CINEPERSONA: CAMERA-AWARE DYNAMIC PERSONA ENCODING FOR IDENTITY-PRESERVING HUMAN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cinematic human video generation must jointly control who appears in frame, what the subject does, and how the camera moves, yet the literature has split into single-axis specialists whose serial composition compounds error on every axis. We argue that the right conditioning object is a chronologically ordered space-time lattice of micro-clips cut from a reference video, and present CINEPERSONA, a camera-aware reference encoder for a frozen 14B video diffusion transformer. Its CHRONO-MOSAIC LATTICE packs nine time-ordered micro-clips into one trunk-compatible tensor, and we show that this ordering is what keeps camera direction decodable: under reversal-compatible acquisitions, shuffling removes directional information up to residual within-cell motion. A dual-pathway encoder pairs a diffusion-synchronized implicit stream with explicit persona, action and camera branches, CINE-ROPE writes the target trajectory into the trunk's rotary phases, and a tri-axis sampler exposes one guidance dial per control, all while training 1.5% of parameters. On CINEHUMAN-BENCH and OpenS2V-Eval, CINEPERSONA natively controls all three axes and achieves state-of-the-art persona similarity, action alignment and camera fidelity, the latter verified with an estimator decoupled from training, on rendered scenes with ground truth, and by 41 human raters.

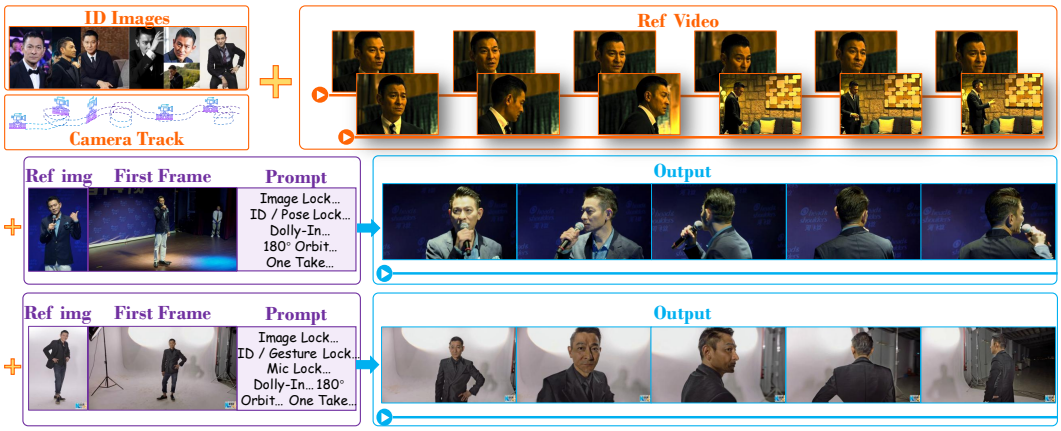

Figure 1: **Unified persona, action, and camera control with CinePersona.** Given identity images, a reference video, a target camera trajectory, a first frame, and a text prompt, CinePersona generates coherent one-take human videos that preserve the subject's facial and body identity, follow the specified pose or gesture, and faithfully realize cinematic camera motions such as dolly-in and 180° orbit.

1 INTRODUCTION

A cinematic shot of a human subject is not a still portrait. From the opening orbit that introduces a film hero to the slow dolly-in on a dancer mid-turn, directors compose three physically coupled signals: *who* appears in frame (persona, spanning face, body and clothing), *what* that subject does

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(action), and *how the camera travels around them* (trajectory). The same actor performing the same motion under a different camera path is a fundamentally different shot, and yet the video-generation literature has fragmented into single-axis specialists. Identity-preserving methods (Yuan et al., 2025b; Liu et al., 2025; Xue et al., 2026; Li et al., 2026; He et al., 2024; Zhang et al., 2025b; Wei et al., 2025; Zhong et al., 2025; Hu et al., 2025; Deng et al., 2026; Fei et al., 2025) inject static facial or subject references and tacitly assume the camera is fixed; camera-controlled methods (Wang et al., 2024b; He et al., 2025a;b; Bahmani et al., 2025a;b; Zheng et al., 2024; Bai et al., 2025a; Ren et al., 2025; Yu et al., 2025; Luo et al., 2025) deliver impressive trajectory accuracy but leave the subject anonymous and exhibit documented identity drift across novel views; human-animation systems retarget a driving motion onto a reference image while delegating the shot's camera to presets or prose (Hu, 2024; Zhu et al., 2024; Zhang et al., 2025a; Men et al., 2025; Fang et al., 2026; Kling Team et al., 2026); and generalist editing trunks unify some signals while explicitly leaving identity and camera unsolved (Jiang et al., 2025; Fei et al., 2025; Deng et al., 2026). Practitioners are therefore left to compose specialists serially, and the cost is severe: in our pilot study a Stand-In (Xue et al., 2026) to CameraCtrl-II (He et al., 2025b) cascade loses 18% of face similarity relative to Stand-In alone and adds 36% rotation error relative to CameraCtrl-II alone. Where in the pipeline the three signals should fuse has remained an open question.

We trace this fragmentation to a shared blind spot in the conditioning interface. Whether the input is a still face image, an ArcFace vector, a CLIP token or an isolated DiT prefix, the conditioning object lacks the temporal structure that cinematic control inherently requires. A short reference video of the desired subject, by contrast, already contains all three signals in densely entangled form, since the face recurs across frames, the body moves, and the surrounding camera shifts. The right object to encode is thus not a point, an image or even a single mosaic, but a *chronologically ordered space-time lattice* of nine short micro-clips: dense, because one tensor carries three signals; trunk-compatible, because it passes through the pretrained 3D VAE and patch embedding unchanged; and order-preserving, because reading the lattice in raster order traces both the action and the camera while every cell offers a fresh persona observation. Contemporary identity mosaics intentionally *shuffle* their cells to enforce permutation invariance. We make the opposite design choice, and we make the stakes precise: for reversal-compatible acquisitions, shuffling obscures direction cues in trajectory pairs such as dolly-in and dolly-out, up to residual within-cell motion (App. A.1), so chronology is exactly what keeps the directional component of camera motion decodable. To our knowledge, no prior encoder has treated a time-ordered reference lattice as the joint carrier of persona, action and camera.

Building on this observation, we present CINEPERSONA, a camera-aware reference-video encoder for human video generation built atop a frozen 14B Wan-2.1 (Team Wan et al., 2025) video DiT. Three architectural pieces work in concert. The CINEPERSONA ENCODER turns the lattice into a dual representation whose *implicit* stream re-noises the lattice on the same flow-matching trajectory as the target video, preserving rich but entangled appearance, motion and camera evidence through Wan's native 36-channel patch embedding, and whose *explicit* stream factorizes the same lattice into three geometry-grounded branches, ArcFace+DINOv2 (Deng et al., 2019; Oquab et al., 2024) for persona, DWPose (Yang et al., 2023) for action and VGGT (Wang et al., 2025)-recovered Plücker rays for camera, fused by a light Persona-Action-Camera transformer. To couple the target trajectory with the trunk's positional geometry rather than with a parallel adapter, CINE-ROPE modulates Wan's split-axis rotary phases by per-frame Plücker geometry, and this single intervention more than halves rotation error in our ablation. A TRI-AXIS GUIDANCE sampler finally factorizes classifier-free guidance into independently re-scaled text, persona and camera-action directions. Only **$\sim$205M** parameters, that is 1.5% of the trunk, are trained.

To evaluate this multi-axis setting we build CINEHUMAN-BENCH, an anonymized license-clean benchmark of 400 cases over 50 identities, each paired with one of ten canonical camera trajectories. Because VGGT supplies camera signal inside our method, we refuse to let it grade its own homework: every camera metric in this paper is computed with MegaSaM (Li et al., 2025), an estimator that never enters our pipeline, both estimators are calibrated against exact ground-truth trajectories on rendered scenes, and expert raters provide an estimator-free check. Under this protocol CINEPERSONA is the only evaluated method that natively supports all three controls jointly, and it Pareto-dominates every single-axis specialist on that specialist's home metric, while on OpenS2V-Eval Human Domain (Yuan et al., 2025a), the same checkpoint tops all metrics, with over 61% preference across all axes in a 41-participant study.

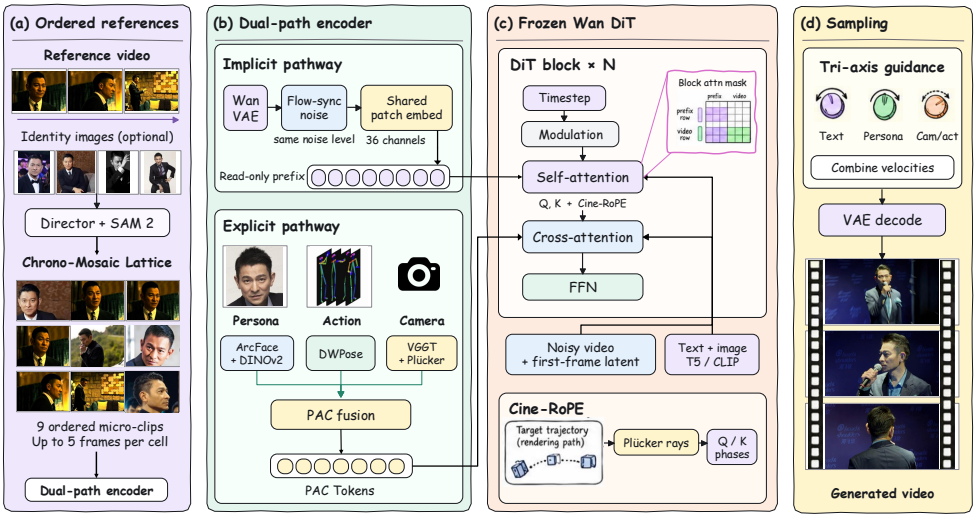

Figure 2: **Overview of CinePersona.** Ordered references feed a flow-synchronized implicit prefix and parallel persona, action, and camera branches. The prefix supplies read-only self-attention context; PAC tokens condition cross-attention, while Cine-RoPE modulates the frozen Wan DiT with camera rays. The tri-axis sampler combines the controls before VAE decoding.

2 Related Work

Identity-preserving video generation. Personalization spans per-identity tuning (Ruiz et al., 2023; Gal et al., 2023; Hu et al., 2022), tuning-free global identity vectors (Ye et al., 2023; He et al., 2024; Yuan et al., 2025b; Zhang et al., 2025b; Wei et al., 2025), reference-latent prefixes (Zhong et al., 2025; Liu et al., 2025; Li et al., 2026; Xue et al., 2026), generalist editing trunks (Jiang et al., 2025; Fei et al., 2025; Hu et al., 2025; Deng et al., 2026), and motion-driven animation (Hu, 2024; Zhu et al., 2024; Zhang et al., 2025a; Men et al., 2025; Fang et al., 2026; Kling Team et al., 2026; Yuan et al., 2025a). Throughout, the camera is a byproduct rather than a metric trajectory, and reference encoders are either implicit and not queryable per axis (Xue et al., 2026; Li et al., 2026) or explicit and starved of dynamic evidence (Yuan et al., 2025b; Ye et al., 2023). We read persona from a chronological lattice carrying action and camera jointly.

Camera control. The dominant recipe injects Plücker rays through adapters (Wang et al., 2024b; He et al., 2025a;b; Bahmani et al., 2025a;b), with epipolar (Zheng et al., 2024) or geometry-prior (Ren et al., 2025; Bai et al., 2025a; Yu et al., 2025) variants; CamCloneMaster clones reference motion because pose estimation on dynamic footage is unreliable (Luo et al., 2025), motivating our estimator-decoupled evaluation (§4.3), yet leaves the subject anonymous. Identity drift and adapter collapse on large DiTs recur (Bahmani et al., 2025b; He et al., 2025b). Since video DiTs already encode geometry in split-axis rotary phases (Su et al., 2024; Team Wan et al., 2025), we deform them directly, zero-initialized (Zhang et al., 2023), and read cameras via feed-forward VGGT (Wang et al., 2025) rather than matching/BA pipelines (Wang et al., 2024a; Leroy et al., 2024; Li et al., 2025; Schönberger & Frahm, 2016) unsuited to a 14B inner loop.

3 Method

We first review the Wan flow-matching DiT and its split-axis 3D RoPE, then introduce the Chrono-Mosaic Lattice, the CinePersona Encoder, Cine-RoPE, and Tri-Axis Guidance together with the training protocol. Throughout, the Wan 2.1 14B trunk remains frozen and only **∼205M** encoder parameters are trained; all derivations, hyperparameters and extended sweeps live in App. A–App. F.

3.1 PRELIMINARIES

Wan 2.1 (Team Wan et al., 2025) is a 14B-parameter Diffusion Transformer trained with the Flow Matching objective (Lipman et al., 2023). Its 3D VAE encodes a video $V \in \mathbb{R}^{3\times T\times H\times W}$ into a clean latent $z_0^v \in \mathbb{R}^{16\times \widetilde{T}\times \widetilde{H}\times \widetilde{W}}$, and for image-to-video conditioning it composes a 36-channel tensor $\Phi_t^v = [z_t^v \parallel m^v \parallel y^v]$ in which $z_t^v = (1-\sigma_t)z_0^v + \sigma_t \boldsymbol{\epsilon}^v$ is the noisified latent, m^v is a temporal mask and y^v is the VAE-encoded first-frame condition. A 3D patch embedding $\mathcal{E}_{\text{patch}}$ projects Φ_t^v into a token sequence, and each attention head splits its head dimension d into three independent one-dimensional rotary axes,

$$d_f = d - 2\lfloor d/3 \rfloor, \quad d_h = d_w = \lfloor d/3 \rfloor, \tag{1}$$

carrying rotary frequencies along the temporal, height and width directions respectively (Su et al., 2024; Team Wan et al., 2025). This split-axis structure is the substrate that CINE-ROPE will exploit.

3.2 CHRONO-MOSAIC LATTICE

The CHRONO-MOSAIC LATTICE folds a reference clip, or a small bank of clips, into a 3×3 space-time tensor whose cells are short, time-ordered micro-clips of the subject, the square side $g{=}3$ attaining the highest measured quality among $g \in \{1,2,3,4\}$ in the sweep of App. B.1 for the reasons set out in App. A.2. Concurrent identity-mosaic methods deliberately shuffle cells to enforce permutation invariance for identity. We keep acquisition-time order instead, and the following statement, proved in App. A.1, makes precise which signal the shuffle forfeits.

Under the observation-regularity and pose-readout assumptions in App. A.1, shuffling nine micro-clips makes a time-reversed trajectory pair distinguishable only through its residual within-cell motion: if that motion is bounded by ε, every decoder has error at least $\frac{1}{2} - O(\varepsilon)$. Chronological order instead recovers an L-Lipschitz camera curve to $O(\eta + L\Delta)$, where η is the pose-readout error and Δ the bin width.

The persona channel is permutation-invariant by construction and the action channel survives up to direction reversal, so the shuffle taxes camera first and hardest. We state the scope plainly: the chronology argument bounds what *any* order-invariant interface can decode in principle, and does not by itself quantify the loss of a particular trained network. The ablation in Tab. 4 supplies the complementary empirical evidence, where shuffling multiplies rotation error by $3.4\times$ while persona barely moves, and where the confusion concentrates on exactly the time-reversed trajectory pairs the proposition predicts (App. B.8).

Director-based cell mining and augmentation. A frozen Qwen3-VL (Bai et al., 2025b) *Mosaic Director* partitions the timeline into nine bins, scores candidate micro-clips within each bin along seven axes including a *camera-informativeness* term over the in-bin optical flow that up-weights bins with strong camera motion, and selects one $k_{\text{raw}}{=}5$-frame micro-clip per bin, cropped to a square face-centered window via GVHMR (Shen et al., 2024) and DWPose (Yang et al., 2023) (full template in App. C). A tiered augmentation pipeline (App. D) combines generative counterfactuals on background, accessories and props with low-level perturbations, while a SAM-2 (Ravi et al., 2025) face mask protects the identity-defining region throughout. The result is a lattice $\mathcal{G} \in \mathbb{R}^{3\times k_{\text{raw}}\times 3H_{\text{id}}\times 3W_{\text{id}}}$.

3.3 CINEPERSONA ENCODER

Implicit pathway. The lattice is encoded by the frozen Wan VAE into $z_0^{\text{id}} \in \mathbb{R}^{16\times n\times \widetilde{H}_{\text{id}}\times \widetilde{W}_{\text{id}}}$ with $n{=}2$, and the latent is then noisified *on the same flow-matching path as the target video*,

$$z_t^{\text{id}} = (1-\sigma_t)\, z_0^{\text{id}} + \sigma_t\, \boldsymbol{\epsilon}^{\text{id}}, \quad \boldsymbol{\epsilon}^{\text{id}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

sharing a generator with $\boldsymbol{\epsilon}^v$. Synchronization places reference and video tokens at the same flow noise level. The analysis in App. A.2 compares equal-dimensional latent patches through an auxiliary Gaussian coupling, rather than subtracting the differently shaped full tensors or assuming identical runtime noise draws. The clean-prefix variant incurs a 22.8% drop in the reported ablation aggregate (Tab. 4). We assemble a Wan-isomorphic 36-channel tensor $\Phi_t^{\text{id}} = [z_t^{\text{id}} \parallel \mathbf{1}_4^{\text{id}} \parallel z_0^{\text{id}}]$, whose all-one mask declares a known reference, and project it through the *same* patch embedding $\mathcal{E}_{\text{patch}}$. The resulting implicit tokens X_t^{id} are prefix-concatenated with the video tokens, and because they share channel semantics with X_t^v no auxiliary adapter is required.

Explicit pathway: three geometry-grounded branches. While the implicit pathway preserves rich but entangled information, the explicit pathway exposes per-axis control by extracting three disentangled descriptors from the same lattice. The persona branch runs ArcFace-R100 (Deng et al., 2019) for a global identity vector and DINOv2-Large (Oquab et al., 2024) for patch-level body tokens on the SAM-2-segmented body region, fused by an MLP into $\boldsymbol{P}_{\text{prs}} \in \mathbb{R}^{N_{\text{prs}} \times d}$. The action branch extracts 134 keypoints per frame with DWPose (Yang et al., 2023) and encodes the sequence with a temporal 1D-CNN into $\boldsymbol{P}_{\text{act}} \in \mathbb{R}^{N_{\text{act}} \times d}$. The camera branch runs VGGT (Wang et al., 2025) per cell for extrinsics (R_j, t_j) and intrinsics K_j, and synthesizes per-pixel Plücker rays

$$\boldsymbol{\rho}_j(u,v) = \big(\boldsymbol{d}_j(u,v),\, \boldsymbol{m}_j(u,v)\big) \in \mathbb{R}^6, \tag{3}$$

with direction $\boldsymbol{d}_j(u,v) = R_j K_j^{-1}[u,v,1]^\top / \|\cdot\|$ and moment $\boldsymbol{m}_j = t_j \times \boldsymbol{d}_j$, after which a ConvNet compresses the dense map into $\boldsymbol{P}_{\text{cam}} \in \mathbb{R}^{N_{\text{cam}} \times d}$; the branch is agnostic to the particular geometry reader, which we exploit for evaluation in §4.3. The pixel-aligned, view-uniform Plücker parameterization alone removes $0.94°$ of rotation error in our ablation.

Persona-Action-Camera fusion. The three branches are concatenated with axis-tagged type embeddings and refined by a two-layer cross-attention transformer,

$$\boldsymbol{P} = \mathcal{T}_{\text{PAC}}\big([\boldsymbol{P}_{\text{prs}} \,\|\, \boldsymbol{P}_{\text{act}} \,\|\, \boldsymbol{P}_{\text{cam}}]\big) \in \mathbb{R}^{N_p \times d}, \tag{4}$$

injected into the DiT through a lightweight cross-attention adapter on every block. Crucially $\boldsymbol{P}$ does not replace the implicit prefix: the prefix carries dynamic visual evidence whereas the explicit tokens carry queryable per-axis control, and Tab. 4 shows that removing either pathway costs more than that pathway contributes in isolation.

Read-only negative-time anchoring. We assign the implicit prefix tokens negative temporal coordinates $\{-n, \ldots, -1\}$ while video tokens occupy $\{0, \ldots, \widetilde{T} - 1\}$, and a block attention mask makes the prefix read-only: video tokens read both prefix and video keys, while prefix tokens read only themselves and never absorb generated content, avoiding the contamination otherwise observed in long-horizon prefix attention.

3.4 CINE-ROPE

We now embed the camera trajectory into the trunk's positional geometry instead of an external adapter. For axis $a \in \{f, h, w\}$, RoPE applies a block-diagonal rotation whose blocks rotate by frequencies determined by the integer position p_a, so that attention sees only relative offsets. Given the camera Plücker ray $\boldsymbol{\rho}(f,h,w) = (\boldsymbol{d}, \boldsymbol{m})$ at a latent grid point, we replace the scalar position by a *camera-modulated* position

$$\widetilde{p}_a(f,h,w) = p_a + \alpha_a \cdot \psi_a\big(\boldsymbol{\rho}(f,h,w)\big), \tag{5}$$

where the per-axis modulator projects the ray onto axis-aligned scalars,

$$\begin{aligned} \psi_f(\boldsymbol{\rho}) &= \boldsymbol{d}^\top \bar{\boldsymbol{d}} && \text{(rotation: cosine to bin-mean direction)}, \\ \psi_h(\boldsymbol{\rho}) &= (\boldsymbol{m})_y && \text{(translation: vertical moment)}, \\ \psi_w(\boldsymbol{\rho}) &= (\boldsymbol{m})_x && \text{(translation: horizontal moment)}. \end{aligned} \tag{6}$$

The learnable scales α_a are zero-initialized so that the model strictly recovers standard RoPE at initialization (Zhang et al., 2023), and their learned depth profile is analyzed in App. E.1. Implicit prefix tokens use each cell's VGGT-recovered pose while main video tokens use the user-specified target trajectory.

At the bin-centered expansion pose $\boldsymbol{t} = \boldsymbol{0}$, with the reference direction $\bar{\boldsymbol{d}}$ held fixed, a small left camera perturbation $\xi = (\boldsymbol{\omega}, \Delta t)$ changes each rotary position to first order as

$$\Delta\widetilde{p}_a = \alpha_a \langle \mathbf{J}_a, \xi \rangle + O(\|\xi\|^2), \qquad \mathbf{J}_f = (\boldsymbol{d} \times \bar{\boldsymbol{d}}, \mathbf{0}_3), \quad \mathbf{J}_{h,w} = (\mathbf{0}_3, \boldsymbol{d} \times \boldsymbol{e}_{y,x}).$$

The temporal modulator responds to rotation and the moment-based modulators to translation at this expansion point; setting $\alpha_a = 0$ recovers standard 3D RoPE exactly.

The expansion is worked out in App. A.3. Its practical reading is that we modulate exactly three scalars because each axis then disentangles a complementary subspace of $\mathfrak{se}(3)$ that matches the

Method	Persona		Action		Camera			Quality	
	Face-Sim↑	Body-Sim↑	Pose-Sim↑	MotSm↑	RotErr↓	TransErr↓	CamTrajSim↑	FVD↓	NatSc.↑
Identity-only baselines (no native camera control)									
VACE-14B (Jiang et al., 2025)	0.624	0.481	0.682	0.971	5.84	0.221	0.482	271.5	0.745
Phantom-14B (Liu et al., 2025)	0.531	0.476	0.694	0.974	6.12	0.232	0.473	296.8	0.731
Stand-In (Xue et al., 2026)	0.748	0.512	0.701	0.969	5.71	0.218	0.491	285.2	0.752
BindWeave (Li et al., 2026)	0.685	0.524	0.708	0.965	5.62	0.214	0.498	274.8	0.738
ConsisID (Yuan et al., 2025b)	0.598	0.418	0.652	0.953	6.41	0.247	0.451	312.4	0.701
Camera-only baselines (single-image identity)									
CameraCtrl (He et al., 2025a)	0.396	0.342	0.514	0.962	2.07	0.094	0.842	318.7	0.683
CameraCtrl-II (He et al., 2025b)	0.412	0.358	0.532	0.961	1.72	0.081	0.873	302.4	0.704
AC3D (Bahmani et al., 2025a)	0.421	0.361	0.546	0.958	1.86	0.087	0.861	308.7	0.692
CamI2V (Zheng et al., 2024)	0.404	0.351	0.524	0.960	1.88	0.089	0.858	309.5	0.694
ReCamMaster (Bai et al., 2025a)	0.452	0.398	0.585	0.960	1.78	0.079	0.870	296.2	0.715
CamCloneMaster (Luo et al., 2025)	0.438	0.385	0.571	0.963	1.65	0.077	0.876	293.8	0.712
Closed-source commercial systems									
Kling-1.6	0.612	0.478	0.672	0.972	4.86	0.182	0.571	274.6	0.751
Kling-MotionControl (Kling Team et al., 2026)	0.605	0.472	0.668	0.971	3.28	0.136	0.694	276.8	0.749
Hailuo	0.634	0.492	0.681	0.974	5.04	0.191	0.556	281.2	0.748
Pika-2.1	0.548	0.421	0.638	0.968	5.52	0.214	0.518	292.4	0.736
Serial cascade									
Stand-In → CameraCtrl-II	0.612	0.471	0.642	0.954	2.34	0.108	0.812	314.6	0.722
CINEPERSONA (Ours)	**0.812**	**0.728**	**0.783**	**0.976**	**1.42**	**0.064**	**0.892**	**248.6**	**0.798**
Δ over best baseline	**+0.064**	**+0.204**	**+0.075**	**+0.002**	**−0.23**	**−0.013**	**+0.016**	**−22.9**	**+0.046**

Table 1: **Comparison on CINEHUMAN-BENCH.** All camera metrics use MegaSaM (§4.3); **green**=best, orange=second; all gains are significant at $p < 0.001$ (App. B.3).

split of Eq. (1), so camera-induced parallax enters attention as an additive positional offset without disturbing the trunk's inductive bias. Switching CINE-ROPE off while keeping token-side camera conditioning more than doubles RotErr from 1.42° to 3.02° and leaves persona metrics nearly intact, direct evidence that the camera signal is transported through the rotary phases rather than through token content.

3.5 TRI-AXIS GUIDANCE AND TRAINING

Three-stage training. Stage 1 pre-trains the three explicit branches and the PAC fuser on subject-image pairs with reconstruction, ArcFace and DINOv2 losses. Stage 2 activates the implicit pathway and the joint cross-attention on 81-frame clips with the standard flow-matching objective, while VGGT poses are mined on the fly and supervise only the explicit camera branch. Stage 3 activates CINE-ROPE with zero-initialized α_a and adds a self-consistency Plücker loss $\|\boldsymbol{\rho}^{\text{gen}} - \boldsymbol{\rho}^{\text{tgt}}\|_2^2$ in which $\boldsymbol{\rho}^{\text{gen}}$ is VGGT-extracted from the partially denoised generation. Because this loss reuses the same reader that powers the camera branch, every camera metric in §4 is computed with an estimator that never enters training (§4.3). We deliberately avoid cross-pair supervision, which is scarce, biased and expensive, and which the chronological mosaic together with the counterfactual cocktail renders unnecessary; substituting 2K mined cross-pair samples costs 5.7 CINESCORE points (App. F). The full training step in algorithmic form, schedules and loss traces are given in App. E.1.

TRI-AXIS GUIDANCE. At inference we expose three guidance scales acting on the text, persona and camera-action axes independently. Let $\Delta_a = u_{+++} - u_{-a}$ denote the difference between the fully conditioned velocity and the velocity with axis a nulled by a shape-preserving zero tensor, matching the 10%-drop training distribution. The final velocity reads

$$\widehat{u}_t = u_{+++} + s_T \Delta_T + s_I(\sigma_t)\rho_t^I \Delta_I + s_C(\sigma_t)\rho_t^C \Delta_C \tag{7}$$

with middle-window schedules $s_I(\sigma_t)$, $s_C(\sigma_t)$ and a per-step normalization ρ_t^a that self-calibrates to the local scale of text guidance (App. E.2). Per-condition guidance scales were pioneered for camera-controlled image-to-video by CamI2V (Zheng et al., 2024); our sampler differs in the σ_t-windowed schedules, the per-step self-normalization, and the observation that Δ_I and Δ_C vary slowly along the trajectory, so the two auxiliary branches may be refreshed every k steps, turning the nominal four network evaluations per step into $2 + 2/k$ and recovering almost all of the benefit at k=2 (App. L).

Method	Total↑	FaceSim↑	NexusSc.↑	NatSc.↑
Hailuo	60.20	57.79	46.31	74.52
Kling-1.6	59.13	41.02	45.92	78.28
Pika-2.1	57.86	48.62	44.05	76.31
VACE-14B (Jiang et al., 2025)	58.57	64.65	44.20	74.33
Phantom-14B (Liu et al., 2025)	58.69	55.02	37.43	68.33
Stand-In (Xue et al., 2026)	58.36	67.31	47.21	73.05
BindWeave (Li et al., 2026)	57.61	53.71	46.84	66.85
MAGREF (Deng et al., 2026)	56.92	58.14	43.86	69.42
SkyReels-A2 (Fei et al., 2025)	55.24	59.31	42.18	66.14
HunyuanCustom (Hu et al., 2025)	55.85	62.25	44.72	67.00
CINEPERSONA (Ours)	**65.04**	**73.42**	**52.18**	**80.16**

Table 2: **OpenS2V-Eval Human Domain.** Camera-passive evaluation with a first-frame condition only, isolating the identity axis. The table retains the overall Total score and the three reported component scores.

Method	MegaSaM RotErr↓	MegaSaM TransErr↓	VGGT RotErr↓	VGGT TransErr↓
CameraCtrl-II	1.72	0.081	1.74	0.082
ReCamMaster	1.78	0.079	1.80	0.080
CamI2V	1.88	0.089	1.91	0.090
CamCloneMaster	1.65	0.077	1.68	0.078
Stand-In→CameraCtrl-II	2.34	0.108	2.37	0.110
CINEPERSONA (Ours)	**1.42**	**0.064**	**1.35**	**0.061**
Estimator floor on synthetic ground truth				
MegaSaM vs. exact GT	0.11	0.007	–	–
VGGT vs. exact GT	–	–	0.19	0.012

Table 3: **Estimator-decoupled camera fidelity.** MegaSaM is excluded from training. The bottom block gives synthetic-ground-truth floors; both estimators preserve the ranking (App. B.7).

4 EXPERIMENTS

4.1 SETUP

We evaluate on OpenS2V-Eval Human Domain (Yuan et al., 2025a) and on our CINEHUMAN-BENCH, 400 tri-modally annotated cases over 50 license-clean identities (Xie et al., 2022; Zhang et al., 2021; Zhu et al., 2022), each pairing a motion prompt with one of ten canonical camera trajectories (construction in App. G). Nine metrics cover persona (Face-Sim, Body-Sim (Deng et al., 2019; Oquab et al., 2024)), action (Pose-Sim, motion smoothness), camera (RotErr, TransErr, CamTrajSim) and quality (FVD (Unterthiner et al., 2018), NaturalScore), summarized by CINESCORE, their normalized mean on $[0, 100]$ (App. B.4); crucially, all camera metrics are computed with MegaSaM (Li et al., 2025), which is fully decoupled from our training, rather than with VGGT, which is not (§4.3). Baselines span identity-only (Jiang et al., 2025; Liu et al., 2025; Xue et al., 2026; Li et al., 2026; Yuan et al., 2025b), camera-only (He et al., 2025a;b; Bahmani et al., 2025a; Zheng et al., 2024; Bai et al., 2025a; Luo et al., 2025), closed-source (Kling-1.6, Kling-MotionControl (Kling Team et al., 2026), Hailuo, Pika-2.1) and the serial cascade Stand-In→CameraCtrl-II; all open baselines are re-run under matched settings, every number averages three seeds, and full protocols, including why driving-video animators (Men et al., 2025; Fang et al., 2026) are discussed rather than tabulated, are in App. H and App. B.2.

4.2 COMPARISON WITH THE STATE OF THE ART

Tab. 1 confirms the central claim of §1: CINEPERSONA beats every single-axis specialist on that specialist's home metric rather than trading one axis for another, while the serial cascade compounds errors exactly as predicted, losing both identity and camera accuracy relative to its own stages. Tab. 2 shows the joint formulation does not tax the identity-only operating point, and Fig. 3 visualizes the complementary failure modes our unified encoder avoids.

4.3 CAMERA FIDELITY WITHOUT GRADING OUR OWN HOMEWORK

Since VGGT feeds both our camera branch and the Stage-3 loss, VGGT-based metrics could reward agreement with its geometric prior rather than genuine camera following (Luo et al., 2025). Tab. 3 therefore referees all camera numbers with the training-decoupled MegaSaM (Li et al., 2025), calibrates both estimators against exact synthetic ground truth, and deliberately exposes the expected mild self-coupling, ours being the only method scoring slightly better under VGGT; rankings and significance are unchanged under either instrument, and estimator-free expert raters agree (§4.5). Fig. 5b adds that the orbits, demanding the largest unseen-view extrapolation, are hardest for every method and our margin is widest exactly there, which is what the chronological lattice was built to buy.

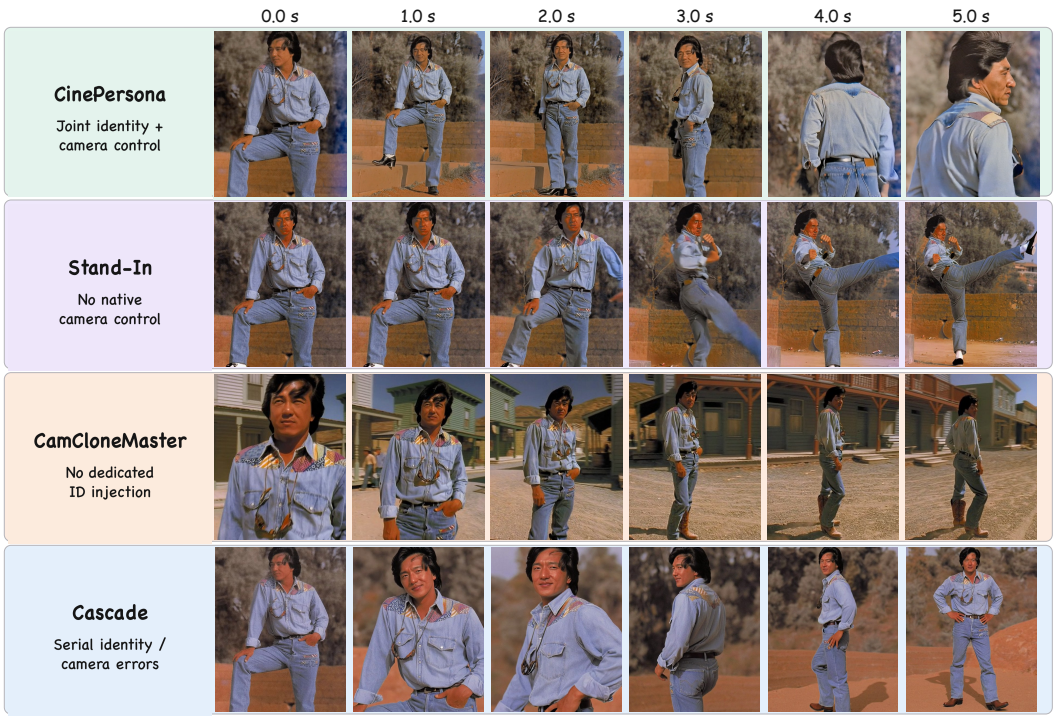

Figure 3: **Qualitative comparison under a compound camera instruction.** Native frames at 0–5 seconds retain their original aspect ratios. CinePersona preserves the denim appearance through the rear-view change; the baselines show unrequested body motion, a changed setting, or a missed terminal rear view. Prompt: Generate video based on the first frame and subject, modify camera movement: quickly pull back to show the full body, then pull back in and rotate the camera to the back of the character.

Fig. 4 traces the mechanism behind the persona gain: the read-only chronological prefix roughly halves frame-wise identity drift because fixed persona observations stay available at every step instead of the model re-encoding its own drifting output; the gains persist on uncurated in-the-wild references (App. B.5).

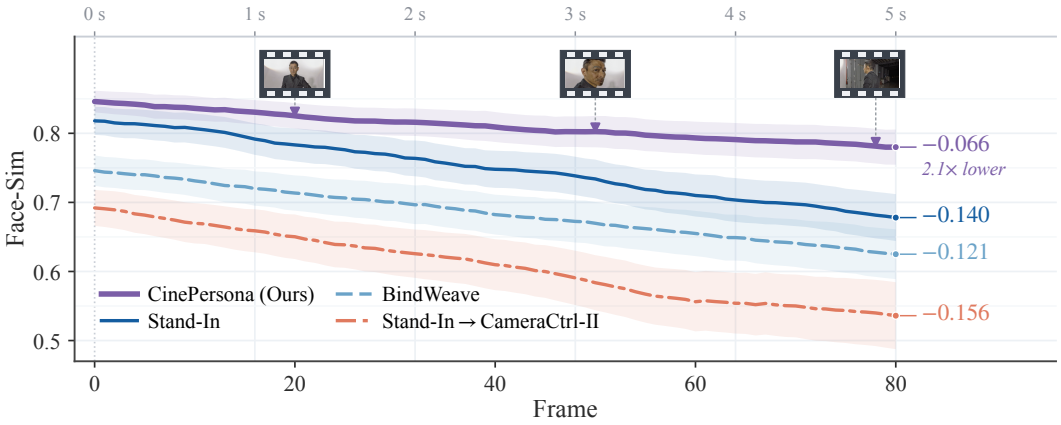

Figure 4: **Frame-wise identity drift on CINEHUMAN-BENCH.** Per-frame face similarity with $\pm 1\sigma$ bands; right-end labels give total drift, which our prefix roughly halves.

4.4 ABLATION STUDIES

Each variant in Tab. 4 breaks exactly where the corresponding claim predicts: shuffling the cells destroys camera control while persona barely moves, the operational signature of the chronology

argument (App. B.8); the two single-pathway variants fail in complementary ways and the unsynchronized clean prefix collapses outright, validating the dual design and the alignment argument of §3.3. Extended variants and training dynamics are in App. B.6 and App. E.1.

Variant	FaceSim↑	PoseSim↑	RotErr↓	FVD↓	ΔAvg
CINEPERSONA (Full)	**0.812**	**0.783**	**1.42**	**248.6**	–
shuffled cells	0.794	0.728	**4.86**	286.5	**−21.1%**
single-image ref.	0.674	0.682	3.21	312.4	**−19.4%**
implicit only	0.762	0.654	**4.42**	268.4	**−19.8%**
explicit only	0.703	0.741	2.18	304.2	**−11.7%**
clean prefix (unsync.)	0.652	0.674	**4.62**	318.4	**−22.8%**
w/o CINE-ROPE	0.802	0.768	3.02	261.8	**−9.6%**
raw (R, t), no Plücker	0.806	0.774	2.36	254.2	**−5.7%**
w/o TRI-AXIS GUIDANCE (CFG)	0.794	0.768	1.74	254.6	**−3.8%**
cross-pair aug.	0.751	0.748	2.13	271.8	**−7.6%**

Table 4: **Compact ablation on CINEHUMAN-BENCH.** Each variant removes one component; "ΔAvg" is the mean relative drop across the four axes. Full 22-variant table in App. B.6.

Disabling CINE-ROPE doubles rotation error while leaving persona intact, localizing the camera signal in the rotary phases as the phase expansion anticipates.

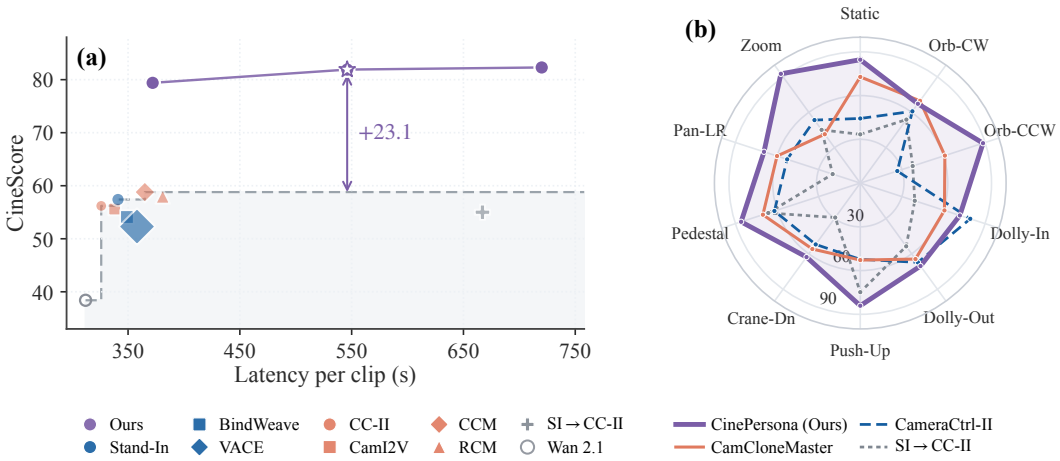

Figure 5: **(a)** Latency versus CINESCORE (bubble area = trainable parameters): all CINEPERSONA operating points sit far above the specialist frontier. **(b)** Per-trajectory rotation error: our margin is widest on the hardest orbit families.

4.5 EFFICIENCY AND HUMAN PREFERENCE

Tri-modal control uses 1.5% trainable parameters and 7.9% additional tokens. Vanilla CFG takes 372 s, 9.4% above the frozen trunk; default k=2 guidance takes 546 s, versus 719 s for full guidance and 667 s for the cascade (Fig. 5a; App. L). Human judgment corroborates the metrics: professional editors prefer our camera and persona over the respective specialists in well over 70% of double-blind pairs, and lay-viewer radar profiles are dominated by no baseline on any axis (App. I).

5 CONCLUSION

We introduced CINEPERSONA, which reformulates reference conditioning as a chronologically ordered space-time lattice read by a dual-pathway encoder, with CINE-ROPE folding the target trajectory into the frozen trunk's rotary phases. Under an evaluation protocol decoupled from the estimator used in training, it achieves state-of-the-art persona, action and camera control simultaneously. We hope this recipe of turning entangled reference material into disentangled, queryable control proves useful beyond human-centric generation.

REPRODUCIBILITY STATEMENT

Every component needed to rebuild CINEPERSONA is specified outside the main text at implementation granularity. App. E.1 records the three-stage training protocol in algorithmic form together with per-stage optimizer settings, learning-rate and noise schedules, EMA decay, loss weights, the parameter budget that yields the reported 1.5% trainable fraction, and the hardware cost of each stage; App. E.2 does the same for inference, giving the middle-window guidance schedules, the per-step normalization and the k-step refresh in pseudo-code. The conditioning interface is reproducible without guesswork because App. C reproduces the Mosaic Director prompt verbatim along with its seven scoring axes and its failure handling, and App. D lists the augmentation tiers, editor prompts and perturbation ranges. CINEHUMAN-BENCH is described constructively in App. G, covering identity sourcing and licensing, the anonymization procedure, the ten trajectory generators and the prompt templates, while App. H gives the checkpoint, resolution and step count used for every baseline, the camera-copy protocol applied to identity-only methods, the single-frame identity protocol applied to camera-only methods and the wiring of the serial cascade, so that the comparison can be reconstructed rather than merely trusted. The evaluation instruments are pinned down in App. B.7, which states the MegaSaM configuration, the Blender protocol behind the synthetic ground-truth calibration and the resulting estimator floors, and in App. B.4, which freezes the constants of CINESCORE; App. B.2 and App. B.3 give the three-seed statistical protocol, the bootstrap intervals and the corrected significance tests behind every claimed gain, and App. L documents the timing harness used for the latency and cost numbers. Proofs of the chronology argument and the phase expansion, with their assumptions stated explicitly and their scope delimited, are given in App. A.1 and App. A.3. All third-party components we build on are publicly released checkpoints, and the encoder implementation together with the benchmark harness will be released.

ETHICS STATEMENT

This work generates video of human subjects, so both the data it consumes and the artifacts it produces carry ethical weight. Every identity in CINEHUMAN-BENCH is drawn from datasets released for research use with documented licensing (Xie et al., 2022; Zhang et al., 2021; Zhu et al., 2022); we did not scrape new footage of private individuals, we redistribute only derived annotations rather than raw identity media, and the anonymization and consent policy is stated in App. G and App. J. The 41-participant preference study was voluntary, collected no personally identifying information beyond coarse expertise level, compensated participants at or above local minimum wage, and presented only material drawn from the same license-clean pool; its recruitment procedure, questionnaires and agreement statistics are reported in App. I. The capability itself is dual-use, since a system that holds a specific person's face and body stable under arbitrary camera motion lowers the cost of producing convincing non-consensual imagery, and the camera axis makes such material harder to dismiss as synthetic precisely because the parallax is geometrically consistent. We therefore treat identity fidelity as a capability to be released carefully rather than an unqualified good: the released artifacts are gated behind an acceptable-use agreement that forbids generating identifiable people without their consent, generated material is watermarked, and App. J discusses residual risks, the demographic coverage of the benchmark and the resulting fairness limits, since persona metrics computed on a benchmark of fifty identities cannot certify uniform behaviour across skin tones, ages or body types. We see no conflict of interest or sponsorship that would bias the reported comparisons.

AI USE STATEMENT

The authors used generative AI tools in a limited and supervised capacity during the preparation of this work, and disclose that use here. Large language models were used to assist with software engineering that surrounds the research contribution, specifically the scaffolding of the evaluation harness, the plotting scripts behind the figures, and routine data-wrangling utilities; all such code was read, corrected where necessary, and validated against manually computed reference values by the authors before any number in this paper was produced from it. Language models were also used for language editing of author-written prose, including grammar correction, tightening of overlong sentences and LaTeX formatting of tables, and the authors verified that no such edit altered a claim, a number or an attribution. Generative AI was not used to originate the research idea, to design the CHRONO-MOSAIC LATTICE or CINE-ROPE, to construct the proofs of the chronology argument and the phase expansion, which were derived and checked by the authors, to retrieve or summarize

related work, or to produce any experimental result, table entry or qualitative sample reported here. Separately, and distinctly from authoring assistance, a frozen vision-language model serves as a component of the proposed method itself in the role of Mosaic Director described in §3.2, and a text-to-image editor is used inside the augmentation pipeline of App. D; both are documented as part of the technical contribution rather than as writing aids. The authors reviewed all AI-assisted work and take full responsibility for the final content of this paper, including its text, its claims and its artifacts.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. AC3D: Analyzing and improving 3D camera control in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: Taming large video diffusion transformers for 3D camera control. In *International Conference on Learning Representations (ICLR)*, 2025b.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhixue Fang, Xu He, Songlin Tang, Haoxian Zhang, Qingfeng Li, Xiaoqiang Liu, Pengfei Wan, and Kun Gai. 3D-aware implicit motion control for view-adaptive human video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2243–2252, 2026.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, and Yahui Zhou. SkyReels-A2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426, 2025b.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. ID-Animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Li Hu. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8153–8163, 2024.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. HunyuanCustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025.

Kling Team, Jialu Chen, Yikang Ding, Zhixue Fang, Kun Gai, Kang He, Xu He, Jingyun Hua, Mingming Lao, Xiaohan Li, Hui Liu, Jiwen Liu, Xiaoqiang Liu, Fan Shi, Xiaoyu Shi, Peiqin Sun, Songlin Tang, Pengfei Wan, Tiancheng Wen, Zhiyong Wu, Haoxian Zhang, Runze Zhao, Yuanxing Zhang, and Yan Zhou. Kling-MotionControl technical report. *arXiv preprint arXiv:2603.03160*, 2026.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024.

Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. BindWeave: Subject-consistent video generation via cross-modal integration. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10486–10496, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, 2025.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. CamCloneMaster: Enabling reference-based camera control for video generation. In *SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. MIMO: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21181–21191, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 2025.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Jiangchuan Wei, Shiyue Yan, Wenfeng Lin, Boyuan Liu, Renjie Chen, and Mingyu Guo. EchoVideo: Identity-preserving human video generation by multimodal feature fusion. *arXiv preprint arXiv:2501.13452*, 2025.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 657–666, 2022.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-In: A lightweight and plug-and-play identity control for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23314–23324, 2026.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4210–4220, 2023.

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*, 2023.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, 2025.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Chongyang Ma, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 38, 2025a.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988, 2025b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, pp. 74896–74910, 2025a.

Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. MagicMirror: ID-preserved video generation in video diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14464–14474, 2025b.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3661–3670, 2021.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Yong Zhong, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Chongxuan Li. Concat-ID: Towards universal identity-preserving video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1927–1936, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 650–667, 2022.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 145–162, 2024.