Research

Avatar V: Scaling Video-Reference Avatar Video Generation

HeyGen Research

Abstract

Generating avatar videos that are not merely visually similar to a target individual but behaviorally recognizable, faithfully reproducing their talking rhythm, gestural tendencies, and expression dynamics, remains an open challenge. Existing methods predominantly condition on single static images, which provide insufficient identity information and cannot capture dynamic motion traits, while standard pixel-level training objectives underserve the perceptually critical facial regions that determine avatar fidelity. To solve these issues, we present Avatar V, a production-scale framework that addresses these limitations through video-reference-conditioned identity modeling. Rather than compressing identity into fixed-size embeddings, the model conditions directly on the full token sequence of a reference video, learning to extract and reproduce both static identity attributes (facial geometry, skin texture) and dynamic behavioral patterns (talking rhythm, micro-expressions) through attention over the reference context. To make this formulation practical, we introduce *Sparse Reference Attention*, an asymmetric mechanism that achieves linear-complexity conditioning on arbitrarily long references. To capture individual-specific motion style, we propose a dedicated motion representation stream that enables *closed-loop talking style transfer*. To recover perceptually critical facial details at production resolution, we design an *identity-aware super-resolution refiner* that inherits the full reference conditioning apparatus. These components are supported by a scalable data engine curating 100M+ training clips from 50M raw videos with cross-clip identity connectivity, and a five-stage progressive training pipeline incorporating flow matching pre-training, personality fine-tuning, two-phase distillation (>10× acceleration), and RLHF alignment, deployed across thousands of GPUs. Avatar V generates 1080p videos of unlimited duration, achieving state-of-the-art performance across identity preservation, lip synchronization, and generation quality on our cross-scene benchmark, consistently outperforming leading systems including Seedance 2.0, Kling O3 Pro, Veo 3.1, and OmniHuman 1.5 in both automated metrics and human evaluation.

Project Page: https://www.heygen.com/research/avatar-v-model

Contents

1 Introduction

Over the past year, the field of video generation has undergone a decisive shift from unimodal synthesis toward multimodal, controllable, and identity-aware generation. Proprietary systems such as Sora [4], Kling [28], and Seedance [5, 6], alongside open-source models including Wan [45], CogVideoX [60], and HunyuanVideo [27], have transformed video generation from a research curiosity into a practical, utility-driven capability. In parallel, audio-driven portrait animation has seen rapid progress through diffusion-based approaches [8, 14, 15, 19, 37, 47, 56], enabling increasingly realistic talking-head synthesis from reference images and driving audio signals.

Despite this remarkable progress, generating production-quality talking avatar videos, where a digital human faithfully reproduces a real person's appearance, expressions, and talking style across diverse scenes, remains an open and multi-faceted challenge. Current systems still fall short in three fundamental aspects:

- **Shallow identity representation.** Nearly all existing methods condition generation on a single static reference image [19, 37, 56], which captures the subject from one viewpoint, under one lighting condition, with one expression. This forces the model to hallucinate unseen views and articulation patterns, leading to identity drift, loss of fine-grained facial details, and an inability to reproduce the individual's characteristic talking style. Recent works explore video-based references [7, 11, 29], but naively concatenating all reference tokens with generation tokens incurs prohibitive quadratic attention cost. Moreover, these approaches lack explicit supervision on both static identity similarity (facial geometry, skin texture) and dynamic motion fidelity (talking rhythm, expression dynamics).
- **Decoupled appearance and motion style.** A convincing avatar must not only *look like* the target person but also *move like* them. Existing systems typically treat identity as a static embedding and motion as a separate conditioning signal, failing to capture the individual's talking rhythm, habitual micro-expressions, and gestural tendencies.
- **Sparse supervision for perceptually critical regions.** Standard diffusion training optimizes a pixel-level loss that distributes learning signal uniformly across the frame, yet the regions most critical for avatar quality (lip shape, teeth, micro-expressions, eye gaze) occupy a small fraction of total pixels. This leads to undertrained facial details and poor lip synchronization, and conventional training pipelines have not been systematically adapted for identity-preserving avatar generation.

To address these challenges, we present Avatar V, a reference-video-based personalized large video generation model for production-scale talking avatar synthesis. The central idea is to formulate personality embedding as a *video-reference conditioning* problem: rather than compressing identity into fixed-size embeddings, the model conditions directly on the full token sequence of the user's reference video, learning to extract and reproduce both static identity attributes and dynamic behavioral patterns through attention over the reference context. Given a short reference video of any individual, Avatar V generates 1080p avatar videos of infinite duration that faithfully preserve the target person's appearance and talking style. Avatar V has been deployed to serve millions of generation requests. Our contributions span model architecture, data curation, and training methodology:

- **Video-reference identity conditioning via Sparse Reference Attention.** We condition generation on full video references of arbitrary length through an asymmetric attention mechanism that models both static identity features (facial geometry, skin texture, accessories) and dynamic behavioral patterns (talking rhythm, habitual expressions, gestural tendencies). Generation tokens attend to all reference tokens for fine-grained identity extraction, while reference tokens only self-attend, reducing complexity from quadratic to linear in reference length (Section 2).
- **Talking style modeling via motion representation.** We introduce a dedicated motion stream that simultaneously serves as a generation target and a conditioning signal, creating a closed-loop training signal for learning each individual's characteristic motion patterns. Through joint optimization of these dual roles, the model develops a unified understanding of the target speaker's motion dynamics (Section 2).
- **Identity-aware super-resolution refiner.** We design a super-resolution module that inherits the full video reference conditioning apparatus, leveraging identity and audio signals to recover fine-grained facial details

lost at base resolution, with efficient sparse temporal attention for practical high-resolution inference (Section 2).

These architectural contributions are supported by a co-designed data and training infrastructure:

- **Scalable data curation with cross-identity connectivity.** We build a data engine producing training data at three quality tiers from over 50M raw videos, with an identity-aware cross-clip connectivity graph that links same-identity clips across visually distinct scenes for disentangling identity from scene-specific details (Section 5).
- **Human-aware progressive training.** We develop auxiliary losses in learned representation spaces (identity, motion, lip-sync, perceptual fidelity) integrated into a five-stage pipeline spanning text-to-video pretraining, audio-to-video pretraining, personality SFT, two-phase distillation for over 10× acceleration, and reinforcement learning from human feedback (Section 3).
- **End-to-end production system.** We deploy the full pipeline across 5,000+ GPUs with inference optimizations including context caching, sequence parallelism, fused kernels, and streaming VAE decode, under a unified multi-cloud infrastructure with QoS-aware scheduling (Sections 4 and 6).

Comprehensive experiments on our cross-scene benchmark demonstrate state-of-the-art performance across all evaluated dimensions, consistently outperforming leading systems including Seedance 2.0, Kling O3 Pro, Veo 3.1, and OmniHuman 1.5 in both automated metrics and human evaluation.

2 Model Design

We present Avatar V, an identity-preserving, audio-driven video generation system built on a Diffusion Transformer (DiT) architecture with the flow matching training framework. Avatar V takes as input a short user video (ranging from a few seconds to several minutes), a target audio track, and a text prompt describing the desired scene, and generates high-fidelity high resolution avatar videos that faithfully reproduce the target person's appearance, expressions, and talking style. The system comprises four major components: an Identity-Preserving Image Engine for scene image generation, a core DiT model with Sparse Reference Attention for personality-embedded video generation, an identity-aware super-resolution refiner for high-resolution output, and a curated data pipeline that enables cross-identity training. The overall architecture is illustrated in Figure 1.

2.1 VideoRef DiT: Video-Reference Personality Modeling

The core of Avatar V is a Diffusion Transformer that formulates personality embedding as a *video-reference conditioning* problem: rather than compressing identity into low-dimensional embeddings or fixed-size feature vectors, the model directly conditions on the full token sequence of the user's reference video, learning to extract and reproduce fine-grained identity details through attention at every transformer layer. The reference video tokens serve as a rich identity context: the model observes the target person's appearance, expressions, and motion patterns, and generates new video that is consistent with these observations. This approach offers key advantages over parametric identity encoding: it scales naturally with reference length (more context yields richer identity information), requires no identity-specific fine-tuning at inference time, and preserves the full visual richness of the reference rather than discarding information through a bottleneck.

Design objectives. The architecture is guided by three objectives that are jointly addressed through the components described below. *Long-form temporal consistency*: the model must maintain stable identity, coherent motion, and consistent scene composition across sequences ranging from several seconds to over a minute. *Audio-visual synchronization*: dedicated audio cross-attention modules (Section 2) align speech content with visual articulation at the phoneme level. *Natural motion dynamics*: beyond lip movements, the model must reproduce co-speech gestures, gaze shifts, and postural sway that collectively determine perceived naturalness.

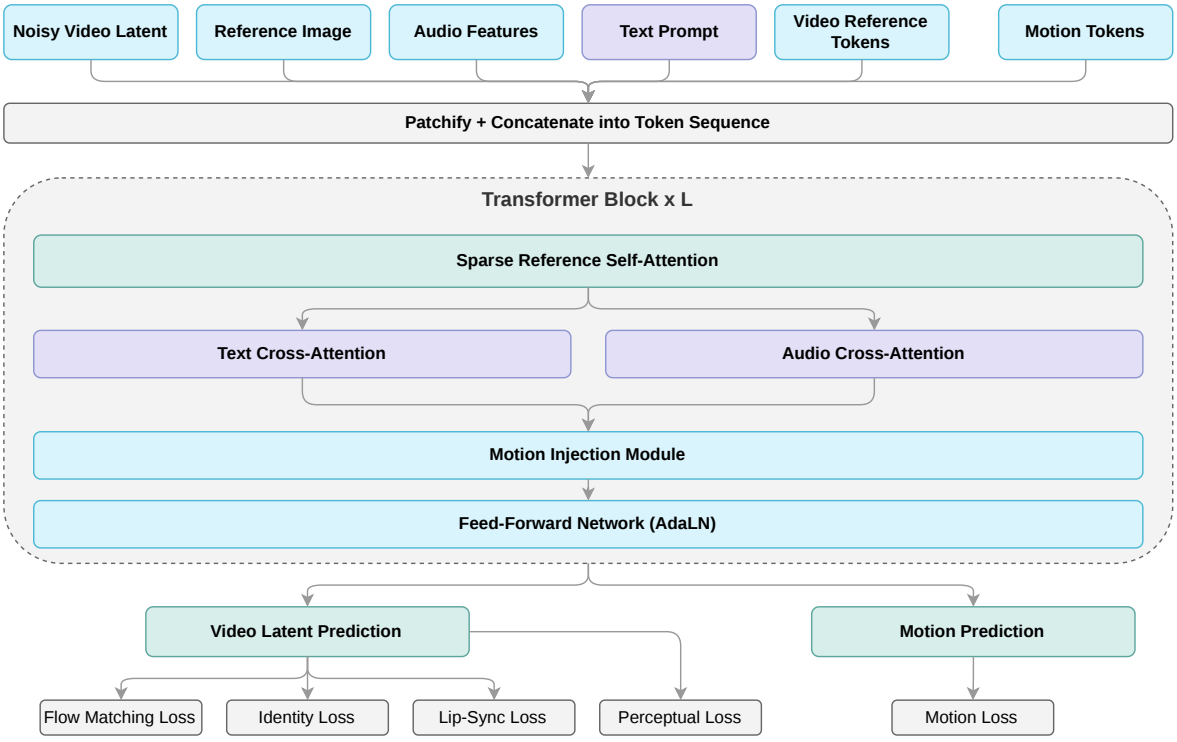

Figure 1 Avatar V Architecture. Multi-modal inputs are patchified into a unified token sequence and processed through L transformer blocks. Each block contains Sparse Reference Self-Attention, text and audio cross-attention, a Motion Injection Module, and an AdaLN-modulated feed-forward network. The model produces a video latent prediction supervised by flow matching and human-aware losses, alongside an auxiliary motion prediction.

Static and Dynamic Identity Modeling. What distinguishes Avatar V from existing systems is its ability to model both *static* and *dynamic* aspects of personal identity. Static features include fine-grained, time-invariant characteristics such as dental structure, skin texture and wrinkles, facial geometry, hair style and color, and accessories. Dynamic features encompass the individual's characteristic behavioral patterns: talking rhythm and mouth movement amplitude, habitual micro-expressions and smile characteristics, and gestural tendencies during speech. The model supports reference videos of arbitrary length: short references provide basic appearance information, while longer references enable the model to observe and internalize the individual's talking cadence and expression dynamics. This scalability is achieved without architectural modification, allowing the system to gracefully adapt to the available reference material. The result is that generated videos are not merely facially similar to the target but are *behaviorally recognizable*: the generated person looks like and acts like the target individual.

Sparse Reference Attention. To fully exploit the dynamic visual features contained in the user's reference video while maintaining computational tractability, we introduce Sparse Reference Attention, a structured sparsity mechanism for video-reference identity conditioning. Standard approaches either compress references into low-dimensional bottlenecks that lose fine-grained identity details, or concatenate all reference tokens with generation tokens incurring prohibitive quadratic cost as reference length grows. Sparse Reference Attention addresses this trade-off through a carefully designed sparsity pattern that preserves full access to identity information during generation while eliminating redundant computation among tokens that do not require mutual interaction. The resulting complexity scales almost linearly with reference length, enabling the model to condition on minutes-long reference footage that captures not only static appearance but also the subject's characteristic expressions, gestures, and talking rhythm, without architectural modification.

Talking Style Modeling via Motion Representation. Talking style, the characteristic temporal pattern of facial movements, mouth shapes, and head gestures during speech, is a crucial but challenging aspect of identity. We observe that talking style can be understood as a temporal variation pattern over motion representations: the same phoneme sequence produces visually different articulation patterns depending on the speaker's individual style. To capture this, we introduce a dedicated motion representation stream that serves as both a learning objective and a conditioning signal within the model. Through joint optimization of these two roles, the model develops a unified understanding of the target speaker's motion dynamics, enabling it to both internalize and reproduce characteristic talking patterns. This integrated design yields faithful style transfer even for unseen speech content, producing generated videos that are behaviorally consistent with the reference speaker.

Human-Aware Auxiliary Losses. Traditional audio-to-video training relies solely on pixel-level diffusion loss, which provides insufficient learning signal for subtle but perceptually critical features like talking style and micro-expressions, which represent a small fraction of total pixel variation yet are essential for identity perception. To address this, we introduce a suite of human-aware auxiliary losses that provide semantically meaningful supervision beyond raw pixels, covering identity consistency, motion fidelity, audio-visual synchronization, and perceptual quality. These losses operate in learned representation spaces rather than pixel space, providing denser and more informative supervision for the human-centric aspects of avatar generation.

2.2 Identity-Preserving Image Engine

The Avatar V pipeline begins with the construction of a high-fidelity, identity-preserving scene image that serves as the visual anchor for subsequent video generation. Given only a short user-provided video, the Image Engine is tasked with generating a photorealistic image of the user in a novel scene while faithfully preserving their unique facial identity. A key design principle is the efficient and thorough exploitation of the user's input video: rather than relying on a single reference frame (which may suffer from suboptimal pose, expression, or occlusion), the pipeline automatically selects a diverse set of frames spanning multiple viewpoints and expressions. This multi-view, multi-expression sampling strategy ensures that the identity representation is both comprehensive and robust, supplying the generation engine with sufficient information to hallucinate consistent identity across novel viewpoints while reproducing subtle identity cues such as smile asymmetry, dimple patterns, and nasolabial fold characteristics. The resulting scene image satisfies several quality criteria: identity fidelity in a learned embedding space, scene diversity through text-prompt-controlled backgrounds and lighting, photorealism with natural skin texture and coherent illumination, and flexible resolution and aspect ratio to accommodate diverse downstream video generation requirements.

2.3 Audio Engine: LLM-Based Voice Cloning

In addition to visual identity, faithful voice reproduction is essential for convincing avatar videos. The Avatar V Audio Engine is a proprietary voice cloning system built on a large language model (LLM) backbone that generates target-speaker speech from arbitrary text input. Given a short audio sample from the user's reference video (as little as 10 seconds), the system extracts a speaker embedding that captures the individual's vocal timbre, prosody patterns, speaking rate, and accent characteristics. The LLM-based architecture models speech generation as a sequence prediction task over discrete audio tokens, enabling it to synthesize natural, expressive speech that faithfully preserves the speaker's voice identity while supporting multilingual output and emotion control. The Audio Engine operates as a standalone module in the Avatar V pipeline: for text-to-avatar use cases, it generates the driving audio track from the user's script before passing it to the VideoRef DiT; for audio-driven use cases, the user-provided audio is used directly. This modular design decouples voice synthesis from video generation, allowing each component to be independently improved and scaled.

2.4 Super-Resolution Refiner with Identity Modeling

While the base DiT operates at low resolution to maintain computational tractability, production deployment demands high-resolution output. The Avatar V Super-Resolution Refiner bridges this gap through an identity-

aware upsampling module that enhances visual fidelity without compromising identity consistency. The Refiner shares the same DiT backbone as the base model but accepts the low-resolution output as additional conditioning input alongside the high-resolution noise, enabling it to leverage the base model's predictions as a strong prior for detail synthesis.

Identity-Aware Conditioning. A naive super-resolution approach would treat upsampling as a purely visual enhancement task, potentially introducing identity-inconsistent artifacts in facial regions. The Avatar V Refiner instead inherits the full identity modeling apparatus from the base DiT, including video reference conditioning, audio features, and motion representations, ensuring that facial identity is preserved with high fidelity throughout the upsampling process.

Efficient Inference via Sparse Temporal Attention. Since the base model has already established strong temporal coherence at low resolution, the Refiner's primary role is local detail enhancement rather than global temporal reasoning. We therefore employ sparse temporal attention that restricts each frame's receptive field to a local neighborhood, significantly reducing the computational cost at high resolution. Through a multi-stage distillation process, the Refiner achieves high-quality upsampling in very few denoising steps, enabling practical inference latency at high resolution.

The model architecture described above places stringent requirements on training data: the Sparse Reference Attention mechanism requires same-identity video pairs across diverse scenes, the motion representation stream demands dense temporal annotations, and the human-aware auxiliary losses rely on per-frame quality signals. The following section describes the scalable data curation pipeline that meets these requirements.

3 Training Strategy

We adopt a progressive multi-stage training paradigm that systematically develops the model's capabilities from general video understanding to identity-specific personality embedding and human-preference alignment. Our training pipeline consists of five major phases: text-to-video general pre-training, audio-to-video pre-training, personality supervised fine-tuning (SFT), distillation, and reinforcement learning from human feedback (RLHF). This structured approach enables the model to learn temporal dynamics, identity preservation, efficient inference, and perceptual quality in a stable and efficient manner. Figure 2 summarizes our training pipeline.

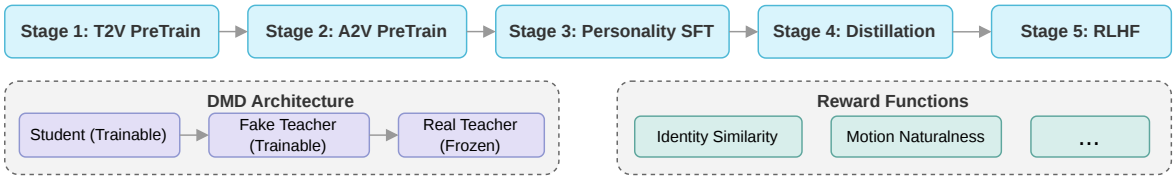

Figure 2 Avatar V Training Pipeline. The five-stage curriculum progressively builds capabilities from general video generation through audio-driven synthesis and identity-preserving personality embedding to distillation and human feedback alignment. Bottom panels detail the DMD distillation architecture and GRPO reward functions.

3.1 Text-to-Video General Pre-Training

The first stage establishes the model's foundational video generation capabilities through large-scale text-to-video (T2V) pre-training on the diverse pretraining corpus described in Section 5. This stage trains only the base-resolution DiT; the super-resolution refiner is not involved.

Progressive resolution and duration scaling. Following the curriculum strategy adopted by recent video generation models [6, 27, 60], we progressively increase spatial resolution and temporal length throughout pre-training. Training begins at low resolution with short clips to allow the model to rapidly converge on basic motion patterns and scene composition, then gradually scales to higher resolution and longer sequences.

This coarse-to-fine strategy improves both training stability and final generation quality compared to directly training at the target resolution.

Multi-task joint training. The T2V pre-training phase jointly trains on text-to-video and image-to-video (I2V) generation tasks within a unified framework. For I2V, the first frame is provided as a conditioning image, and the model learns to generate temporally coherent video continuations. This multi-task formulation enables the model to learn both unconditional video dynamics from T2V and frame-conditioned temporal extension from I2V, providing a versatile foundation for downstream avatar-specific fine-tuning.

Training objective. We adopt rectified flow matching [17, 33] as the training objective, where the model learns to predict the velocity field that transports noise to data along straight-line trajectories. A logit-normal timestep distribution concentrates training signal on intermediate noise levels where the learning signal is most informative. We use a distributed Muon optimizer [26] for 2D+ weight tensors for improved convergence on large-scale models, and AdamW for embeddings and 1D parameters. A cosine learning rate schedule with linear warmup is applied throughout.

3.2 Audio-to-Video Pre-Training

Building on the T2V foundation, the second stage specializes the model for audio-conditioned avatar video generation, following the line of audio-driven portrait animation works [15, 37, 47, 56]. Starting from the pre-trained T2V checkpoint, the model is adapted to accept a conditioning image (the first frame) and a driving audio track, learning to generate temporally coherent video continuations with synchronized lip movements and natural head motion. This stage introduces the audio cross-attention modules and trains them jointly with the visual backbone on a broad corpus of talking-head video data spanning diverse speakers, languages, and speaking styles. Progressive resolution scaling and dynamic sequence length sampling build robust multi-scale generation capabilities.

3.3 Personality Supervised Fine-Tuning

The supervised fine-tuning stage transforms the general-purpose video model into an identity-aware avatar generator by training on the curated same-identity-different-scene dataset described in Section 5. During SFT, the model receives reference videos of the target identity through the Sparse Reference Attention mechanism, teaching the model to extract and utilize identity information from video references. The motion representation pathways are activated during this stage, enabling the model to learn talking style transfer. Leveraging the cross-clip connectivity from the data pipeline, each training example consists of a target video clip paired with reference clips from different scenes but the same identity, forcing the model to extract identity-invariant features rather than simply copying scene-specific details. The human-aware auxiliary loss suite is progressively activated, providing dense semantic supervision that guides the model toward identity-faithful, expressively accurate, and well-synchronized generation.

3.4 Model Distillation

To enable practical deployment with low inference latency, we apply a two-phase distillation strategy [42] that reduces both the number of classifier-free guidance (CFG) evaluations and the number of denoising steps required for generation.

CFG Distillation. The VideoRef DiT employs multiple classifier-free guidance [22] streams to control different conditioning aspects of the generated video. At inference time, each CFG stream requires a separate forward pass with the corresponding condition dropped, resulting in a multiplicative increase in computational cost. To eliminate this overhead, we distill the multi-stream CFG behavior into a single forward pass [36], reducing the per-step cost by a factor proportional to the number of guidance streams while preserving generation quality.

DMD Distillation. Following CFG distillation, we further reduce the number of denoising steps using an improved Distribution Matching Distillation (DMD) framework [44, 62]. Our implementation employs a three-model architecture: a trainable *student* that generates video from pure noise in a single forward pass, a trainable *fake teacher* that models the student's output distribution, and a frozen *real teacher* that provides the target distribution from the original multi-step model. The student learns to match the real teacher's distribution through a combination of distribution matching gradients and progressive distillation objectives, with an optional adversarial loss for sharpening fine-grained details. Our implementation incorporates several stability improvements over vanilla DMD, resulting in more reliable convergence. The combined two-phase distillation pipeline reduces the total inference cost by over an order of magnitude while maintaining generation quality comparable to the original multi-step, multi-CFG model.

3.5 Human Feedback Alignment

The final training stage aligns the model with human perceptual preferences through reinforcement learning from human feedback [2]. We employ multiple reward signals that jointly cover identity fidelity, motion naturalness, and visual quality. We adopt Group Relative Policy Optimization (GRPO) [25, 43, 64] as the primary RL algorithm, adapted with a flow-matching-compatible formulation for efficient policy gradient computation within the diffusion framework. KL regularization against the pre-RLHF model prevents quality degradation on previously learned capabilities. As a complementary approach, we also support Direct Preference Optimization (DPO) [41, 48] training, which learns directly from human-annotated preference pairs without requiring online generation. The preference pairs used for DPO training are collected through the human annotation system described in Section 5.

4 Inference

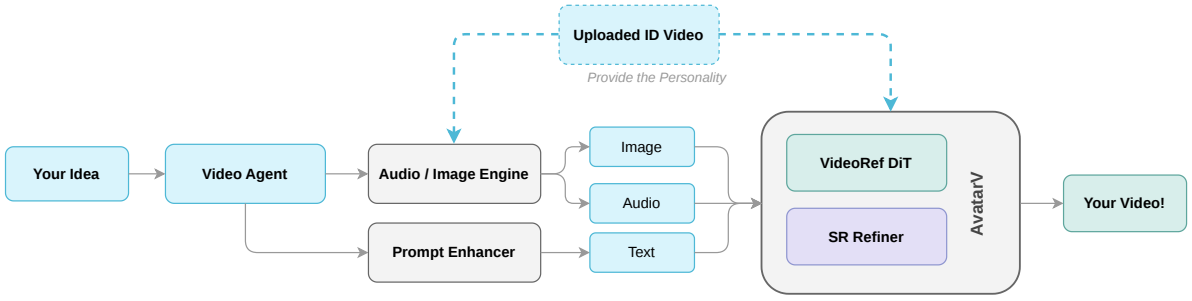

Figure 3 Avatar V Inference Pipeline. The user's identity video is processed once into a reusable personality embedding. Scene image generation and prompt engineering proceed in parallel, then all signals are combined for low-resolution DiT generation followed by identity-aware super-resolution to high-resolution.

This section describes the inference pipeline and the optimizations that enable Avatar V to generate high-fidelity, high-resolution talking avatar videos with practical latency. All inference uses the distilled model from Section 3, which internalizes classifier-free guidance and operates in a reduced number of denoising steps (24 steps after two-phase distillation), eliminating the need for multiple conditional and unconditional forward passes at each step.

End-to-end pipeline overview. As illustrated in Figure 3, inference proceeds in four stages. (1) *Preprocessing*: the user's reference video is encoded once into video reference tokens, identity embeddings, and expression embeddings; audio features are extracted from the target audio track; and scene description prompts are encoded into text embeddings. The Identity-Preserving Image Engine generates a scene image conditioned on the reference identity (Section 2). These preprocessing steps run in parallel. (2) *DiT generation*: the base-resolution DiT performs chunk-based autoregressive generation, conditioning on all preprocessed signals through Sparse Reference Attention. (3) *Super-resolution*: an identity-aware SR refiner upscales the output

to high resolution in a single denoising step. (4) *Streaming decode*: a streaming VAE decoder converts latents to pixels incrementally, producing output frames before the full video is complete.

4.1 Chunk-Based Long-Form Generation

Avatar V generates videos through a chunk-based autoregressive pipeline that enables arbitrarily long video synthesis while maintaining temporal coherence. Each chunk produces 41 latent frames (corresponding to 161 pixel frames at 25 fps, approximately 6.4 seconds of video), and chunks are connected through a prefix conditioning mechanism.

Ref2V + Prefix2V Strategy. The first chunk operates in *ref2v mode*, where the reference video frame is encoded as the identity conditioning signal. Subsequent chunks operate in *prefix2v mode*: the last frames of the previous chunk serve as the prefix condition for the next chunk, providing a smooth temporal bridge. Adjacent chunks share a 2-frame overlap to ensure seamless transitions. This sequential strategy eliminates the need for separate interpolation chunks, simplifying the generation flow while maintaining temporal consistency.

Global Appearance Anchor. For multi-chunk generation, we extract a global appearance anchor from the first generated chunk. This anchor, combined with motion frame propagation between consecutive chunks, ensures that the subject's identity remains consistent across arbitrarily long videos.

4.2 Diffusion Sampling

Improved Stochastic Euler. Deterministic ODE-based samplers for flow matching models struggle with high-frequency details at reduced step counts, manifesting as unstable hand clarity, blurred dental structures, and temporally inconsistent fine textures. We adopt a stochastic overshoot-and-renoise strategy: at each step, the sample is advanced beyond the target noise level by a controlled overshoot factor, then stochastically renoised back to the correct level with fresh Gaussian noise. This controlled stochasticity improves detail recovery in high-frequency regions and prevents the accumulation of discretization errors, enabling high-quality generation in 24 steps with stable hand, teeth, and facial detail quality comparable to much longer sampling schedules.

4.3 VideoRef Context Caching and Sparse Attention

A key observation is that video reference tokens, encoded from clean, non-noisy reference frames, remain invariant across denoising steps. We exploit this through a two-level caching strategy:

Context-Level Caching. At the first denoising step ($t = 0$), the full video reference context (latents, audio features, validity masks, expression and identity embeddings) is computed and cached. For all subsequent steps ($t > 0$), this cached context is reused without recomputation, avoiding the substantial cost of re-encoding the reference video at each of the 24 denoising steps.

Attention-Level KV Caching. Within each transformer block's reference self-attention layer, the key and value projections of video reference tokens are computed once at the first denoising step and cached in GPU memory. Subsequent steps directly concatenate the cached reference KV tensors with freshly computed generation token KV tensors for the asymmetric attention computation, eliminating redundant linear projections and RoPE computations across all steps.

Sparse Validity Masking. Video reference sequences can contain invalid tokens (e.g., frames where the face is not visible). We implement sparse attention masks that skip computation for these invalid tokens. The validity masks are precomputed per-rank for the sequence-parallel layout and applied to self-attention, cross-attention, and FFN layers, avoiding wasted computation on non-informative reference tokens.

4.4 Distributed Inference with Sequence Parallelism

Avatar V employs Ulysses Sequence Parallelism (USP) across 8 GPUs within a single node to distribute the computation of long token sequences. The input sequence, comprising video latents, reference tokens, high-resolution face tokens, and conditioning tokens, is partitioned along the sequence dimension across GPU ranks, with all-to-all communication for attention operations that require cross-rank token interaction.

FSDP2 with CPU Offloading. Model parameters are sharded across GPUs using FSDP2 with CPU offloading for inactive parameter shards. This frees GPU memory for the large intermediate activations required by the DiT and enables multi-model co-location: multiple model variants can reside on a single machine with rapid switching via CPU-to-GPU loading rather than full reloading from disk. Forward prefetching overlaps the AllGather of the next block's parameters with the current block's computation, hiding communication latency. Processes are pinned to the NUMA node of their assigned GPU to ensure optimal memory bandwidth for these CPU-GPU transfers.

4.5 Inference Acceleration

Deploying Avatar V at production scale surfaces three latency bottlenecks: (1) kernel launch overhead and memory bandwidth waste from thousands of small operators per transformer block, (2) coarse-grained inter-GPU synchronization in sequence-parallel attention, and (3) hardware-level frequency variance across GPU ranks causing straggler effects at collective boundaries. We address each through a custom compiler with agentic kernel synthesis, NVSHMEM-based sequence parallelism, and targeted system-level tuning, achieving a combined 3× latency reduction over the unoptimized baseline and 33% latency reduction over the torch inductor compiler optimized version.

4.5.1 Custom Compiler with Agentic Kernel Synthesis

Limitations of existing compilers. While `torch.compile` with the Inductor backend provides a general-purpose optimization path, we find it insufficient for production diffusion inference at our scale. First, Inductor's pattern matcher fails to capture many cross-operator fusion opportunities, leaving significant memory bandwidth on the table. Second, the Triton kernels it generates are suboptimal for the specific tensor shapes and access patterns in our model. Third, Inductor handles dynamic shapes and dynamic tensor memory operations poorly, relying on a guard-and-recompile strategy that triggers excessive recompilation. Fourth, on a complex production codebase with many control-flow paths, the compilation overhead itself becomes prohibitive.

Agentic kernel synthesis. We introduce a compiler workflow that combines human expertise with LLM-based kernel generation. In the first phase, engineers profile the end-to-end forward pass and define fusion scopes, identifying which operator subgraphs should be merged into single kernel launches. In the second phase, an LLM-based agent takes each fusion specification and generates CUDA/Triton kernel candidates through an iterative evolution process.

A key challenge is that kernel-level profiling is inherently noisy due to GPU thermal state, memory allocator behavior, and scheduling variance. To mitigate this, we adopt an evolution island strategy: 2–3 islands run in parallel, each exploring different tiling strategies and memory access patterns across 4 candidates per generation. Fitness is evaluated on both kernel latency and numerical accuracy, and the best candidate is selected across islands after a fixed iteration budget. This avoids overfitting to noisy single-point measurements while keeping the search tractable.

Results. The agentic workflow produces mega kernels that fuse entire non-attention portions of each transformer block into single kernel launches, eliminating intermediate tensor materializations and minimizing memory round-trips. The compiled forward pass reduces from thousands of small kernels to only Flash Attention calls, cuBLAS GEMMs, and a handful of fused mega kernels. This yields approximately 3× latency reduction over the unoptimized baseline and 33% improvement over the `torch.compile` Inductor backend.

4.5.2 NVSHMEM-Based Sequence Parallelism

Avatar V distributes attention computation across 8 GPUs via Ulysses Sequence Parallelism, which requires all-to-all communication at every transformer block. The standard approach using NCCL operates at kernel-level granularity: each all-to-all is a monolithic operation that must fully complete before downstream compute can begin. We replace this with NVSHMEM-based communication that exploits NVLink for direct GPU-to-GPU data movement with tile-level dataflow control.

With NVSHMEM, individual data tiles can be sent, received, and synchronized within a single fused kernel, enabling sub-tensor pipelining: the all-to-all scatter, cuBLAS GEMM computation, and all-to-all gather overlap at fine granularity rather than executing sequentially. As each tile arrives from a remote rank, it is immediately available for computation without waiting for the full transfer to complete.

A critical design consideration is SM partitioning between communication and compute. Naively dedicating SMs to communication leaves the compute GEMM with a non-round number of thread block waves, causing up to 1.5× slowdown from the wave quantization effect, where SMs sit idle during the final partial wave. We determine the optimal partition through extensive profiling-guided auto-tuning that balances NVLink bandwidth utilization against GEMM wave efficiency.

4.5.3 System-Level Optimization

NUMA-aware process placement. Each inference rank is pinned to the CPU cores and memory controllers on the same NUMA node as its assigned GPU, ensuring optimal bandwidth for the CPU-GPU parameter transfers required by FSDP2 offloading.

GPU clock locking. In distributed inference where all ranks synchronize at collective operations, the slowest rank determines overall latency. Default GPU boost clocking allows frequency to vary across GPUs depending on thermal and power state, creating straggler ranks that gate every synchronization point. We lock all GPUs to a stable frequency below the boost ceiling, eliminating this variance and reducing overall block latency by approximately 3%.

4.6 Super-Resolution

The base DiT generates video at low resolution in latent space. To produce the final high-resolution output, we apply a single-step adversarial SR model that focuses computational resources on high-detail regions, particularly the mouth area for lip-sync fidelity. Low-resolution latents are noised at $\sigma = 0.6$ before the SR step, providing the model sufficient room for detail enhancement while preserving structural content from the base generation.

4.7 Streaming VAE Decode

The VAE decoder employs causal 3D convolutions with temporal feature caching, enabling chunk-by-chunk decoding without requiring the full video to be held in memory. Decoded frames are piped directly into an asynchronous streaming video encoder that writes the output file incrementally. This streaming pipeline ensures bounded memory consumption regardless of video length and enables the first frames to be available before the full video has been decoded.

5 Data Curation

Training Avatar V requires two distinct data regimes: a large-scale *pretraining* corpus of diverse human-centric video for learning general motion and appearance priors, and a curated *audio-to-video (A2V) fine-tuning* corpus with dense avatar-specific annotations for talking-head generation. Both corpora are produced by a unified distributed pipeline that orchestrates 25+ processing stages and 20+ specialized AI models across heterogeneous CPU and GPU infrastructure managed by our custom-built data processing engine (Section 6). Through this meticulous curation process, we ultimately obtain over 100M clips for pretraining and 10M+ clips for avatar fine-tuning.

5.1 Pretraining Data

The pretraining data is designed to capture general human-centric video priors at scale. The data is processed through a multi-stage pipeline for filtering, annotation, and feature extraction.

5.1.1 Segment-Level Curation

Raw segments pass through a 10-stage cascade. **Normalization** standardizes resolution (longest side 640px) and frame rate (25 fps), while **temporal pre-filtering** rejects choppy or static content via frame-difference analysis and perceptual hashing, both CPU-only, eliminating degenerate content before model inference. **Human detection** uses a joint object detector and face analysis model to verify human presence and define eligible temporal intervals. **Optical flow** quantifies motion statistics that feed into the clipping optimizer, and **visual quality assessment** via Q-Align scores keyframes with continuous quality scores calibrated to human opinion.

Smart clipping formulates clip selection as a constrained optimization problem, jointly maximizing clip duration while satisfying constraints on visual quality, motion, and face presence ratio, replacing brittle independent threshold cascades used in prior work. **Scene-cut detection** and **content filtering** via VLMs identify scene boundaries and reject screencasts, game footage, and static photo content. Finally, clips are **categorized** across 15 semantic dimensions for distribution balancing, and **video embeddings** are extracted for deduplication.

5.1.2 Deduplication and Feature Extraction

GPU-accelerated nearest-neighbor indexing over video embeddings groups near-duplicate clips into clusters; within each cluster, only the highest-quality clip is retained. Rule-based derived categories enable distribution rebalancing across content types.

Deduplicated clips are then processed through 13 parallel extraction stages: (1) **visual analysis**: OCR text detection, lip-sync scoring, whole-body pose estimation with dense keypoints, anatomical quality scoring, and synthetic audio detection; (2) **audio analysis**: language identification, speaker diarization, and ASR with word-level timestamps; (3) **captioning and embeddings**: an in-house audio-video understanding captioner producing rich descriptions, plus text embedding pre-encoding for diffusion conditioning; and (4) **latent pre-encoding**: multiple video VAE architectures producing latent representations for direct diffusion transformer training.

5.2 Audio-to-Video Fine-Tuning Data

The audio-to-video fine-tuning data applies additional avatar-specific curation to produce training data tailored for talking-head and portrait animation. Ten additional fine-grained quality signals are computed per-clip: eye gaze and blink patterns, face clarity, teeth and hand quality, mouth openness, camera shake, choppy frame detection, lighting consistency, and secondary speaker presence. These composable signals enable flexible quality tier construction without re-running inference.

5.3 Human Data Annotation System

Achieving the highest quality thresholds, particularly for RLHF and quality model training, requires reliable human judgment at scale. We build a distributed annotation platform supporting 100+ concurrent freelance annotators across multiple geographic regions.

Annotation tasks span five categories: (1) *quality scoring* of curated clips across perceptual dimensions (visual quality, facial naturalness, lip-sync accuracy, motion smoothness) using calibrated Likert scales, serving as ground truth for automated quality models; (2) *preference labeling* via pairwise comparisons of generated outputs along axes of identity preservation, expression naturalness, and audio-visual synchronization, producing training data for DPO and GRPO reward models (Section 3); (3) *bad case filtration* to identify subtle artifacts escaping automated filters, temporal identity drift, teeth deformation, occlusion glitches, asymmetric blinking, with flagged samples feeding back into both corpus cleaning and new detector development; (4) *generation evaluation and competitive benchmarking* through blinded side-by-side comparisons of model

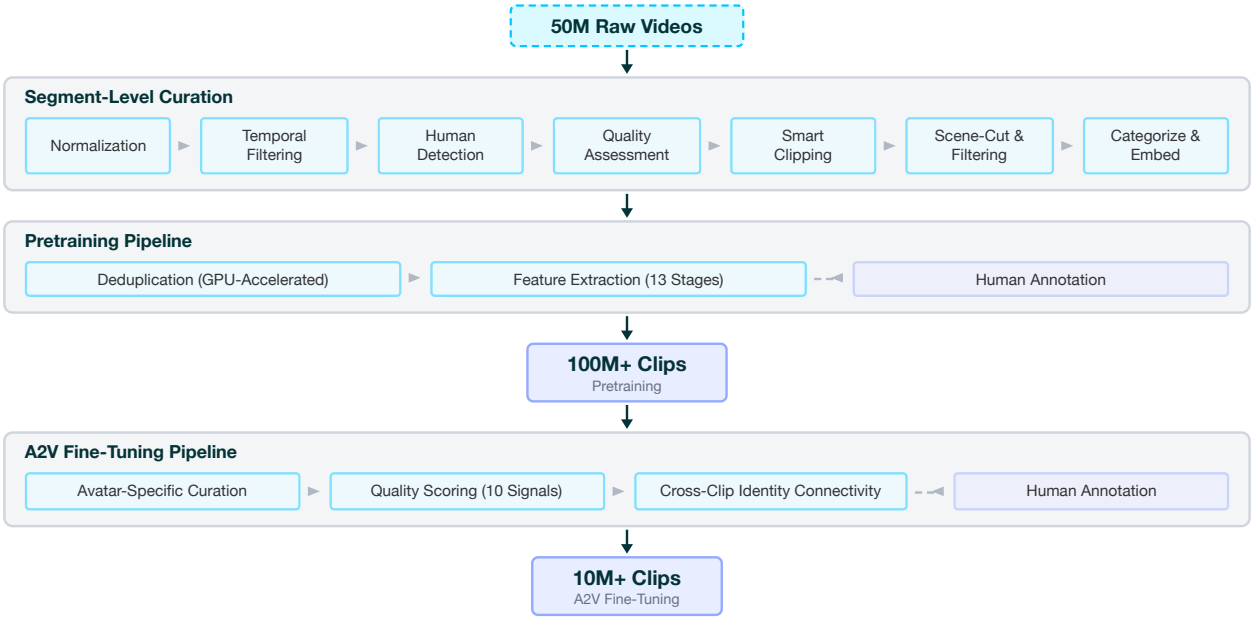

Figure 4 Data Curation Pipeline Overview. Starting from 50M raw videos, our pipeline applies shared segment-level curation before branching into pretraining (100M+ clips) and A2V fine-tuning (10M+ clips) paths, with human annotation and cross-clip identity connectivity feeding into the avatar-specific branch.

iterations and competitor systems, producing statistically grounded win-rate matrices that guide model improvement priorities and release decisions; and (5) *attribute annotation* of gaze direction, emotion, gesture type, and speaking style for conditional generation supervision.

Organization. The workforce follows a three-tier hierarchy: *Tier 1* annotators (100+ freelancers) complete qualification pipelines with calibration exercises before admission to production tasks; *Tier 2* reviewers perform random audits on annotation batches and adjudicate disagreements to produce gold-standard labels; *Tier 3* task designers define schemas, write guidelines, and manage the feedback loop between annotation results and model improvement.

Incentive system. Base compensation is supplemented by quality bonuses tied to agreement rates with reviewer audits. A leaderboard tracks accuracy and consistency, with top performers receiving priority access to higher-paying tasks. Annotators falling below thresholds are assigned re-calibration exercises. This closed-loop system maintains inter-annotator agreement above 85% across all task types.

5.4 Cross-Clip Identity Connectivity

Avatar video synthesis demands paired clips depicting the same individual across visually distinct contexts, enabling the model to disentangle identity from background, lighting, and pose. We establish cross-clip connectivity through joint filtering: two clips are linked if they depict the same individual (verified by high face similarity) in visually distinct scenes (verified by low background similarity), with sufficient duration for learning dynamic features. The resulting connectivity graph enables efficient sampling of cross-scene reference pairs during training, organized into resolution-duration groups with balanced demographic representation. This curated data, together with the annotation signals described above, feeds directly into the progressive training pipeline described next.

6 Infrastructure

Training and data processing for Avatar V require coordinating large-scale GPU workloads across heterogeneous, multi-cloud clusters while maintaining high utilization and fault tolerance. This section describes the two core infrastructure systems that support Avatar V: *HELIOS*, a unified GPU infrastructure platform for multi-cloud orchestration, and our custom-built *data processing engine* that replaced Ray to handle the scale and scheduling requirements of our video data pipelines.

6.1 HELIOS: Unified GPU Infrastructure Platform

As our models and products scaled, training, inference, and data processing increasingly competed for the same scarce GPU resources. The challenge was not merely acquiring more GPUs, but making resources from different providers, regions, and supply models behave as a single usable platform, without requiring every team to understand the underlying complexity.

We built **HELIOS** (**H**eyGen **E**ngine for **L**arge-scale **I**nfrastructure **O**rchestration **S**ervice), a unified GPU infrastructure platform for multi-cloud and large-scale operations. Today, HELIOS manages more than 5,000 GPUs across 5+ providers, 10+ regions, and 15+ standardized cells, supporting reserved, on-demand, and preemptible capacity under one system.

Standardized Onboarding. Before HELIOS, onboarding a new GPU provider or region required repeating the same engineering work (adapting networking, storage, cluster management, monitoring, and operational workflows) for each new environment. HELIOS replaces this with a standard onboarding path: a new provider or region undergoes common validation, acceptance checks, and baseline infrastructure setup before joining the platform. Once admitted, its resources are exposed through the same management model as the rest of the fleet. This reduced the average time from initial validation to production availability from two weeks to three days.

Cell-Based Architecture. Rather than building one monolithic cluster, HELIOS organizes the fleet into standardized cells, typically aligned by provider and region. Each cell is a Kubernetes cluster with a validated size boundary and a common operational baseline. This design limits blast radius: a problem in one cell is far less likely to propagate fleet-wide. It also provides a clean growth path by adding capacity through new standard cells rather than stretching a single control plane.

Two-Stage QoS-Aware Scheduling. Inference workloads require higher priority and faster response; training workloads need larger, more stable allocations over longer periods; data processing is more flexible and can tolerate interruption. Treating all workloads identically would either waste expensive capacity or create contention for critical services. HELIOS employs a two-stage scheduling model: a global scheduler makes capacity decisions based on workload QoS class, GPU type, request size, and supply model, while the selected cell handles local deployment and placement. This improved overall GPU utilization by 15% and reduced non-productive GPU time by approximately 20%.

Continuous Resource Governance. HELIOS continuously monitors key health signals across the fleet, including GPU, PCIe, and NCCL-related conditions. Unhealthy nodes are automatically isolated and routed through operational recovery workflows. The platform also detects long-idle or low-utilization resources by combining signals such as GPU utilization, memory usage, task state, and runtime progress, reclaiming and reallocating capacity according to workload priority.

Unified Observability. The platform collects signals from infrastructure, clusters, workloads, and applications, applying different sampling and retention strategies depending on the use case. In addition to standard metrics, traces, and logs, HELIOS adds finer-grained network-side observability on key nodes to identify communication bottlenecks and performance jitter in training and inference scenarios. This shared operational view shortens debugging loops, improves capacity planning, and makes cost attribution tractable across teams.

6.2 Data Processing Engine

Our video AI data processing pipelines run across multiple GPU clusters, sharing compute resources with online inference serving, where nodes are routinely preempted, rescheduled, and terminated. As data processing demand surged to over 100K concurrent tasks, we encountered fundamental scalability limitations in our initial Ray-based infrastructure and built a purpose-built replacement.

6.2.1 Design Constraints

Four constraints define the operating environment for our data processing workloads:

Heterogeneous Pipeline Stages. A video processing pipeline is a DAG of stages with mixed resource profiles: IO-bound media decode and network transfer, GPU-bound model inference, and CPU-bound processing and encoding. If a GPU stage blocks waiting on an upstream IO stage, or CPU stages over-consume capacity and starve the GPU, utilization collapses. The scheduler must be resource-profile-aware rather than treating every stage as a homogeneous unit of work.

Priority-Based Scheduling. Latency-critical pipelines feed downstream systems with tight SLAs and require preemptive priority, while background pipelines (*e.g.,* batch processing, data enrichment, training data preparation) optimize for throughput. Lower-priority pipelines must yield resources immediately when higher-priority work arrives.

GPU Fragmentation. The cluster runs multiple GPU types. When a stage requires 4 GPUs but available capacity is scattered as single-GPU slots across different nodes, those GPUs are effectively stranded. The engine must adapt placement strategy to hardware type, workload mix, and real-time capacity.

Constant Node Preemption. Data processing is colocated with production inference, so a GPU machine running a data pipeline can be reclaimed at any time. The engine must treat node failure as a continuous operating condition, detecting it within seconds rather than minutes, and redistribute work without impacting running pipelines on healthy nodes.

6.2.2 From Ray to a Custom Engine

We initially adopted Ray for its actor model, unified Python API, and built-in resource management. As concurrent pipeline counts grew, we built an external scheduler on top of Ray for priority scheduling and pipeline-aware resource allocation. However, when the data processing cluster grew to more than 2k nodes, Ray's Global Control Store (GCS) became the bottleneck: it consumed over 100 GB RSS and 400% CPU. Profiling revealed that the GCS could not keep up with message volume at this scale: stale metadata accumulated indefinitely with no eviction, and under sustained load, the GCS suffered process crashes that took down the entire coordination layer.

The root cause was architectural: the GCS is a centralized coordination point where every state change must be broadcast to every node, generating traffic that scales quadratically with cluster size. After extensive stabilization attempts (parameter tuning, cache capping, timeout reduction, telemetry disabling, and metric backend switching), each yielding only marginal improvement, we concluded we had outgrown Ray's architecture and built a purpose-built replacement.

6.2.3 Declarative Reconciliation Architecture

The core architectural decision is a *declarative* engine, drawing from the same principles behind Kubernetes. Instead of issuing imperative commands (*e.g.,* "start actor X on node Y"), the control plane declares desired state in a distributed key-value store. Nodes independently converge local state toward the declared target through an *observe* $\rightarrow$ *diff* $\rightarrow$ *reconcile* loop, replacing the entire command dispatch, acknowledgment, retry, and rollback complexity, as illustrated in Figure 5.

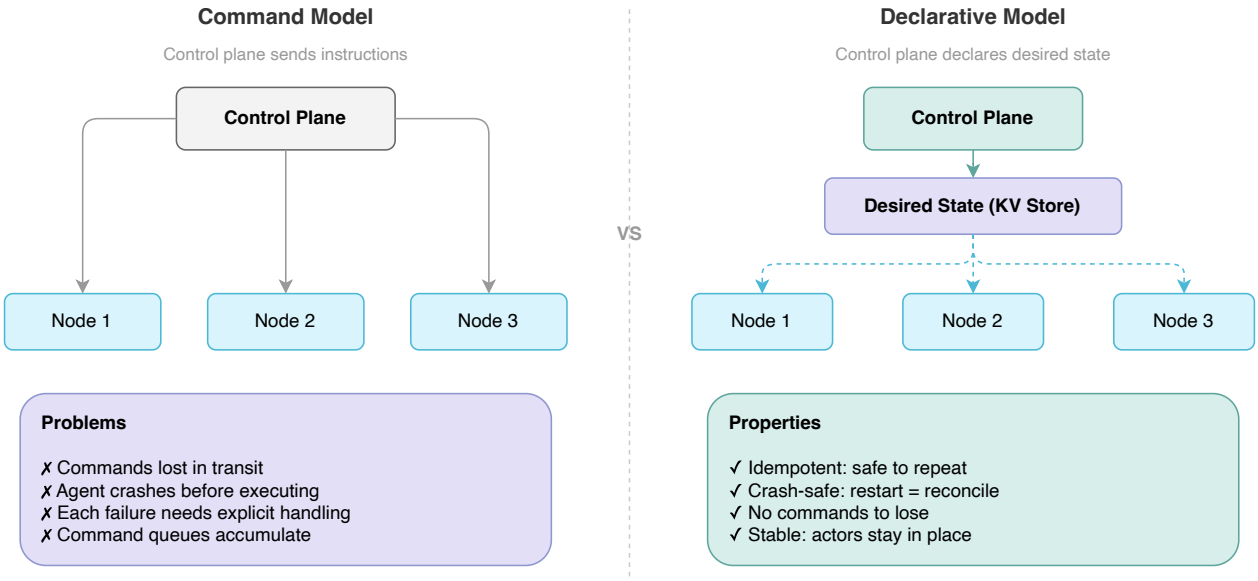

Figure 5 Command model vs. declarative model. In the command model (left), the control plane sends instructions directly to nodes, leading to problems such as lost commands and accumulated queues. In our declarative model (right), the control plane publishes desired state to a KV store, and nodes independently observe, diff, and reconcile, yielding idempotent, crash-safe operation.

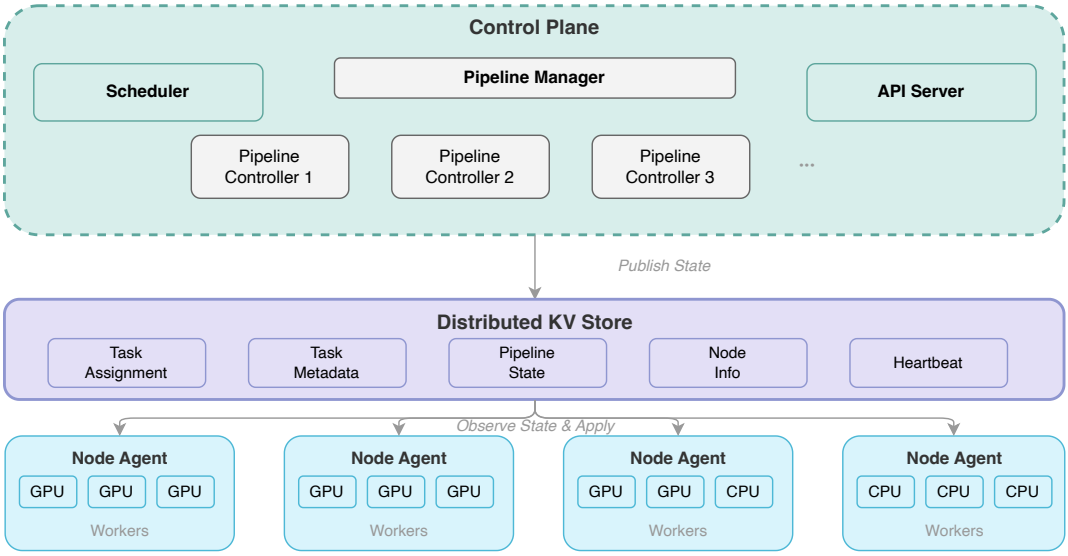

Figure 6 Four-layer architecture of the data processing engine. The control plane (scheduler and pipeline controllers) publishes desired state to a distributed KV store. Node agents observe state changes and reconcile local workers accordingly. Each layer has a single responsibility and can be independently restarted without affecting running pipelines.

This design is particularly well-suited to our workload: actors load large models onto GPUs with initialization costs measured in seconds to minutes. Once running, actors should remain in place to amortize that cost. The system optimizes for placement stability, not churn.

The engine is organized into four layers, each with a single responsibility (Figure 6):

Layer 1: Scheduler. Reads demand, metrics, and capacity from the KV store and publishes placement assignments. GPU-bound stages are packed densely to minimize fragmentation; IO/CPU-bound stages are spread for throughput balance. Latency-critical pipelines are placed first; background pipelines expand into remaining capacity and contract on demand.

Layer 2: Pipeline Controllers. One controller per pipeline declares demand for its stages, publishes workload metrics, wires the pipeline DAG, and manages lifecycle from code delivery to graceful shutdown. Controllers are fault-isolated: a controller crash affects only its pipeline, and restart re-reads state from the KV store and reconverges idempotently.

Layer 3: Node Agents. One agent per worker node runs the core reconciliation loop (read desired state, compare with actual running processes, spawn or kill workers to close the gap) and reports status to the control plane periodically.

Layer 4: Workers. Stateless, single-purpose, disposable processes. Each hosts one task instance, pulls work from a task queue, and writes output. Workers contain zero coordination logic and no awareness of the scheduler, peers, or topology. Scale-out and crash replacement use the same code path.

6.2.4 Results

The custom engine achieves several key operational improvements:

- **GPU utilization above 95%.** Priority-aware scheduling, hardware-aware bin-packing, and dynamic capacity reallocation eliminate GPU fragmentation. Background pipelines absorb every idle cycle that inference and latency-critical pipelines leave behind.
- **Highly available control plane.** Any transient failure (scheduler restart, KV failover, or network partition) does not affect running tasks. Workers continue processing, node agents continue reconciling, and when the control plane recovers, it reads current state and resumes without losing in-flight work.
- **Node failure detection in under 30 seconds.** Down from 5–10 minutes. On a cluster with constant inference preemption, this directly recovers GPU-hours that would otherwise be silently lost.
- **Linear scalability.** The control plane scales linearly with cluster size, supporting 5,000+ GPU nodes and 200K+ concurrent tasks without becoming the bottleneck, a capability that was structurally impossible with a centralized GCS.
- **Zero-downtime deployments.** Because all components are stateless and state lives in the KV store, any layer can be rolling-restarted without interrupting running pipelines.

7 Evaluation

We evaluate Avatar V through both objective automated metrics and subjective human evaluation, designed to assess identity-preserving avatar video generation across multiple perceptual dimensions. Our evaluation compares against four state-of-the-art systems spanning both avatar-specialized and general video generation models, using a diverse cross-scene benchmark that tests generalization beyond the training distribution.

7.1 Benchmark Construction

Test set. We construct a cross-scene evaluation benchmark comprising 70 test cases sourced from publicly available online videos, with a focus on talking-video scenarios. For each test case, we collect two video clips depicting the same individual in different scenes. One clip serves as the *reference video* providing identity context, while the first frame and audio track of the other clip serve as the driving signals. This cross-scene setup is deliberately more challenging than same-identity-same-scene evaluation, as it requires the model to transfer identity information across visual contexts rather than simply reproducing the reference scene.

Figure 7 Qualitative comparison: same-scene condition. The reference video (cyan border, top row) provides identity context. All five methods generate from the same driving audio and scene image. Avatar V produces the most faithful identity and natural motion.[1]

Scene conditions. To evaluate robustness under varying scene configurations, we test three scene generation modes: (1) *Same-scene*, where the target scene image is drawn from the same scene as the reference video, providing an upper-bound on scene familiarity; (2) *Cross-scene*, where the target scene image comes from a different real video of the same individual, testing cross-context generalization; and (3) *Generated-scene*, where the scene image is produced by our Identity-Preserving Image Engine (Section 2), representing the fully automated production pipeline. Figures 7, 8, and 9 show representative examples from each condition.

In the cross-scene condition (Figure 8), the target scene image is drawn from a different video of the same individual, requiring the model to disentangle identity from scene-specific details. Methods that rely on scene cues rather than genuine identity features exhibit noticeable quality degradation, while Avatar V maintains faithful identity transfer across visually distinct contexts.

The generated-scene condition (Figure 9) represents the fully automated production pipeline, where the scene image is synthesized by the Identity-Preserving Image Engine. This is the most challenging setting as it combines identity transfer with a novel, unseen background. Avatar V produces temporally coherent outputs with natural expressions, while several competing methods struggle with identity consistency or introduce visible artifacts.

Competing methods. We compare against four systems: Kling O3 Pro, Veo 3.1, OmniHuman 1.5, and Seedance 2.0, accessed through their latest publicly available versions at the time of evaluation. Some competitors generate audio that does not match the original speech; in these cases, we evaluate the visual output independently.

[1] All videos shown in this report are for research demonstration purposes only. HeyGen's platform enforces consent verification for all digital twin creation.

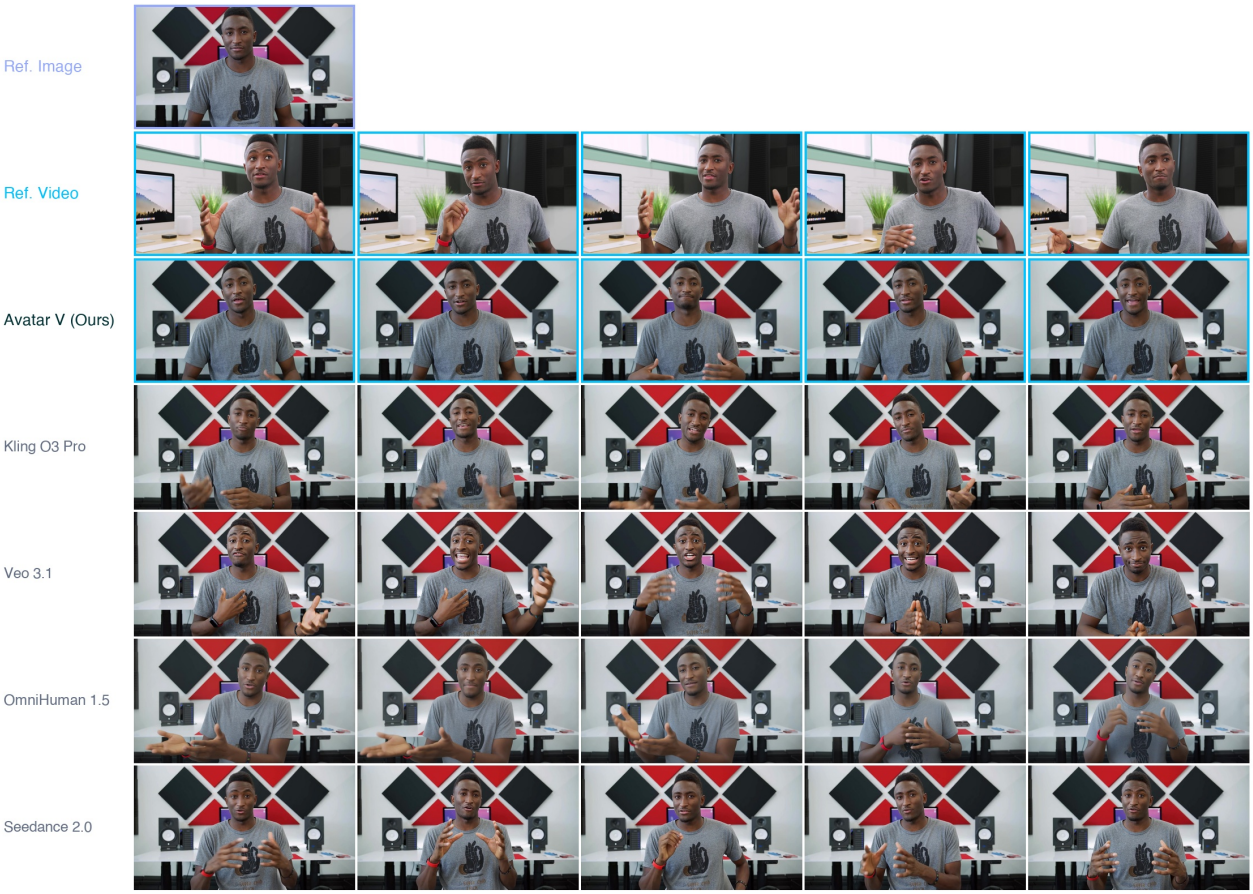

Figure 8 Qualitative comparison: cross-scene condition. The driving scene image (purple border, top-left) differs from the reference video scene, requiring cross-context identity transfer.

7.2 Objective Evaluation

We compute four automated metrics, each targeting a distinct aspect of generation quality. All per-frame metrics are sampled at 2 fps and computed as 10% trimmed means (removing the top and bottom 10% of frames) for robustness.

SyncNet score. We measure audio-visual synchronization at the video level using SyncNet [12]. The model detects face tracks across frames, computes audio-visual embeddings for each track segment, and reports the synchronization confidence (Sync-C, higher is better) and minimum distance (Sync-D, lower is better).

Face similarity. Identity preservation is measured as the cosine similarity between ArcFace [16] embeddings of detected faces in each generated frame and the reference image.

Q-Align IQA. We assess overall frame-level perceptual quality using Q-Align [53], a vision-language model that produces quality scores calibrated to human mean opinion scores.

7.2.1 Results

Avatar V achieves the strongest overall performance across all four automated metrics. On lip synchronization, Avatar V achieves the highest SyncNet confidence (8.97) and the lowest Sync-D (6.75), surpassing even ground truth recordings (7.93 / 6.76) and demonstrating superior audio-visual alignment. On identity preservation, Avatar V achieves the highest Face Similarity (0.840) among all methods, closely approaching

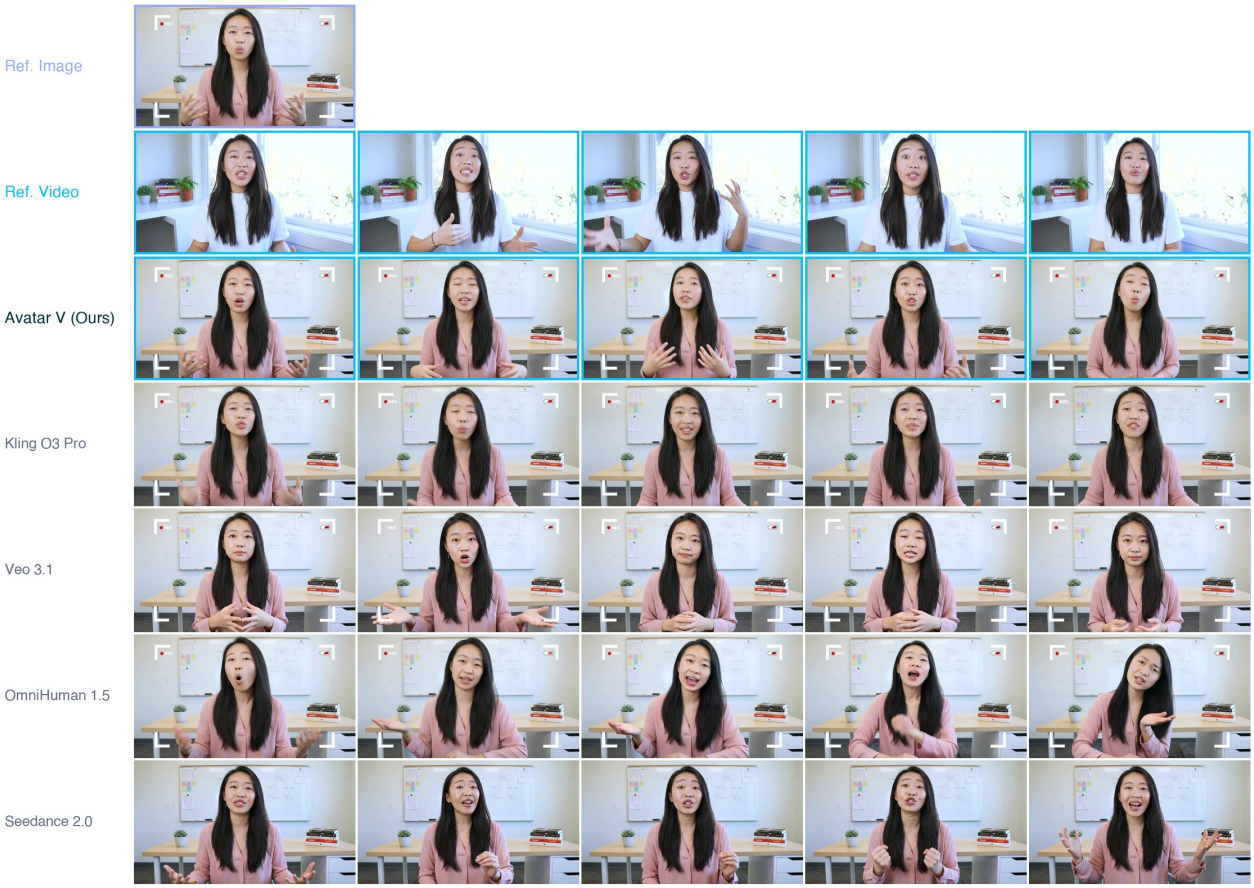

Figure 9 Qualitative comparison: generated-scene condition. The scene image (purple border, top-left) is produced by the Identity-Preserving Image Engine, representing the fully automated production pipeline.

Table 1 Objective metrics comparison on the 36 matched test cases where all five methods produced valid outputs. Per-frame metrics (Face Sim, Q-Align) are computed as 10% trimmed means for robustness. Higher is better for all metrics (↑) except Sync-D (↓). Best results are in **bold**, second best underlined.

Method	Sync-C↑	Sync-D↓	Face Sim↑	Q-Align↑
Ground Truth	7.93	6.76	0.861	4.75
Kling O3 Pro (2026)	5.16	10.07	0.838	4.80
Veo 3.1 (2025)	8.05	7.28	0.714	**4.95**
OmniHuman 1.5 (2025)	7.53	8.25	0.732	4.70
Seedance 2.0 (2026)	8.86	6.99	0.823	4.85
Avatar V (Ours)	**8.97**	**6.75**	**0.840**	4.85

ground truth (0.861) and substantially outperforming Veo 3.1 (0.714) and OmniHuman 1.5 (0.732). On Q-Align perceptual quality, Avatar V ties with Seedance 2.0 for second place (4.85). Veo 3.1 achieves the highest Q-Align score (4.95) but at the cost of severely degraded identity preservation (Face Sim = 0.714); we also observe that Veo 3.1 outputs exhibit noticeable over-sharpening, which can inflate perceptual quality scores without corresponding subjective improvement. This analysis reveals a key trade-off among competing methods: systems optimized for visual quality metrics may sacrifice identity fidelity or resort to aggressive post-processing. Avatar V uniquely maintains top-tier performance across identity, synchronization, and quality axes simultaneously.

Table 2 MOS comparison (5-point Likert scale). Higher is better. Best results are in **bold**, second best underlined.

Method	Identity↑	Lip Sync↑	Motion Nat.↑	Motion Con.↑	Artifacts↑	Visual↑
Kling O3 Pro (2026)	4.18	4.40	4.21	4.12	4.19	4.45
Veo 3.1 (2025)	4.34	4.62	3.88	4.05	4.66	4.76
OmniHuman 1.5 (2025)	4.70	4.04	3.59	3.87	3.89	3.81
Seedance 2.0 (2026)	4.84	4.64	4.13	4.44	4.61	4.17
Avatar V (Ours)	**4.98**	**4.69**	**4.48**	**4.57**	**4.75**	**4.78**

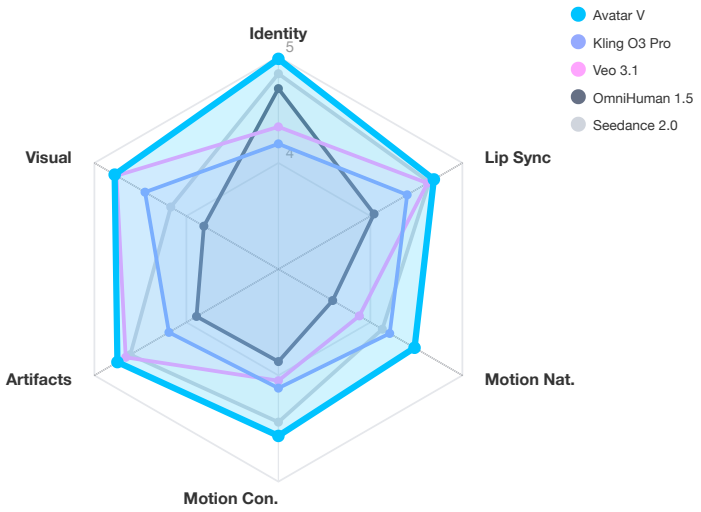

Figure 10 MOS radar chart. Avatar V (cyan) consistently achieves the highest human ratings across all six perceptual dimensions, demonstrating balanced excellence rather than specialization in a single aspect.

7.3 Subjective Evaluation

All subjective evaluations are conducted by trained human annotators following the annotation system described in Section 5.

7.3.1 Mean Opinion Score (MOS)

Each generated video is independently rated on a 5-point Likert scale (1 = very poor, 5 = excellent) across six perceptual dimensions: identity consistency, lip-sync accuracy, motion naturalness, motion consistency, artifact control, and visual quality. Each video is rated by at least two annotators, blinded to model identity, in randomized order. The final score is the arithmetic mean.

Avatar V achieves the highest MOS scores on all six dimensions (Table 2, Figure 10). On identity consistency, Avatar V scores 4.98 out of 5, nearly perfect and substantially higher than all competitors. The advantage is particularly pronounced on motion naturalness (4.48 vs. second-best 4.21) and motion consistency (4.57 vs. 4.44), reflecting the effectiveness of the dedicated motion representation in capturing individual-specific behavioral patterns. Unlike competing methods that excel on one or two aspects while underperforming on others (e.g., Veo 3.1 scores high on visual quality but low on motion naturalness; OmniHuman 1.5 preserves identity but produces artifacts), Avatar V maintains uniformly high quality across all dimensions.

7.3.2 Pairwise Win Rate

For pairwise relative evaluation, we present annotators with side-by-side video pairs: one from Avatar V and one from a competing method, both generated from identical inputs. Annotators select which video they prefer as an overall quality judgment. Each pair is rated by three annotators, and the final winner is determined by majority vote.

Table 3 Pairwise win rate. Each row shows the percentage of test cases where Avatar V is preferred (Win) or the competitor is preferred (Lose), determined by majority vote of three annotators. The number of evaluated cases varies across competitors because certain commercial APIs reject inputs containing celebrity likenesses due to portrait rights restrictions.

Competitor	Win↑	Lose↓	#Cases
Kling O3 Pro	69.6%	30.4%	69
Seedance 2.0	68.9%	31.1%	45
Veo 3.1	72.5%	27.5%	40
OmniHuman 1.5	85.7%	14.3%	70

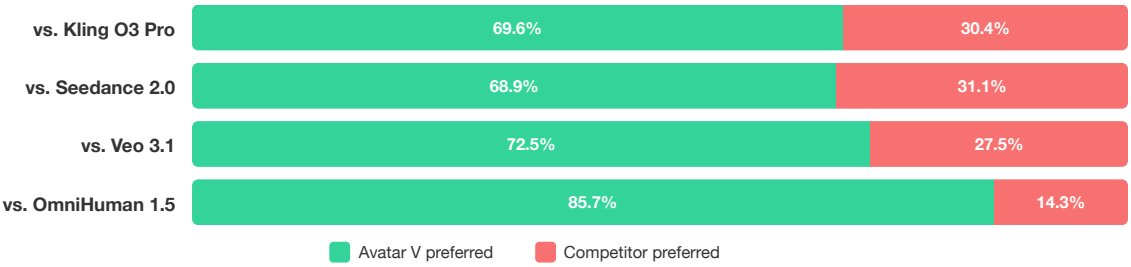

Figure 11 Pairwise win rate. Each bar shows the percentage of cases where Avatar V is preferred (green) or the competitor is preferred (red) relative to each competitor.

Table 4 Avatar Turing Test results. Three annotators each evaluated 18 pairs of Avatar V outputs and ground truth videos, yielding 54 total judgments. Lower identification accuracy indicates more realistic generation.

Metric	Value
Real identification accuracy	77.8%
Fooled rate (Avatar V mistaken as real)	22.2%
Chance level (random guessing)	50.0%
Cases fooling ≥1 annotator	11/18 (61.1%)

As shown in Table 3, Avatar V is consistently preferred across all four competitors, achieving win rates ranging from 68.9% to 85.7%. The advantage is most pronounced against OmniHuman 1.5 (85.7%) and Veo 3.1 (72.5%), while Avatar V maintains a clear lead against Kling O3 Pro (69.6%) and Seedance 2.0 (68.9%).

Qualitative analysis of annotator feedback reveals two consistent strengths of Avatar V: (1) *generation stability*, where Avatar V rarely produces visual artifacts across diverse test cases, and (2) *holistic identity consistency*, where both the subject's appearance and behavioral patterns and talking style closely match the reference video, yielding high perceived similarity under visual inspection.

In contrast, competing methods exhibit several recurring failure modes: excessive skin smoothing and over-enhancement that reduces perceived realism; poor behavioral consistency, particularly in motion dynamics and gestural patterns that deviate from the reference; and unstable generation quality with various visual artifacts across frames.

7.3.3 Avatar Turing Test

To evaluate perceptual realism, we conduct an Avatar Turing Test in which annotators are shown pairs of videos consisting of an Avatar V generation and its corresponding ground-truth recording, and are asked to determine which video is real. In this setup, a system approaching 50% identification accuracy would be considered perceptually indistinguishable from real footage.

Across all judgments, annotators correctly selected the real video 77.8% of the time, indicating that the

generated videos remain distinguishable overall. However, the results also show strong realism: in 61.1% of test cases, at least one of three annotators identified the generated video as the real one. This suggests that Avatar V is able to produce highly realistic talking-avatar videos that can frequently deceive trained evaluators on a case-by-case basis, even though a measurable gap from full perceptual indistinguishability still remains.

8 Related Work

8.1 Video Diffusion Models

Diffusion-based video generation has advanced rapidly since Sora [4] demonstrated that Diffusion Transformers (DiT) [39] can generate high-fidelity minute-long videos. A wave of large-scale models followed, including CogVideoX [60] with joint spatiotemporal 3D VAE compression, HunyuanVideo [27] as an open-source 13B model competitive with closed-source systems, Wan [45] validating scaling laws at both 14B and 1.3B scales, Movie Gen [40] unifying video and audio generation at 30B parameters, Kling [28], Lumiere [1], Step-Video [34], and SkyReels-V2 [18]. Architecturally, DiT has largely replaced U-Net as the predominant backbone [9, 35], while Flow Matching [33, 59] has emerged as a practical alternative to DDPM [23] training objectives. Open-Sora [69, 70] demonstrated commercial-grade quality at limited budget, and Cosmos [38] positioned video generation as a world foundation model platform. For controllable generation, Stable Video Diffusion [3] studied transfer from image to video diffusion, DynamiCrafter [54] animates open-domain images, and AnimateDiff [21] proposed plug-and-play motion modules compatible with personalized text-to-image models. Avatar V builds upon this DiT and flow matching foundation but diverges in two key ways: it introduces an asymmetric attention mechanism tailored for identity-conditioned generation, and it employs a progressive training pipeline that extends beyond standard text-to-video objectives to audio-driven, personality-preserving synthesis.

8.2 Portrait Video Generation

Audio-driven talking head generation. Early work such as SadTalker [65] mapped audio to 3DMM motion coefficients for decoupled head motion and expression synthesis. Diffusion-based methods brought significant quality improvements: EMO [46] pioneered direct audio-to-video generation without intermediate representations; Hallo [55] introduced hierarchical audio conditioning for separate control of lip, expression, and pose, with Hallo2 [13] extending to longer durations and higher resolutions, and Hallo3 [15] and Hallo4 [14] further advancing controllability and visual quality; EchoMimic [10] added editable landmark control, followed by EchoMimicV3 [37] which unified multiple generation tasks in a single architecture; V-Express [49] addressed weak-signal suppression in multi-condition models; and VASA-1 [57] achieved real-time generation via a latent-space DiT. More recent works have tackled multi-condition orchestration: OmniHuman [31, 32] proposed a one-stage conditioned human animation framework that scales to high-quality generation with cognitive simulation, OmniAvatar [19] proposed a unified framework for diverse avatar tasks, HuMo [8] introduced collaborative multi-modal conditioning for synchronized body and face generation, and StableAvatar [47] addressed infinite-length generation with identity consistency. A common limitation of these methods is their reliance on single-image references, which constrains identity information to one viewpoint and expression. Avatar V overcomes this by conditioning on full video references through Sparse Reference Attention, extracting rich static and dynamic identity cues without per-identity fine-tuning.

Single-image portrait animation. LivePortrait [20] achieves real-time inference through implicit keypoints with stitching and retargeting modules. AniPortrait [52] combines facial landmarks with audio for controllable portrait animation. X-Portrait [63] handles large head movements via hierarchical motion attention, and MuseTalk [68] enables efficient real-time lip synchronization through latent-space inpainting. While these single-image approaches are efficient, they fundamentally lack the dynamic identity signals (talking rhythm, habitual expressions) that video references provide.

Video-reference-based approaches. Several recent works explore conditioning generation on video references rather than single images. WanAnimate [11] extends Wan with video-guided motion transfer, SlotID [29]

uses slot attention for identity-disentangled control, and Seedance 2.0 [7] incorporates full-clip reference conditioning. However, these approaches either compress references through bottleneck encoders that discard fine-grained details, or concatenate all reference tokens incurring quadratic attention cost. Avatar V addresses both issues through its asymmetric sparse attention design, where generation tokens attend to reference tokens while reference tokens only self-attend, achieving linear complexity in reference length.

8.3 Human Body Video Generation

Pose-guided human animation. Animate Anyone [24] established the paradigm of ReferenceNet combined with Pose Guider and Temporal Attention for single-image pose-driven animation, while MagicAnimate [58] introduced an appearance encoder with video fusion for temporal consistency. Subsequent works enriched the guidance signal: Champ [71] incorporated SMPL-derived 3D conditions (depth, normals, semantics), MimicMotion [67] introduced confidence-aware pose guidance for stable long-sequence generation, and UniAnimate [51] unified reference, pose, and video in a shared latent space for minute-length synthesis. MIMO [66] extended to scene-controllable multi-character generation via spatial decomposition.

Identity preservation. Maintaining character consistency remains a core challenge. IP-Adapter [61] achieves image-text prompt compatibility through decoupled cross-attention, InstantID [50] combines identity embeddings with landmark guidance for zero-shot preservation, and PhotoMaker [30] supports multi-reference identity fusion via stacked embeddings. These techniques provide critical foundations for reference-based conditioning in human video generation. Unlike bottleneck-based identity encoders, Avatar V retains the full visual richness of reference tokens through its video-reference attention mechanism, avoiding the information loss inherent in fixed-size identity embeddings.

8.4 Training Efficiency and Alignment

Diffusion distillation. Reducing the inference cost of diffusion models has been widely studied. Progressive distillation [42] halves the number of sampling steps iteratively, while consistency models [44] learn to map any noisy sample directly to the clean output. Distribution Matching Distillation (DMD) [62] introduces a regression loss combined with an adversarial objective for few-step generation. Classifier-free guidance (CFG) distillation [36] internalizes the conditional and unconditional score combination, eliminating the need for multiple forward passes. Avatar V combines CFG distillation with DMD in a two-phase pipeline, achieving over 10× inference acceleration while maintaining generation quality.

Reinforcement learning for generative models. Aligning generative models with human preferences has drawn increasing attention. DPO [41] provides a reference-model-free approach to preference optimization, and Diffusion-DPO [48] adapts it to diffusion training. RLHF for diffusion [2] directly optimizes reward functions through policy gradients. More recently, GRPO [43], originally proposed for language models, has been adapted to visual generation: DanceGRPO [25] and FlowGRPO [64] apply group-relative advantage estimation to video diffusion, demonstrating the feasibility of multi-reward RL for visual content. Avatar V extends this line of work with identity, motion, and visual quality reward functions tailored to avatar generation, combined with DPO for complementary preference alignment.

9 Ethics and Safety

Avatar generation raises important considerations around consent and content safety. Our production platform addresses these through two mechanisms. First, creating a custom avatar requires explicit verification from the individual being represented; the depicted individual retains the right to request removal of their likeness at any time. Second, all content uploaded to or generated by the platform passes through a two-stage moderation pipeline combining automated review powered by machine learning with manual review by human moderators, covering categories including but not limited to fraud, harassment, child safety, misinformation, and intellectual property infringement. Violations may result in content removal, account suspension, or reporting to legal authorities. The full policy is available at https://www.heygen.com/moderation-policy.

10 Conclusion

We have presented Avatar V, a system for generating high-fidelity talking avatar videos from short video references. The central insight is to formulate identity conditioning as a video-reference conditioning problem: by letting the model attend directly to the full token sequence of a reference video through Sparse Reference Attention, Avatar V captures both static appearance features (facial geometry, skin texture, accessories) and dynamic behavioral patterns (talking rhythm, habitual expressions, gestural tendencies) without per-identity fine-tuning or information-lossy bottleneck encoders.

Around this core mechanism, we contribute a suite of techniques spanning model design, data, training, and deployment: a dedicated motion representation stream that creates closed-loop supervision for person-specific talking style through joint generation and conditioning; an identity-aware super-resolution refiner with sparse temporal attention that recovers fine facial details at high resolution; an LLM-based voice cloning engine that reproduces the target speaker's vocal identity from a short audio sample; a scalable data curation pipeline processing 50M+ raw videos into 100M+ pretraining clips and 10M+ avatar fine-tuning clips with cross-clip identity connectivity; a five-stage progressive training pipeline incorporating text-to-video pretraining, audio-to-video pretraining, personality SFT, two-phase distillation for over 10× inference acceleration, and reinforcement learning from human feedback; and a comprehensive inference optimization stack deployed across thousands of GPUs under a unified multi-cloud infrastructure.

Experiments on our cross-scene benchmark demonstrate that Avatar V achieves state-of-the-art performance across all evaluated dimensions, including identity preservation, lip synchronization, expression naturalness, motion quality, and visual fidelity, as measured by both automated metrics and comprehensive human evaluation.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.

[3] 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.

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. OpenAI Technical Report.

[5] ByteDance Seed. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

[6] ByteDance Seed. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[7] ByteDance Seed. Seedance 2.0: A native multi-modal video generation model for real human interactions. *arXiv preprint*, 2025.

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

[9] Shoufa Chen, Mengmeng Ge, Jiawei Qu, Jiaxin Jia, Jia-Bin Liu, Jiatao Han, Hao Yang, Zhifeng Cai, and Dahua Lin. GenTron: Diffusion transformers for image and video generation. *CVPR*, 2024.

[10] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. EchoMimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

[11] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Li, Jinwei Meng, Penchong Qi, and others Qiao. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

[12] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Asian Conference on Computer Vision (ACCV) Workshops*, 2017.

[13] Jiahao Cui, Hui Li, Yao Yao, Hanlin Shang, Kaihui Zhu, Jingdong Wang, and Siyu Zhu. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.

[14] 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.

[15] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *CVPR*, 2025.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *ICML*, 2024.

[18] Guibin Feng, Lei Yang, Zhixiang Huang, Jiaxi Li, Xiaoyu Ma, Jie Zhou, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

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

[20] 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.

[21] 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. *ICLR*, 2024.

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

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[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. *CVPR*, 2024.

[25] Zeyue Huang, Zhicheng Yu, Weilin Chen, Yixiao Zhang, Jiashi Liu, and Limin Feng. DanceGRPO: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[26] Keller Jordan, Yuchen Li, Angus Nabarro, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. *arXiv preprint arXiv:2502.16982*, 2025.

[27] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Xu, Ziwei Gu, Senyan Chen, Yatian Liu, Qianru Wang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[28] Kuaishou Technology. Kling: A text-to-video generation model. *Kuaishou Technical Report*, 2024.

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

[30] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. *CVPR*, 2024.

[31] Gaojie Lin, Jianwen Bai, Jiaqi Chen, Yanbo Zeng, Yue Wang, Yuting Ge, Heng-Da Guo, Zhe Wan, Xiang Zhang, Jingdong Liu, Ming Yang, and Ying-Cong Zhang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025.

[32] Gaojie Lin et al. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. *arXiv preprint*, 2025.

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *ICLR*, 2022.

[34] Guoqing Ma, Haoyang Li, Kun Zhang, Yufeng Zheng, Yiming Zhao, Xin Liu, et al. Step-Video-T2V: Technical report. *arXiv preprint arXiv:2502.10248*, 2025.

[35] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2025.

[36] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.

[37] Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. EchomimicV3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. *AAAI*, 2026.

[38] NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

[40] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[41] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

[42] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

[43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[44] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

[45] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[46] 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 *ECCV*, 2024.

[47] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. StableAvatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

[48] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senera Purber, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024.

[49] 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.

[50] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[51] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Wang, Changxin Gao, and Nong Sang. UniAnimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024.

[52] Huawei Wei, Zejun Yang, and Zhisheng Wang. AniPortrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

[53] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[54] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. DynamiCrafter: Animating open-domain images with video diffusion priors. *ECCV*, 2024.

[55] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

[56] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

[57] 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.

[58] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. *CVPR*, 2024.

[59] Zhijie Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Chen, Xiaohan Zhu, et al. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2025.

[60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Chen, Xiaohan Zhu, Yuxuan Zeng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

[61] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2024.

[62] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.

[63] Jianwen You, Hao Yu, Lantao Feng, Xiaoming Hu, and Yuxin Jiang. X-Portrait: Expressive portrait animation with hierarchical motion attention. *arXiv preprint arXiv:2403.15931*, 2024.

[64] Jie Zhang, Zhihong Huang, Xin Wang, and Yu Zhou. Flow-GRPO: Training flow matching models with online rl. *arXiv preprint arXiv:2505.05470*, 2025.

[65] 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*, 2023.

[66] Yifang Zhang, Hao Li, Meng Liu, Liqian Zhang, and Dacheng Tao. MIMO: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024.

[67] Yuang Zhang, Jiaxi Li, Liang Liu, Gongwei Huang, Yongping Zeng, Xiangyu Xue, Zhengjun Miao, Xiao Lu, and Siyu Zhu. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

[68] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yingjie He, Chao Zhan, and Wenjiang Yao. MuseTalk: Real-time high quality lip synchronization with latent space inpainting. *arXiv preprint arXiv:2410.10122*, 2024.

[69] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Cheng, Shenggui Wang, Ang Li, Bin Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2403.02962*, 2024.

[70] Zangwei Zheng, Tianji Yang, Xiangyu Peng, Chenhui Cheng, Shenggui Wang, Ang Li, Bin Li, and Yang You. Open-Sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025.

[71] Shenhao Zhu, Junming Xu, Zuozhuo Liu, Boyuan Qin, Yongming Ouyang, Lei Shi, Yao Yao, Ce Liu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *ECCV*, 2024.