.
- [24] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.
- [25] Wentao Hu, Shunkai Li, Ziqiao Peng, Haoxian Zhang, Fan Shi, Xiaoqiang Liu, Pengfei Wan, Di Zhang, and Hui Tian. Ggtalker: Talking head synthesis with generalizable gaussian priors and identity-specific adaptation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.
- [26] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv: 2505.14357*, 2025.
- [27] Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, et al. Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. *arXiv preprint arXiv:2506.08797*, 2025.
- [28] Jianwen Jiang, Gaojie Lin, Zhengkun Rong, Chao Liang, Yongming Zhu, Jiaqi Yang, and Tianyun Zhong. Mobileportrait: Real-time one-shot neural head avatars on mobile devices. *arXiv preprint arXiv:2407.05712*, 2024.
- [29] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. In *The Thirteenth International Conference on Learning Representations*, 2025. <https://openreview.net/forum?id=weM4YBicIP>.
- [30] Jianwen Jiang, Weihong Zeng, Zerong Zheng, Jiaqi Yang, Chao Liang, Wang Liao, Han Liang, Yuan Zhang, and Mingyuan Gao. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint arXiv:2508.19209*, 2025.
- [31] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. <https://github.com/ultralytics/ultralytics>.
- [32] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [33] 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. *arXiv preprint arXiv:2505.22647*, 2025.
- [34] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, Zerong Zheng, and Yanbo Zheng. Cyberhost: A one-stage diffusion framework for audio-driven talking body generation. In *The Thirteenth International Conference on Learning Representations*, 2025. <https://openreview.net/forum?id=vaEPihQsAA>.
- [35] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025.
- [36] Dongyang Liu, Shicheng Li, Yutong Liu, Zhen Li, Kai Wang, Xinyue Li, Qi Qin, Yufei Liu, Yi Xin, Zhongyu Li, Bin Fu, Chenyang Si, Yuewen Cao, Conghui He, Ziwei Liu, Yu Qiao, Qibin Hou, Hongsheng Li, and Peng Gao. Lumina-video: Efficient and flexible video generation with multi-scale next-dit. *arXiv preprint arXiv:2502.06782*, 2025.
- [37] Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K. Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. <https://arxiv.org/abs/2506.03123>, 2025.

- [38] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.
- [39] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *CVPR*, pages 5489–5498, 2025.
- [40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.
- [41] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025.
- [42] Di Qiu, Zhengcong Fei, Rui Wang, Jialin Bai, Changqian Yu, Mingyuan Fan, Guibin Chen, and Xiang Wen. Skyreels-a1: Expressive portrait animation in video diffusion transformers. *arXiv preprint arXiv:2502.10841*, 2025.
- [43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, volume 32, 2019.
- [47] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, pages 13653–13662, 2021.
- [48] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2022.
- [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [50] Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024.
- [51] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.
- [52] Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.
- [53] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.
- [54] Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation. *arXiv preprint arXiv:2501.10687*, 2025.
- [55] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2025.
- [56] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Fei Wu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [57] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024.

- [58] Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Ke-qiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency model. *arXiv preprint arXiv:2405.18407*, 2024.
- [59] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.
- [60] Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025.
- [61] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049, 2021.
- [62] Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. *arXiv preprint arXiv:2506.09984*, 2025.
- [63] Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025.
- [64] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.
- [65] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024.
- [66] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023.
- [67] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.
- [68] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. <https://arxiv.org/abs/2311.18828>, 2024.
- [69] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *ICCV*, 2025.
- [70] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.
- [71] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, pages 8652–8661, 2023.
- [72] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, pages 3657–3666, 2022.
- [73] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. In *ICLR*, 2025.
- [74] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.
- [75] Tianyun Zhong, Chao Liang, Jianwen Jiang, Gaojie Lin, Jiaqi Yang, and Zhou Zhao. Fada: Fast diffusion avatar synthesis with mixed-supervised multi-cfg distillation. *arXiv preprint arXiv:2412.16915*, 2024.