3D-Aware Implicit Motion Control for View-Adaptive Human Video Generation

Supplementary Material

Overview

In this supplementary document, we provide additional details to support the main paper, organized as follows:

- **Section A:** Additional quantitative evaluations on view-adaptive control, 3D consistency, and cross-identity motion transfer.
- **Section B:** Demonstrations of broader applications, including single-image novel view synthesis, video stabilization, and automatic motion alignment.
- **Section C:** The composition of the full training set and the data mixing strategy used during training.
- **Section D:** Detailed specifications of our in-house data collection pipeline and the definition of camera trajectories.
- **Section E:** A discussion on current limitations and potential directions for future work.

A. Additional Quantitative Evaluation

To further complement the main paper, we provide additional quantitative evaluations on view-adaptive motion control, 3D consistency, and cross-identity motion transfer.

A.1. View-Adaptive Control and 3D Consistency

We first evaluate view-adaptive motion control under moving-camera scenarios. We compare our method with a composed baseline, **Wan-Animate (W-A) + ReCamMaster (ReCam)**, using FID, FVD, and VBench metrics without ground-truth references. For fairness, we extract camera extrinsics from our generated results and use them to condition ReCam, ensuring aligned camera trajectories across methods.

In addition, following prior work, we assess 3D consistency across views under fixed poses with moving cameras using CLIP similarity and matching pixel metrics. As shown in Tab. A1, our method significantly outperforms the composed baseline in visual quality, motion fidelity, subject consistency, and cross-view 3D consistency. In contrast, the baseline tends to exhibit more noticeable motion distortion and weaker structural coherence, suggesting that generic camera-control pipelines struggle to preserve human motion and geometry as effectively as our implicit motion representation.

A.2. Cross-Identity Motion Transfer

We further evaluate cross-identity motion transfer, where the driving motion and the reference subject come from different identities. Under this setting, we compare our method against **Wan-Animate** in the static-camera scenario.

As shown in Tab. A1, our method consistently achieves better results across all reported metrics. This more challenging setting further highlights the advantage of our implicit motion representation: instead of rigidly copying 2D pose patterns, it captures genuine motion dynamics that generalize more effectively across identities.

B. Broader Applications

Benefiting from the implicit 3D motion reasoning and flexible text-driven camera control of 3DiMo, our approach generalizes effectively to specific downstream tasks, as shown in Fig. 1.

Human Novel View Synthesis From a Single Image. While traditional novel view synthesis (NVS) typically requires reconstructing 3D scenes from reference images to render new angles, 3DiMo achieves human-specific single-image NVS through a straightforward inference strategy. We construct a driving video by repeating the reference frame (implying zero motion) and pair it with text prompts describing camera trajectories (e.g., "*camera rotates in a circular path around the woman*"). Although pre-trained I2V foundation models theoretically support this via prompts specifying camera movement alongside "*static subject*," they suffer from significant limitations in practice. As noted by [1], these models tend to hallucinate motion, failing to keep the subject strictly stationary. Moreover, we observe that base I2V models often confuse camera control with background animation rather than performing true geometric view synthesis. By leveraging our view-agnostic motion representation and the model's improved 3D awareness, 3DiMo overcomes these ambiguities to produce consistent novel-view generations.

Video Stabilization. Capturing stable footage during dynamic recording is often challenging. Video stabilization aims to smooth out camera jitters to obtain high-quality, steady sequences. In human-centric scenarios, 3DiMo effectively performs this task. By utilizing the first frame of the shaky video as the reference image and the full video as the driving signal, we can feed the model a prompt such as "*camera remains static*." This instructs the generator to reconstruct the underlying human motion from a fixed viewpoint, effectively eliminating the original camera shake while preserving the subject's dynamics.

Automatic Motion-Image Alignment. Conventional motion transfer methods, particularly 2D-based approaches, rigidly impose the absolute orientation of the driving video onto the reference subject. This often leads to unnatural transitions when the driving and reference subjects have dif-

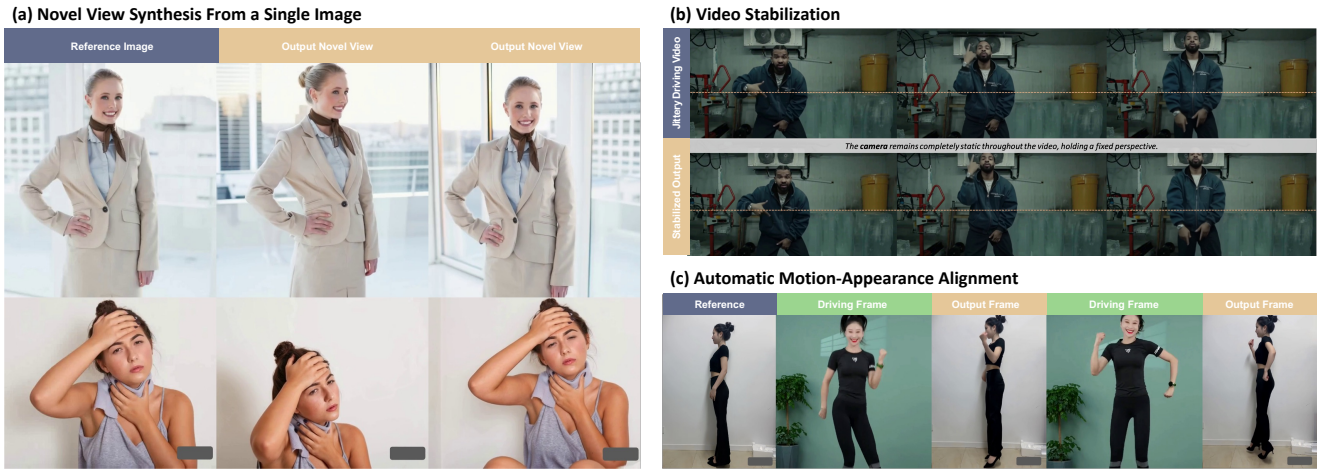

Figure 1. **Broader applications of 3DiMo.** We demonstrate the versatility of our framework on three downstream tasks: (a) single-image novel view synthesis by enforcing static motion; (b) video stabilization by suppressing camera jitter from the driving video; and (c) automatic motion-appearance alignment without explicit calibration.

Table A1. Quantitative comparison on view-adaptive motion control and cross-identity motion transfer. For view-adaptive control, we compare **Ours (moving)** against **W-A + ReCam** under moving-camera settings. For cross-identity transfer, we compare **Ours (static)** against **Wan-Animate** under static-camera settings.

Method	FID ↓	FVD ↓	Aesthetic Quality ↑	Motion Smooth. ↑	Subject Consist. ↑	CLIP Sim. ↑	Mat. Pix. ↑
Ours (moving)	**39.5**	**355.6**	**56.1**	**99.2**	**92.1**	**88.7**	**862**
W-A + ReCam	52.6	558.3	48.2	98.3	86.9	64.5	499
Ours (static)	**37.9**	**302.1**	**56.3**	**99.3**	**93.5**	-	-
Wan-Animate	40.9	385.9	53.2	98.9	87.8	-	-

ferent initial facing directions (e.g., a side-view driver controlling a front-view reference). In contrast, because 3DiMo extracts a view-agnostic implicit motion representation, it naturally aligns the driving motion with the reference subject's initial orientation. Our model transfers the relative 3D dynamics rather than the absolute 2D projection, eliminating the need for manual camera calibration or explicit root-rotation alignment required by SMPL-based methods.

C. Training Data Composition and Mixing Strategy

Our full training set is composed of both large-scale single-view videos and a smaller subset of view-rich videos. The single-view portion provides broad coverage of identities, motions, appearances, and scenes, which is important for maintaining diversity and generalization. The view-rich portion, including multi-view and moving-camera videos, provides the critical supervision needed to learn view-agnostic motion representations and foster 3D-aware generation.

In the final training recipe, we use a mixture of **80% single-view data** and **20% view-rich data**. This ratio provides a practical balance between motion and appearance diversity on the one hand, and geometric supervision on the other.

Empirically, we find that the model is relatively robust to moderate changes in this ratio. A noticeable degradation in view-adaptive camera control is observed only when the proportion of view-rich data is reduced to a very low level (e.g., below 5%). This suggests that while view-rich supervision is essential for activating 3D-aware motion learning, only a moderate amount is needed to effectively unlock the pretrained video generator's inherent 3D spatial priors.

D. In-House Data Acquisition Setup

Our in-house data capture involves a three-camera array positioned at diverse angles relative to the subject. For every captured performance, each camera is assigned a camera motion type sampled randomly from the following categories:

1. **Static Variants:** Static, Handheld Static.
2. **Linear Translations:** Move Forward, Move Back, Move Left, Move Right, Move Up, Move Down.
3. **Zoom Actions:** Zoom In, Zoom Out, Rapid Zoom In,

Rapid Zoom Out, Handheld Zoom In, Aerial Pull-out.

4. **Complex Trajectories:** Vertigo In, Vertigo Out, Dynamic Zoom Swing, Arc Left (variable angles, e.g., $30°, 45°$), and Arc Right (variable angles, e.g., $30°, 45°$).

By pairing identical human motions with diverse, non-correlated camera trajectories across three views, we maximize the supervision signal for view-agnostic motion learning.

E. Limitations and Future Work

Despite the significant advancements 3DiMo achieves in view-adaptive human video generation, several limitations remain to be addressed in future research.

Resolution and Fine-Grained Details. Currently, our framework operates at a resolution of 480p. While this is sufficient for capturing global motion dynamics, it imposes a bottleneck on high-frequency details. Specifically, in full-body shots where the subject occupies a relatively small proportion of the frame, the limited pixel budget can lead to artifacts, such as blurred facial features or a lack of texture in hand details. Future iterations could address this by scaling up the framework to higher-resolution DiT backbones (e.g., 720p or 1080p) or incorporating cascaded super-resolution modules to enhance local details in small-scale regions.

Complex Human-Object Interactions. Since our motion encoders are explicitly designed to distill human body and hand dynamics, the current framework does not explicitly model the motion of external objects or props (e.g., a person holding a bag or riding a bicycle). Consequently, while the human motion is faithfully reproduced, the interaction with held objects may sometimes be hallucinated. Extending the implicit motion encoding mechanism to handle general dynamic objects or human-scene interactions represents a promising direction for future work.

References

[1] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 1