OMNIHUMANOID: STREAMING CROSS-EMBODIMENT VIDEO GENERATION WITH PAIRED-FREE ADAPTATION

Anonymous authors
Paper under double-blind review

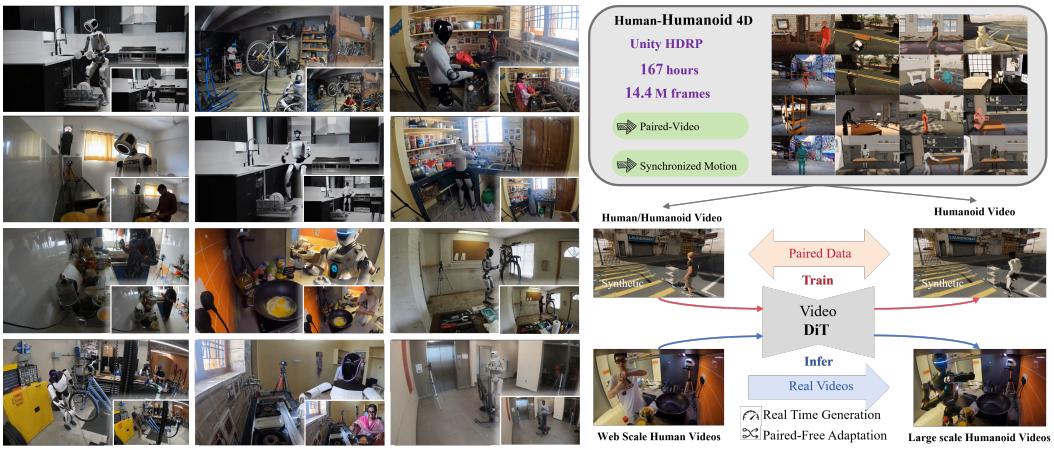

Figure 1: OmniHumanoid enables scalable cross-embodiment video generation by decoupling transferable motion from embodiment-specific appearance. Built on a motion-aligned synthetic dataset spanning diverse humanoids, scenes, and actions, our framework transfers motion to new embodiments and synthesizes consistent target videos with only lightweight adaptation.

ABSTRACT

Cross-embodiment video generation aims to transfer motions across different humanoid embodiments, such as human-to-robot and robot-to-robot, enabling scalable data generation for embodied intelligence. A major challenge in this setting is that motion dynamics are partly transferable across embodiments, whereas appearance and morphology remain embodiment-specific. Existing approaches often entangle these factors, and many require paired data for every target embodiment, which limits scalability to new robots. We present OmniHumanoid, a framework that factorizes transferable motion learning and embodiment-specific adaptation. Our method learns a shared motion transfer model from motion-aligned paired videos spanning multiple embodiments, while adapting to a new embodiment using only unpaired videos through lightweight embodiment-specific adapters. To reduce interference between motion transfer and embodiment adaptation, we further introduce a branch-isolated attention design that separates motion conditioning from embodiment-specific modulation. In addition, we construct a synthetic cross-embodiment dataset with motion-aligned paired videos rendered across diverse humanoid assets, scenes, and viewpoints. Experiments on both synthetic and real-world benchmarks show that OmniHumanoid achieves strong motion fidelity and embodiment consistency, while enabling scalable adaptation to unseen humanoid embodiments without retraining the shared motion model.

1 INTRODUCTION

Learning from videos has become a promising paradigm for embodied intelligence, as videos naturally capture rich motion, interaction, and scene dynamics. However, collecting high-quality demonstrations for every robot embodiment is expensive and difficult to scale. A more scalable alternative is to transform existing demonstrations, such as human videos or videos of other robots, into videos of a target robot embodiment, and use the generated data for downstream robot learning.

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.

This motivates cross-embodiment video generation: given a source video and a target humanoid embodiment, synthesize a target video that preserves the source motion while rendering the desired robot appearance and morphology.

Despite its potential, cross-embodiment video generation remains far from solved. Existing methods face four major obstacles: motion is often entangled with embodiment-specific geometry and kinematics; paired source-target videos are expensive to collect for every new robot; generic video editing models struggle to preserve robot identity, structural details, and temporal consistency for high-DOF robotic bodies; and high-quality video-to-video generation is typically too slow for interactive use or large-scale data production. These limitations make it difficult to build a cross-embodiment generator that is both scalable and practically deployable.

To address these issues, we propose **OmniHumanoid** under the **TAPE** principle: **T**ransferable motion, paired-free **A**daptation, embodiment **P**reservation, and generation **E**fficiency. OmniHumanoid decomposes cross-embodiment generation into two components: a shared motion component capturing how actions and scenes evolve, and an embodiment-specific appearance component determining the target humanoid's look and movement. This design allows motion knowledge to be reused across humanoids, while new robots are introduced through unpaired adaptation.

To realize this factorization, OmniHumanoid introduces a unidirectional information-flow design separating motion understanding from embodiment rendering. The model learns transferable motion from paired multi-embodiment videos, while robot-specific appearance and morphology are captured by lightweight embodiment adapters trained from unpaired videos. By preventing embodiment-specific priors from entering the shared motion pathway, OmniHumanoid preserves general motion transfer ability while allowing efficient adaptation to unseen humanoids.

To support this learning paradigm, we construct a motion-aligned synthetic cross-embodiment dataset. Diverse motions are retargeted to multiple digital humans and robots under aligned scenes, cameras, and temporal structures. The resulting videos share underlying motion and scene dynamics while differing in appearance and morphology, providing controlled supervision for motion transfer and systematic evaluation on unseen robots, actions, and environments.

Beyond paired-free adaptation, OmniHumanoid targets efficient deployment. We introduce a streaming video-to-video distillation pipeline turning the expensive bidirectional generator into an efficient causal generator. This reduces generation from 50 denoising steps to 4 and enables autoregressive cross-embodiment video generation at 720p and 5 FPS on a single NVIDIA H200 GPU.

Extensive experiments on synthetic and real-world benchmarks show that OmniHumanoid achieves strong motion fidelity and embodiment consistency, supports paired-free adaptation to unseen humanoids, and enables efficient long-horizon generation. Compared with generic video editing and offline cross-embodiment pipelines, OmniHumanoid offers a more scalable and practical path toward robot-centric video data generation.

Our contributions are threefold:

- We propose **OmniHumanoid**, a factorized cross-embodiment video generation framework built upon the **TAPE** principle. Through a unidirectional information-flow design, OmniHumanoid decouples transferable motion from embodiment-specific appearance, enabling paired-free adaptation while preserving target embodiment identity.
- We introduce a streaming video-to-video distillation pipeline that converts the bidirectional generator into a causal attention-based model, enabling efficient long-horizon and real-time cross-embodiment video generation.
- We construct a motion-aligned synthetic cross-embodiment dataset across diverse humanoids, motions, scenes, and viewpoints, and validate generalization to unseen embodiments, actions, and environments on both synthetic and real-world benchmarks.

2 Related Work

2.1 Video Generation Models

In recent years, diffusion models have been widely applied to image synthesis (Rombach et al., 2022), image editing (Brooks et al., 2023; Hertz et al., 2022; Huang et al., 2025), and video generation

(Blattmann et al., 2023; Song et al., 2025a; Chen et al., 2025; Guo et al., 2023; Ma et al., 2024b;a; 2025a;b; Li et al., 2025b). Within video generation, the currently prevalent DiT architecture (Ho et al., 2020) has progressively surpassed earlier GAN-based (Pan et al., 2017) and UNet-based methods (Guo et al., 2023; Song et al., 2024) by significantly enhancing visual fidelity and temporal consistency. Today, rapidly evolving DiT frameworks form the foundation of cutting-edge models like WAN (Wan et al., 2025), Sora (Liu et al., 2024), and HunyuanVideo (Kong et al., 2024), pushing generation quality to new heights. Recent studies have further leveraged large pre-trained video generation models for robotic and manipulation tasks (Fu et al., 2025; Ci et al., 2025; Yuan et al., 2026; Kim et al., 2026; Li et al., 2025a), highlighting their potential for cross-domain generalization and interactive learning.

2.2 Robotizing Human Videos

In embodied intelligence, the scarcity of large-scale robot image and video datasets remains a major bottleneck, motivating recent efforts to bridge the embodiment gap by converting human-centric videos into robot-centric observations. In first-person settings, methods jointly train on egocentric human and robot data (Yang et al., 2025b; Kareer et al., 2025; Qiu et al., 2025), where EgoVLA (Yang et al., 2025b) trains vision-language models on both domains, and EgoMimic (Kareer et al., 2025) and PH2D (Qiu et al., 2025) align wearable-camera human demonstrations with teleoperated robot trajectories through shared backbones and cross-domain losses. Data-editing pipelines further scale supervision by converting in-the-wild egocentric human videos into robot-centric observations (Lepert et al., 2025a). Extending robotization to third-person videos is more challenging due to full-body motion, dynamic backgrounds, complex interactions, and self-occlusion. Phantom (Lepert et al., 2025b) overlays rendered robotic arms guided by estimated hand poses, but rendering-based pipelines are limited by lighting, depth, scene geometry, and generalization to unseen robot embodiments. Recent diffusion-based human-to-robot video translation methods address third-person robotization with full-body motion and occlusion handling beyond simple overlays (Yang et al., 2025a; Song et al., 2025b). Building on this direction, we propose OmniHumanoid, a modularly decoupled framework that learns general motion representations with a shared motion model and uses enhanced embodiment LoRA modules for robot appearance replacement under structural guidance, enabling robust human-to-robot video translation in complex motions and cluttered scenes.

2.3 Decoupled Learning

Decoupled learning has become an effective strategy for improving the generalization and controllability of generative visual models by separating factors of variation with different statistical structures. In video generation, AnimateDiff (Guo et al., 2023) decouples spatial and temporal modeling with lightweight Temporal Modules, enabling image generation models to propagate information across time. MotionDirector (Zhao et al., 2024) uses dual-path LoRA to disentangle motion and appearance, separating "how to move" from "what to look like" for better transfer across appearances. In image stylization, many methods decouple style and content to balance stylistic consistency and content preservation (Wang et al., 2024; Xing et al., 2024). OmniConsistency (Song et al., 2025c) further reduces the conflict between style learning and consistency preservation through a two-stage progressive LoRA design. Overall, these methods reduce interference between different factors, leading to improved generation quality and stronger generalization.

3 Methods

We present **OmniHumanoid**, a scalable and streaming framework for cross-embodiment video generation following TAPE. Sec. 3.1 defines paired training and paired-free adaptation; Sec. 3.2 introduces our branch-decoupled architecture; Sec. 3.3 describes two-stage training and unseen-embodiment adaptation; and Sec. 3.4 presents streaming video-to-video distillation for efficient long-horizon inference.

3.1 Problem Formulation

Given a source video V^{src} and a target embodiment e, cross-embodiment video generation synthesizes a target video $\hat{V}^e$ that preserves the motion and scene dynamics of V^{src} while rendering the appearance of embodiment e. During training, we assume access to a motion-aligned paired dataset:

$$\mathcal{D}_{\mathrm{pair}} = \{(V_i^{\mathrm{src}}, V_i^{\mathrm{tgt}}, e_i)\}_{i=1}^{N}, \tag{1}$$

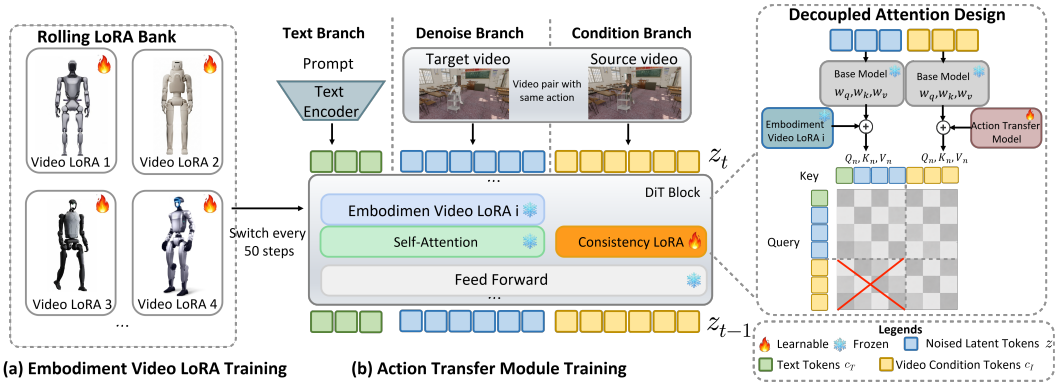

Figure 2: Overview of OmniHumanoid. A Shared Motion Transfer Model learns transferable motion from motion-aligned paired videos, while each Embodiment Video LoRA captures the appearance and morphology of a specific humanoid embodiment from unpaired videos. During paired training, the target embodiment LoRA is activated on the denoising branch, while the shared motion transfer parameters are optimized on the conditioning branch. This branch-decoupled design enables paired-free adaptation to unseen embodiments while preserving transferable motion learning.

where V_i^{src} is a source video, V_i^{tgt} is the corresponding target video under embodiment e_i, and the paired videos share the same underlying motion and scene dynamics. These paired samples provide supervision for learning motion transfer across embodiments. At adaptation time, for a new embodiment e^*, paired source–target videos are unavailable. Instead, we only assume access to unpaired videos:

$$\mathcal{U}_{e^*} = \{V_j^{e^*}\}_{j=1}^{M}. \tag{2}$$

The goal is to adapt the model to e^* using $\mathcal{U}_{e^*}$ only, without retraining the shared motion transfer parameters. This setting motivates a factorized design that separates transferable motion learning from embodiment-specific appearance and morphology modeling.

3.2 Architecture and Unidirectional Motion-Appearance Decoupling

OmniHumanoid is built on a Diffusion Transformer (DiT) video generation backbone. The central design is to decompose cross-embodiment generation into two complementary components: a **Shared Motion Transfer Model** that captures embodiment-invariant motion dynamics, and a bank of **Embodiment Video LoRAs** that model embodiment-specific appearance and morphology.

Shared Motion Transfer Model. The shared motion transfer model operates on the conditioning branch. It takes source video V^{src} to encode transferable motion cues like temporal dynamics, object interactions, and scene evolution. Being shared across all embodiments, this module learns common humanoid behaviors: how actions unfold over time.

Embodiment Video LoRA. For each embodiment e, we introduce a lightweight LoRA module Φ_e attached to the denoising branch. Φ_e captures embodiment-specific geometry, texture, and structural priors. At inference time, the target video is generated by loading the corresponding embodiment LoRA while keeping the shared motion transfer model fixed.

Branch-specific parameterization. We denote the token streams in the DiT backbone by

$$X^{\text{text}}, \qquad X^{\text{den}}, \qquad X^{\text{cond}},$$

corresponding to text tokens, denoising latent tokens, and source-video conditioning tokens. For branch $b \in \{\text{text}, \text{den}, \text{cond}\}$, the standard attention projections are:

$$Q^b = X^b W_Q, \qquad K^b = X^b W_K, \qquad V^b = X^b W_V. \tag{3}$$

For a target embodiment e, we apply LoRA updates only to the denoising branch:

$$W_\bullet^{\text{den},e} = W_\bullet + \Delta W_\bullet^{(e)}, \qquad \Delta W_\bullet^{(e)} = B_\bullet^{(e)} A_\bullet^{(e)}, \qquad \bullet \in \{Q, K, V\}, \tag{4}$$

where $A_\bullet^{(e)} \in \mathbb{R}^{r\times d}$, $B_\bullet^{(e)} \in \mathbb{R}^{d\times r}$, and $r \ll d$. The text and conditioning branches use the shared backbone projections without embodiment-specific updates.

Unidirectional information flow. The key design is not only where the embodiment LoRA is attached, but also how information is allowed to flow. To prevent embodiment-specific appearance priors from directly affecting the shared motion pathway, we apply an asymmetric attention mask:

$$\mathcal{M}(\text{den} \to \text{cond}) = 1, \qquad \mathcal{M}(\text{cond} \to \text{den}) = 0, \tag{5}$$

where 1 and 0 denote visible and masked attention, respectively. Thus, the denoising branch can read motion information from the conditioning branch, while the conditioning branch is isolated from embodiment-specific denoising updates. This unidirectional information flow helps preserve transferable motion representations while allowing the denoising branch to render target-specific morphology and appearance.

3.3 Two-Stage Training and Paired-Free Adaptation

The factorized architecture naturally leads to a two-stage training strategy. The first stage teaches each embodiment LoRA what a target embodiment looks like, while the second stage teaches the shared model how motion transfers across embodiments.

Stage I: Embodiment Video LoRA Pretraining. For each seen embodiment e, we train Φ_e using only unpaired videos $\mathcal{U}_e$. The DiT backbone is frozen, and only the LoRA parameters are optimized under the standard video diffusion denoising objective. Since Φ_e is attached to the denoising branch, it learns embodiment-specific appearance, geometry, and morphology directly in the generative pathway without requiring paired motion supervision.

Stage II: Shared Motion Transfer Training. We then train the Shared Motion Transfer Model on the paired dataset $\mathcal{D}_{\text{pair}}$ while freezing all Embodiment Video LoRAs. For each paired sample $(V^{\text{src}}, V^{\text{tgt}}, e)$, we activate the LoRA Φ_e corresponding to the target embodiment and optimize only the shared motion transfer parameters. To reduce bias toward any single embodiment, we adopt a *rolling LoRA loading* strategy: the active LoRA changes across training iterations according to the target embodiment of the current batch. This encourages the shared model to focus on cross-embodiment invariances, namely motion and scene dynamics, while leaving embodiment-specific rendering to the LoRAs.

Adaptation to Unseen Embodiments. For a new humanoid embodiment e^*, we instantiate a new LoRA Φ_{e^*} and train it using only a small collection of unpaired videos $\mathcal{U}_{e^*}$, typically on the order of tens of videos, without requiring any motion-aligned source–target pairs. The Shared Motion Transfer Model is kept frozen throughout this adaptation stage. During inference, OmniHumanoid takes a source video V^{src} and the target embodiment e^*, loads the adapted LoRA Φ_{e^*}, and synthesizes a target video that transfers the source motion to the new robot embodiment.

3.4 Streaming Video-to-Video Distillation

While the decoupled bidirectional generator provides high-quality cross-embodiment synthesis, standard DiT-based generation requires iterative full-sequence denoising, making it computationally expensive for interactive use and large-scale data production. To improve efficiency, we distill the bidirectional OmniHumanoid generator into a causal streaming student.

We organize the reference, source-video condition, and target-video tokens into an interleaved super-chunk layout:

$$[\text{ref} \mid \text{cond}_0 \mid \text{tgt}_0 \mid \cdots \mid \text{cond}_M \mid \text{tgt}_M]. \tag{6}$$

During inference, target chunk i attends only to the reference, previous chunks, and the current condition chunk, while future chunks are masked out. This block-wise causal structure enables autoregressive rollout with cached key-value states.

We train the causal student with a two-stage distillation objective. First, the student is initialized from the bidirectional teacher and optimized with teacher-forcing denoising score matching under the causal mask, which preserves basic token prediction fidelity. We then apply self-forcing few-step distillation, where the student is optimized under its own autoregressive rollout trajectory with

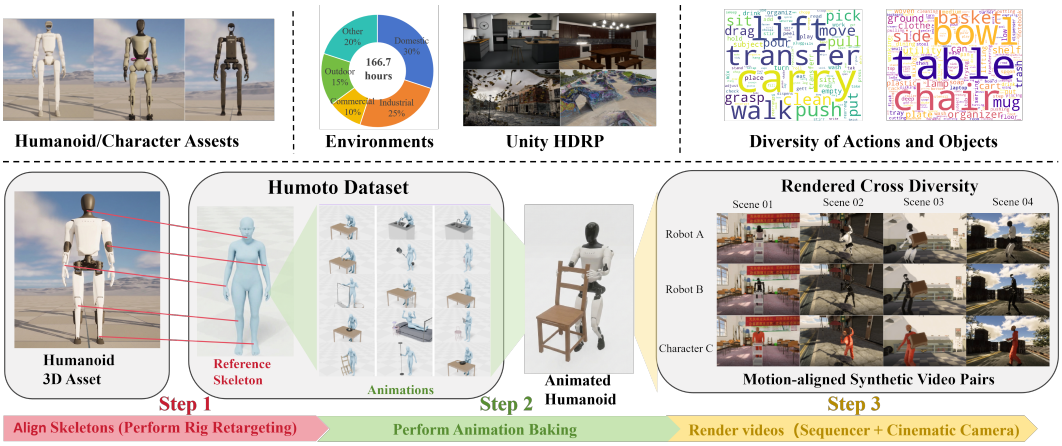

Figure 3: Synthetic data construction pipeline. We create paired human-humanoid videos from community character, animation, and environment assets by aligning skeletons, retargeting the same motions to different embodiments, and recording them in diverse scenes with identical camera setups.

teacher-guided distribution matching:

$$\mathcal{L}_{\text{stream}} = \mathcal{L}_{\text{DSM}} + \lambda_{\text{vsd}}\mathcal{L}_{\text{VSD}} + \lambda_{\text{gan}}\mathcal{L}_{\text{GAN}}. \tag{7}$$

Here, $\mathcal{L}_{\text{DSM}}$ denotes the denoising score matching loss under causal attention, $\mathcal{L}_{\text{VSD}}$ aligns the fast causal student with the frozen bidirectional teacher's score guidance, and $\mathcal{L}_{\text{GAN}}$ improves local sharpness and reduces detail degradation in few-step generation. This distillation converts the 50-step bidirectional generator into a 4-step causal streaming model, enabling efficient long-horizon cross-embodiment video generation.

4 Synthetic Cross-Embodiment Dataset

To support the factorized learning paradigm and evaluate cross-embodiment generalization, we construct a motion-aligned synthetic video dataset with a Unity-based rendering pipeline. The dataset is built on the Humoto motion library, which provides over 700 humanoid motion sequences covering object manipulation, environment interaction, locomotion, and daily full-body activities.

Motion-aligned rendering. We curate ten humanoid assets, including five humanoid robots and five digital human avatars. To ensure reliable cross-embodiment motion alignment, all assets are equipped with compatible skeletons and animation controllers. We align their skeletons to a unified topology in Blender and perform animation retargeting in Unity, establishing joint-level correspondence across embodiments under a shared motion source.

Scene and embodiment variation. Each selected motion sequence is retargeted to all training embodiments. We render videos in 100 diverse scenes from online 3D platforms, covering offices, factories, and outdoor spaces. Within each scene, we vary only the humanoid asset while keeping the scene layout, camera viewpoint, and motion sequence unchanged. Thus, paired videos share identical motion and context, differing primarily in embodiment appearance and morphology. All videos are rendered at 1920×1080 resolution and 30 FPS, each containing roughly 300 frames.

Paired sample construction and evaluation split. To construct paired supervision, we form ordered source–target pairs among the training embodiments under matched motion and scene configurations, resulting in 7,200 paired training samples across 100 scenes. To evaluate generalization to unseen embodiments, one robot embodiment is fully held out from all training stages and used exclusively for testing. The test set also includes unseen motion tasks and scene configurations, ensuring that evaluation reflects cross-embodiment, cross-motion, and cross-environment generalization rather than memorization of seen embodiment–motion combinations.

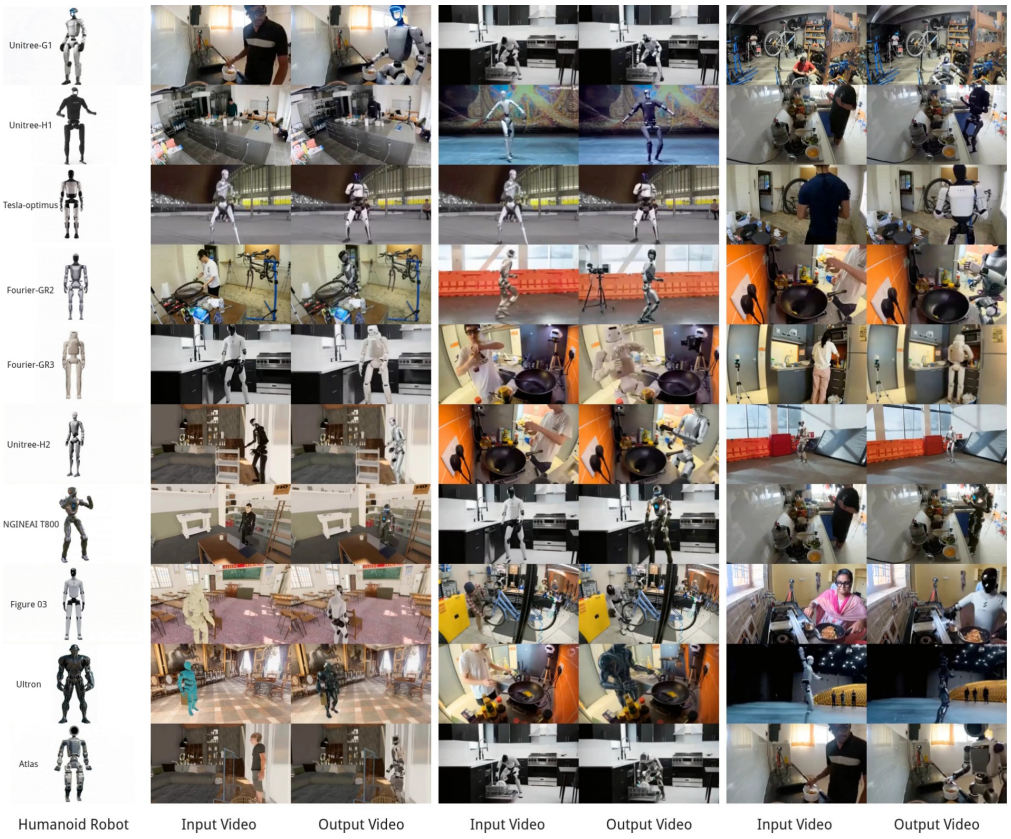

Figure 4: Generalization to new scenes, tasks, and both seen and unseen robot embodiments. The top five rows show embodiments seen during training, while the bottom five show unseen ones. Given a reference video and a target humanoid-embodiment LoRA, our model transfers the motion to the target robot under novel environments and action configurations absent from paired supervision. The generated videos preserve action consistency, embodiment identity, and scene-consistent relighting. Readers can click and play the clips in this figure using **Adobe Acrobat**.

5 Experiments

5.1 Experimental Setup

Implementation Details. We finetune our model based on the Wan2.2-TI2V-5B backbone using rank-64 LoRA adaptation. Training is performed for 10000 optimization steps on 2 NVIDIA H200 GPUs, with a batch size of 1 per GPU. We adopt the AdamW optimizer with a learning rate of $2e-4$. More implementation details are provided in the supplementary material.

Baselines. We compare with state-of-the-art V2V generation and editing systems, including commercial APIs such as Runway Gen-4 (Runway AI, 2025), Kling O1 (Team et al., 2025), and Kling O3 (Kuaishou Technology, 2026), as well as open-source methods such as Wan2.1-VACE (Wan et al., 2025) and the closely related X-Humanoid (Yang et al., 2025a). Commercial APIs are evaluated as zero-shot reference-guided models without embodiment-specific adaptation, while trainable open-source baselines, including Wan2.1-VACE (Wan et al., 2025) and X-Humanoid (Yang et al., 2025a), are fine-tuned on the same synthetic dataset following their original training strategies.

Benchmarks. We evaluate OmniHumanoid on two complementary benchmarks. The first is the Synthetic Held-out Embodiment Benchmark, where we reserve one unseen robot model, Unitree-G1, from training and use 50 motion-aligned videos for testing. Since ground-truth target videos are available in this synthetic setting, it supports reference-based quantitative evaluation. The second is the Real-world Benchmark, which contains 50 in-the-wild videos collected from human daily activities and online robot demonstrations. These videos cover diverse indoor tasks and scenes such

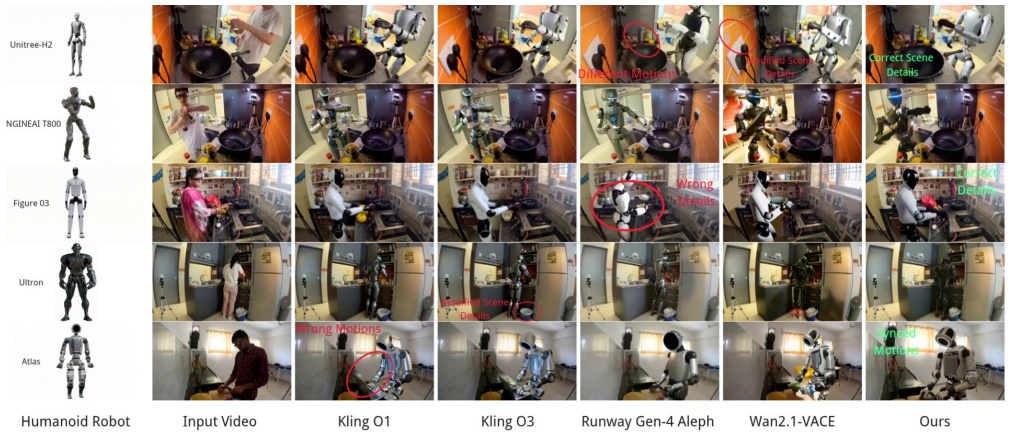

Figure 5: Qualitative comparison of motion transfer on various robot embodiments. Compared to general video generation baselines, our method generates significantly higher fidelity motions with fewer visual artifacts. Our framework enables transferable motion synthesis across arbitrary humanoid types without retraining the core model, achieving embodiment-consistent generation at 720P resolution. Readers can click and play the video clips in this figure using **Adobe Acrobat**.

as kitchens, garages, laboratories, and theaters, enabling reference-free evaluation under realistic visual conditions.

Metrics. We evaluate OmniHumanoid on two benchmarks: the Synthetic Held-out Embodiment Benchmark with ground-truth target videos, and the Real-world Benchmark without paired ground truth. For the synthetic benchmark, we report PSNR, SSIM, and MSE to measure reconstruction quality and motion-aligned fidelity between generated and ground-truth videos. For both benchmarks, we use Gemini-3 Flash as a vision-language evaluator to assess reference-free consistency: background consistency between the input and generated videos, motion consistency between the source and generated robot motions, and embodiment identity consistency assessing coherent robot-specific appearance. Thus, the synthetic benchmark supports reconstruction and reference-free consistency metrics, while the real-world benchmark uses only reference-free metrics.

5.2 Comparison and Evaluation

Table 1 reports quantitative results under the unseen robot adaptation setting. Our method consistently outperforms V2V baselines across all metrics, especially in motion fidelity and embodiment correctness. Figure. 5 further show that generic V2V models often suffer from motion drift, inconsistent limb articulation, and weak embodiment fidelity, whereas our adapted model better preserves motion trajectories while accurately reflecting the target robot embodiment.

Table 1: Comparison of unseen robot adaptation on the Synthetic Held-out Embodiment Benchmark and Real-world Benchmark. Higher is better except MSE.

	Synthetic Held-out Benchmark							Real-world Benchmark			
Method	PSNR ↑	SSIM ↑	MSE ↓	Motion. ↑	Embod. ↑	BG. ↑	Overall ↑	Motion. ↑	Embod. ↑	BG. ↑	Overall ↑
Kling O1	22.70	0.8951	0.0067	8.06	6.94	9.52	7.08	7.49	8.46	9.91	**8.53**
Kling O3	22.76	0.8959	0.0065	8.76	7.90	9.32	7.42	7.47	8.34	9.82	8.21
Runway Gen4	18.83	0.6575	0.0181	7.26	7.50	8.14	7.22	6.79	5.07	8.61	7.22
Wan2.1-VACE	22.44	0.8599	0.0066	6.40	5.88	8.68	6.22	5.60	5.65	7.85	6.45
X-Humanoid	23.03	0.8891	0.0057	8.94	8.04	9.78	7.53	–	–	–	–
Ours	**25.47**	**0.9039**	**0.0033**	**9.06**	**8.43**	**9.94**	**7.92**	**8.47**	**8.56**	**9.95**	8.39

5.3 Ablation Studies

We conduct system-level ablation studies on the core design of OmniHumanoid. Since our architecture combines several coupled components, including branch-isolated attention, rolling embodiment-LoRA loading, two-stage training, and streaming distillation, some low-level components cannot be

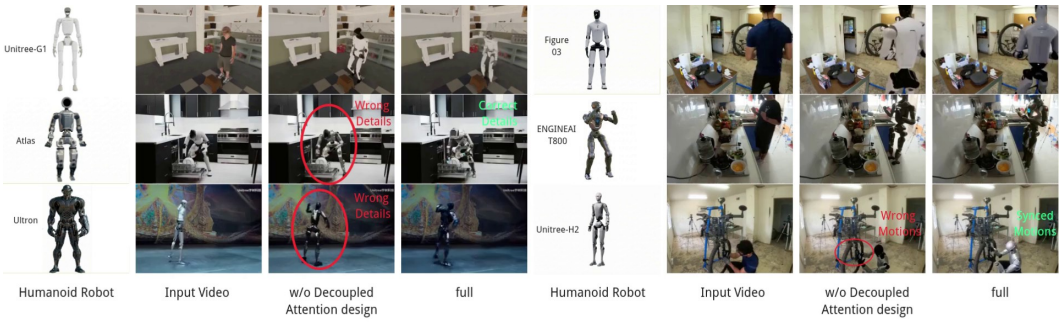

Figure 6: **Effectiveness of Decoupled Attention.** We compare our full model with a baseline without decoupled attention. As highlighted by red circles, the baseline suffers from rendering errors ("Wrong Details") and physical inconsistencies ("Wrong Motion"), while our full model produces faithful embodiments and accurate motion transfer for both seen and unseen robot identities. Readers can click and play the clips using **Adobe Acrobat**.

Table 2: Ablation study on OmniHumanoid variants across decoupling and distillation stages. Best results are in bold.

Method	FPS ↑	PSNR ↑	SSIM ↑	MSE ↓	Motion. ↑	Embod. ↑	BG. ↑	Overall ↑
Teacher Model w/o Decoupling	0.21	21.21	0.8928	0.0098	6.35	2.53	8.56	6.80
Teacher Model	0.10	**25.47**	**0.9039**	**0.0033**	**9.06**	**8.43**	**9.94**	**7.92**
Causal Student	4.96	23.35	0.8841	0.0058	8.82	8.07	9.58	7.60
Full Streaming Student	**4.96**	23.34	0.8900	0.0053	8.90	8.09	9.64	7.73

removed in isolation without breaking the training or adaptation pipeline. We therefore focus on two main factors: motion-appearance decoupling and streaming distillation. As shown in Table 2, we compare four variants: a teacher model without motion-appearance decoupling, the full bidirectional teacher, a causal student trained without self-forcing, and the full streaming student.

The results show that motion-appearance decoupling is the most critical component. Without decoupling, the embodiment score drops from 8.43 to 2.53, while the motion score decreases from 9.06 to 6.35. This indicates that directly mixing source-motion conditioning with target-embodiment rendering causes strong interference, making it difficult to preserve both transferable motion and robot appearance. In contrast, the bidirectional teacher with our decoupled design achieves the best reconstruction quality and the strongest motion, embodiment, and background consistency, validating the importance of explicitly separating transferable motion from embodiment-specific appearance.

Streaming distillation mainly improves efficiency at the cost of some generation quality. Compared with the bidirectional teacher, the causal student increases the speed from 0.10 FPS to 4.85 FPS, but its reconstruction and consistency metrics decrease. The full streaming student further improves efficiency to 4.96 FPS and slightly recovers quality compared with the causal student, improving SSIM, MSE, motion, embodiment, background, and overall scores. These results show that streaming distillation provides a practical speed-quality trade-off, while motion-appearance decoupling remains the key factor for stable cross-embodiment generation.

6 Conclusion

We present a scalable framework for human-to-robot video generation that explicitly decouples transferable motion learning from robot-specific appearance modeling. By combining paired pretraining on a limited set of robot embodiments with lightweight, paired-free adaptation via RobotAppearance LoRA modules, our approach enables efficient generalization to unseen robots without retraining the core model. A decoupled attention architecture and rolling LoRA training strategy ensure minimal interference between motion transfer and appearance adaptation. Together with a large-scale synthetic dataset constructed through motion-aligned combinatorial rendering, our method reduces data collection cost while delivering strong motion fidelity and embodiment consistency, offering a practical path toward scalable robot data generation for embodied intelligence.

AI USE STATEMENT

As described in the evaluation methodology, Gemini-3 Flash was used as a vision-language evaluator for generated videos. Claude Code assisted with portions of the experimental workflow and automated evaluation. OpenAI Codex was used to convert the LaTeX source to the ICLR 2027 template and adjust document formatting. The authors reviewed the resulting outputs and key code components and take full responsibility for the code, experimental results, and final manuscript.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 4 describes the synthetic dataset construction, Section 5.1 specifies the training configuration and evaluation setup, and Appendix A provides the complete prompt for automated evaluation. The complete implementation has been released as open-source software. To preserve double-blind anonymity, we omit the public repository link from this submission.

REFERENCES

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.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Xuewei Chen, Zhimin Chen, and Yiren Song. Transanimate: Taming layer diffusion to generate rgba video. *arXiv preprint arXiv:2503.17934*, 2025.

Hai Ci, Xiaokang Liu, Pei Yang, Yiren Song, and Mike Zheng Shou. H2r-grounder: A paired-data-free paradigm for translating human interaction videos into physically grounded robot videos. *arXiv preprint arXiv:2512.09406*, 2025.

Xiao Fu, Xintao Wang, Xian Liu, Jianhong Bai, Runsen Xu, Pengfei Wan, Di Zhang, and Dahua Lin. Learning video generation for robotic manipulation with collaborative trajectory control. *arXiv preprint arXiv:2506.01943*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, Mike Zheng Shou, and Jiaming Liu. Photodoodle: Learning artistic image editing from few-shot pairwise data. *arXiv preprint arXiv:2502.14397*, 2025.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Kuaishou Technology. All-new klingai 3.0 series. https://klingai.com/, 2026.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025a.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*, 2025b.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025a.

Yuanhang Li, Yiren Song, Junzhe Bai, Xinran Liang, Hu Yang, Libiao Jin, and Qi Mao. Ic-effect: Precise and efficient video effects editing via in-context learning. *arXiv preprint arXiv:2512.15635*, 2025b.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4117–4125, 2024a.

Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024b.

Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6018–6026, 2025a.

Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025b.

Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1789–1798, 2017.

Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Runway AI. Runway gen-4: Ai video generation with world consistency. https://runwayml.com/research/introducing-runway-gen-4, 2025.

Quanjian Song, Yiren Song, Kelly Peng, Yuan Gao, and Mike Zheng Shou. Worldwander: Bridging egocentric and exocentric worlds in video generation. *arXiv preprint arXiv:2511.22098*, 2025a.

Yiren Song, Shijie Huang, Chen Yao, Xiaojun Ye, Hai Ci, Jiaming Liu, Yuxuan Zhang, and Mike Zheng Shou. Processpainter: Learn painting process from sequence data. *arXiv preprint arXiv:2406.06062*, 2024.

Yiren Song, Cheng Liu, Weijia Mao, and Mike Zheng Shou. Mitty: Diffusion-based human-to-robot video generation. *arXiv preprint arXiv:2512.17253*, 2025b.

Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025c.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024.

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. X-humanoid: Robotize human videos to generate humanoid videos at scale. *arXiv preprint arXiv:2512.04537*, 2025a.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024.

APPENDIX

A PROMPT FOR AUTOMATED EVALUATION

We employ a comprehensive prompt structure to evaluate the generated videos across four distinct dimensions. The complete prompt template is shown below:

Comprehensive Evaluation Prompt

```
# ========================
# PROMPT: MOTION FIDELITY
# ========================
You are an expert evaluator for video generation and motion
    tracking.

You will see TWO videos:
(1) Condition video (Human): the original reference clip showing
    the driving motion.
(2) Generated video (Robot): the synthesized output.

TASK: Evaluate MOTION FIDELITY of the generated robot compared to
    the human condition video.
Focus on:
- Movement accuracy: Does the generated robot's motion (pose,
    trajectory, speed, and timing) precisely match the human's
    motion?
- Articulation: Are the joint movements, limb articulations, and
    subtle actions faithful to the human reference?
- Smoothness: Is the motion temporally smooth without unnatural
    jitter or sudden jumps?

SCORING (1-10):
10: Motion perfectly matches the condition video; fluid and
    flawless.
8-9: Very accurate motion; only extremely minor differences in
    timing or subtle jitter.
6-7: Noticeable differences in motion or occasional jitter, but the
     general action matches.
4-5: Frequent motion errors, stiffness, or desynchronization.
1-3: Motion completely breaks down or fails to follow the condition
     video.

OUTPUT FORMAT:
Return ONLY a valid JSON object exactly like:
{"score": <number 1-10>, "reason": "<one short sentence with the
    main evidence>"}


# ========================
# PROMPT: EMBODIMENT SIMILARITY
# ========================
You are an expert evaluator for identity and embodiment
    preservation in video synthesis.

You will see TWO files:
(1) Reference Image (Robot): The ground-truth appearance of the
    target robot.
(2) Generated Video (Robot): The synthesized output video.

TASK: Evaluate EMBODIMENT SIMILARITY of the generated robot
    compared to the reference image.
Focus on:
```

```
- Identity preservation: Does the robot in the video look EXACTLY
    like the robot in the reference image?
- Structural integrity: Are body proportions, mechanical features,
    colors, and structural details faithfully preserved?
- Temporal consistency: Does the robot morph, warp, or change shape/
    identity across frames in the video?

SCORING (1-10):
10: Perfect preservation of identity, proportions, and appearance;
    zero morphing.
8-9: Very high similarity; minor softening or tiny structural
    shifts.
6-7: Noticeable morphing or differences in details, but still
    clearly the same robot.
4-5: Severe morphing; proportions feel wrong; identity is barely
    maintained.
1-3: Robot is unrecognizable or structurally completely broken
    compared to the image.

OUTPUT FORMAT:
Return ONLY a valid JSON object exactly like:
{"score": <number 1-10>, "reason": "<one short sentence with the
    main evidence>"}


# =========================
# PROMPT: BACKGROUND CONSISTENCY
# =========================
You are an expert evaluator for video background stability.

You will see TWO videos:
(1) Condition video (Human): the original reference clip.
(2) Generated video (Robot): the synthesized output.

TASK: Evaluate BACKGROUND CONSISTENCY of the generated video.
Focus on:
- Stability: Does the static background remain completely rigid and
     still over time?
- Fidelity: Are the background details, textures, and structures
    preserved exactly as in the condition video?
- Artifacts: Check for background flickering, texture crawling,
    unwanted warping near the moving robot, or "hallucinated"
    elements.

SCORING (1-10):
10: Background is perfectly stable, artifact-free, and identical to
     the condition video.
8-9: Mostly stable; very minor flickering or tiny artifacts in
    complex areas.
6-7: Noticeable background warping or flickering, especially near
    the moving robot.
4-5: Frequent background distortion, crawling textures, or loss of
    detail.
1-3: Background is extremely unstable, totally warped, or
    completely hallucinated.

OUTPUT FORMAT:
Return ONLY a valid JSON object exactly like:
{"score": <number 1-10>, "reason": "<one short sentence with the
    main evidence>"}


# =========================
```

```
# PROMPT: OVERALL QUALITY
# ========================
You are an expert evaluator for cinematic video quality.

You will see ONE video:
(1) Generated video (Robot): the synthesized output.

TASK: Evaluate OVERALL QUALITY of the generated video.
Focus on:
- Visual fidelity: Sharpness, clarity, and lack of noise/blur.
- Realism: Does the lighting, rendering, and compositing look
    professional and aesthetically pleasing?
- Artifact absence: Are there any glaring AI generation artifacts (
    e.g., weird pixelation, smearing, color shifts)?
- Holistic impression: How pleasing and professional is the final
    result?

SCORING (1-10):
10: Flawless cinematic quality; hyper-realistic and sharp.
8-9: High quality; very clean with only barely noticeable AI
    artifacts.
6-7: Acceptable quality; some visible noise, blur, or slight AI
    artifacts, but highly watchable.
4-5: Poor quality; obvious generation artifacts, severe blur, or
    color degradation.
1-3: Unwatchable; completely destroyed by artifacts or extreme blur.


OUTPUT FORMAT:
Return ONLY a valid JSON object exactly like:
{"score": <number 1-10>, "reason": "<one short sentence with the
    main evidence>"}
```

B User Study Results

We conducted a user study to evaluate the generated videos across four key dimensions: Motion Fidelity, Embodiment Similarity, Background (BG) Consistency, and Overall Quality. Participants were asked to vote for the best method in each category. The results, representing the percentage of total votes each method received, are summarized in Table 3. As shown, our method significantly outperforms existing baselines across all metrics.

Table 3: User study results. Numbers indicate the percentage of user preference votes each method received across four evaluation metrics.

Method	Motion Fidelity	Embodiment Similarity	BG Consistency	Overall Quality
Kling O1	7.1%	14.1%	11.1%	8.1%
Kling O3	12.1%	16.2%	17.2%	20.2%
Ours	**72.7%**	**65.7%**	**62.6%**	**63.6%**
Runway Aleph	8.1%	4.0%	9.1%	8.1%

C Limitations

While our proposed method significantly improves generation efficiency, it still possesses certain limitations. Specifically, we observe a noticeable performance degradation in the few-step distillation regime compared to the original teacher model. When operating at extremely low inference steps (e.g.4 steps), the generated videos may exhibit slight compromises in fine-grained details, temporal smoothness, or complex motion fidelity when directly compared to the high-quality outputs generated by the full-step teacher model. This phenomenon highlights an inherent trade-off between inference

speed and generation quality within current distillation frameworks. The distilled student model, although highly efficient in reducing computational overhead, still struggles to perfectly approximate the complex data distribution captured by the teacher model over dozens of denoising steps. In future work, we plan to explore more advanced knowledge distillation strategies to mitigate this issue.