This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Zhixue Fang[1,*] Xu He[2,*] Songlin Tang[1,*] Haoxian Zhang[1,*,✉]
Qingfeng Li[1] Xiaoqiang Liu[1] Pengfei Wan[1] Kun Gai[1]
[1]Kuaishou Technology [2]Tsinghua University
[*]Equal contribution [✉]Corresponding author

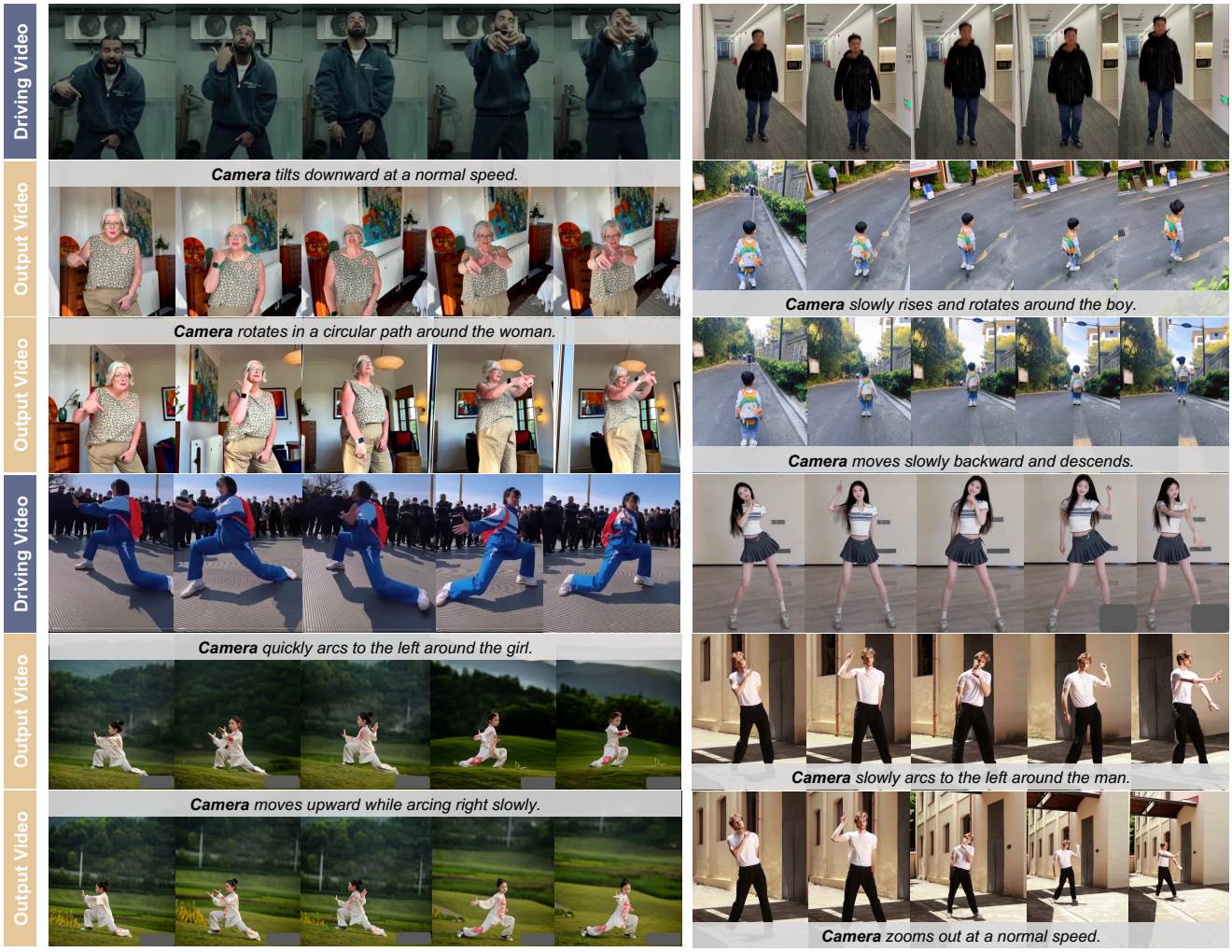

Figure 1. **3DiMo** can faithfully reproduce the 3D spatial motion from a 2D driving video, supporting flexible text-guided camera control.

Abstract

Existing methods for human motion control in video generation typically rely on either 2D poses or explicit 3D parametric models (e.g., SMPL) as control signals. However, 2D poses rigidly bind motion to the driving viewpoint, precluding novel-view synthesis. Explicit 3D models, though structurally informative, suffer from inherent inaccuracies (e.g., depth ambiguity and inaccurate dynamics) which, when used as a strong constraint, override the powerful intrinsic 3D awareness of large-scale video generators. In this work, we revisit motion control from a 3D-aware perspective, advocating for an implicit, view-agnostic motion representation that naturally aligns with the generator's spatial priors rather than depending on externally reconstructed constraints. We introduce 3DiMo, which jointly trains a motion encoder with a pretrained video generator to distill driving frames into compact, view-agnostic motion to-

kens, injected semantically via cross-attention. To foster 3D awareness, we train with view-rich supervision—single-view, multi-view, and moving-camera videos—forcing motion consistency across diverse viewpoints. Additionally, we use auxiliary geometric supervision that leverages SMPL only for early initialization and is annealed to zero, enabling the model to transition from external 3D guidance to learning genuine 3D spatial motion understanding from the data and the generator's priors. Experiments confirm that 3DiMo faithfully reproduces driving motions with flexible, text-driven camera control, significantly surpassing existing methods in both motion fidelity and visual quality. More demos and resources are available at https://hjrphoebus.github.io/3DiMo.

1. Introduction

Recent advances show that large-scale video generation models possess strong 3D spatial awareness and motion reasoning [29, 32], enabling text-guided novel-view synthesis and human reposing with consistent 3D geometry [11, 13, 30]. Meanwhile, controllable video generation has emerged as a research focus, with human motion control being one of its central challenges, which aims to animate a reference image according to motion cues from a driving video. Existing approaches typically extract 2D-rendered pose images from driving frames [4, 5, 7, 35, 37] and inject them via pixel-aligned conditioning [9, 34]. However, such 2D conditioning rigidly binds motion to the driving viewpoint, preventing the model from reasoning about motion in its inherently 3D nature. As a result, generated videos collapse to the 2D projection of the driving view, losing viewpoint flexibility and limiting applications such as novel-view human synthesis or cinematic camera motion.

In this work, we revisit the task of human motion control from a 3D-aware perspective. Our goal is to enable the model to reproduce the underlying 3D motion implied in 2D driving frames, while maintaining independent, text-guided camera control during generation. To achieve similar goals, recent works attempt to explicitly separate motion and camera control via 3D reconstruction—recovering SMPL(-X) [16, 21] sequences from driving videos and conditioning generation through mesh rendering [4] or projected keypoints [5] under predefined camera trajectories. This line of work indeed moves beyond 2D constraints by recognizing that human motion inherently occurs in 3D space. However, it still faces a fundamental limitation: the motion representation is fully determined by externally estimated parametric reconstructions such as SMPL. These estimates, while structurally stable, suffer from depth ambiguities [8] (e.g., forward tilting, incorrect inter-limb contact, or distorted Z-axis motion), and their inaccurate reconstruction of natural motion dynamics further limits expressiveness. When such biased 3D signals are injected into the generator—especially through rigid projection-based 2D alignment—they impose strong geometric constraints that override native 3D priors of large-scale video models, ultimately limiting the generator's ability to produce spatially coherent and physically plausible motion.

To address these limitations, we propose a new paradigm of *implicit 3D reasoning for motion control* that leverages the video generator's intrinsic spatial and motion understanding. We argue two key principles. First, we advocate for end-to-end learning of a motion encoder jointly with the generator, extracting implicit 3D motion representations directly from 2D frames while naturally aligning with the model's spatial priors, which requires encoder designs that encourage view-agnostic motion discovery with semantic conditioning rather than rigid projection. Second, effective 3D awareness demands supervision beyond conventional same-view reconstruction, which merely learns 2D projection patterns. Instead, view-rich data across diverse viewpoints and camera trajectories forces the extraction of the essential 3D spatial motion. Unlike works such as Uni3C [4], which emphasize precise camera trajectory control in animation scenarios, our work focuses on modeling and reproducing 3D motion from 2D observations. In our framework, camera control is not an explicit objective but a natural byproduct of the model's learned 3D awareness. We therefore leverage the generator's native text-driven camera control—rather than predefined camera parameters—as it both aligns with the model's intrinsic spatial understanding and serves as evidence of whether genuine 3D awareness has emerged in the learned motion representations.

Building on this paradigm, we present **3DiMo**, an end-to-end framework to learn **3D**-aware **i**mplicit **mo**tion control for view-adaptive human video generation under view-rich supervision. Specifically, we design a Transformer-based motion encoder that distills 2D driving frames into compact 1D tokens, intentionally discarding spatial layout to encourage viewpoint-agnostic, semantic motion abstraction. The encoder is jointly optimized with a pretrained video generator to align with the generator's generative capability. The resulting motion tokens are injected through cross-attention, enabling flexible semantic conditioning in place of rigid projection-based alignment. To achieve genuine 3D awareness, we collect and train on a comprehensive view-rich dataset spanning single-view, multi-view, and moving-camera videos. Each clip produces motion cues that condition the generator, which is supervised to either reconstruct the same video or reproduce the motion from alternative viewpoints or camera trajectories guided by text prompts. This dual-objective training encourages the emergence of expressive, 3D-aware motion representations.

To accelerate spatial understanding during early training, we further introduce lightweight auxiliary decoders that provide geometric supervision by aligning motion features

with parametric 3D reconstruction results (*i.e.*, SMPL and MANO [25]). Although these external estimates are imperfect, they supply 3D human priors that offer a reliable initialization. As training progresses, the auxiliary loss is gradually annealed to zero, allowing the model to shift from externally guided geometry to the generator's inherent 3D priors and the richness of view-rich data—ultimately enabling expressive and truly 3D-aware motion representations.

Our contributions can be summarized as:

- **3D-aware motion control.** We reformulate human motion control for video generation as a 3D-aware task that recovers underlying 3D motion from 2D frames while naturally supporting flexible text-driven camera control.
- **End-to-end implicit motion framework.** We propose **3DiMo**, an end-to-end framework that jointly learns a view-agnostic implicit motion encoder with a powerful pretrained DiT-based video generator. This design encourages motion representations that align with the generator's intrinsic 3D spatial priors and enables semantically rich motion conditioning via cross-attention.
- **View-rich supervision for 3D learning.** We collect a large-scale human motion dataset spanning single-view, multi-view, and moving-camera videos, enabling viewpoint-agnostic 3D motion learning aligned with the generator's inherent 3D reasoning.

Through extensive experiments, we demonstrate that 3DiMo faithfully reproduces driving motions while preserving 3D consistency across varying viewpoints, validating that the learned motion representations are both expressive and 3D-aware, effectively conditioning the DiT-based video model to generate high-fidelity human motion videos.

2. Related Work

Diffusion-Based Video Generation. Diffusion models have achieved remarkable success in high-fidelity image and video synthesis. Latent Diffusion Models (LDMs) [3, 24] improve efficiency by operating in compressed latent spaces, while DiT-based architectures [22] further enhance scalability and spatiotemporal consistency for video generation. Recent advances [12, 13, 19, 30] show that large-scale pretrained video diffusion models exhibit strong awareness and reasoning capabilities over both dynamics and 3D space [29, 32]. In this work, we focus on learning an implicit 3D-aware motion representation that aligns with a pretrained video generator's spatial and motion priors—thereby eliciting their intrinsic 3D understanding and enabling high-quality, spatially consistent human animation.

Motion Control for Human Image Animation. Human image animation aims to animate a reference image according to motion cues from a driving video, as pioneered by early works such as FOMM [26] and MRAA [27]. Recent diffusion-based approaches [9, 35] achieve impressive quality by injecting explicit motion signals (e.g., 2D poses or DensePose), but their 2D formulations inherently lose spatial information, causing depth ambiguities. To overcome this, 3D-based methods [4, 5, 8, 17, 37] introduce SMPL [16] or SMPL-X [21] models as control conditions—typically rendered or projected into 2D space, or mapped as camera-space joint trajectories for explicit 3D control. However, such approaches rely heavily on externally reconstructed representations, which, though structurally stable, lack the expressiveness and 3D reasoning priors present in large-scale pretrained video generators. In contrast, we learn an implicit, end-to-end motion representation aligned with these priors to achieve expressive and 3D-aware motion modeling. While X-Nemo [36] and X-UniMotion [28] explore implicit motion representations, they remain limited to 2D spatial patterns and cannot generalize to true 3D motion or camera control—challenges that our work directly addresses.

3. Our Approach

Given a reference image $\boldsymbol{I}_R$ of a subject and a driving video $\boldsymbol{V}_D = \{\boldsymbol{I}_D^t\}_{t=0}^T$ providing motion cues, our proposed 3D-aware motion control aims to transfer the driving video's motion—which inherently exists in 3D space—to the reference subject, while preserving flexible, text-guided camera control. This task is highly challenging as human motion and camera trajectories are entangled within the 2D projections of driving frames, obscuring the underlying 3D motion essence that truly exist in physical space.

To achieve this goal, we leverage a pretrained DiT-based video generation model with rich 3D spatial and motion priors as our backbone, which generates videos from a reference image guided by text prompts (Sec. 3.1). The core of our framework lies in an implicit motion encoder jointly optimized with the pretrained video generator, which distills view-agnostic motion tokens from 2D driving frames and injects them via cross-attention for semantical motion control compatible with text-driven camera manipulation (Sec. 3.2). To endow the learned motion representation with 3D awareness, we train our framework on a view-rich human video dataset encompassing diverse viewpoints and camera movements, supplemented with auxiliary decoders that provide early-stage geometric alignment supervision to accelerate spatial understanding (Sec. 3.3).

3.1. Preliminary

Video Generation Backbone. Our video generation model adopts the latent diffusion model (LDM) paradigm, utilizing a causal 3D VAE for video compression into latent space and a generative backbone for latent sequence modeling. The backbone is a DiT-based architecture [6] pretrained on text-to-video and image-to-video tasks, comprising multiple DiT blocks that interleave full self-attention and Feed-Forward Networks (FFN). The reference image is

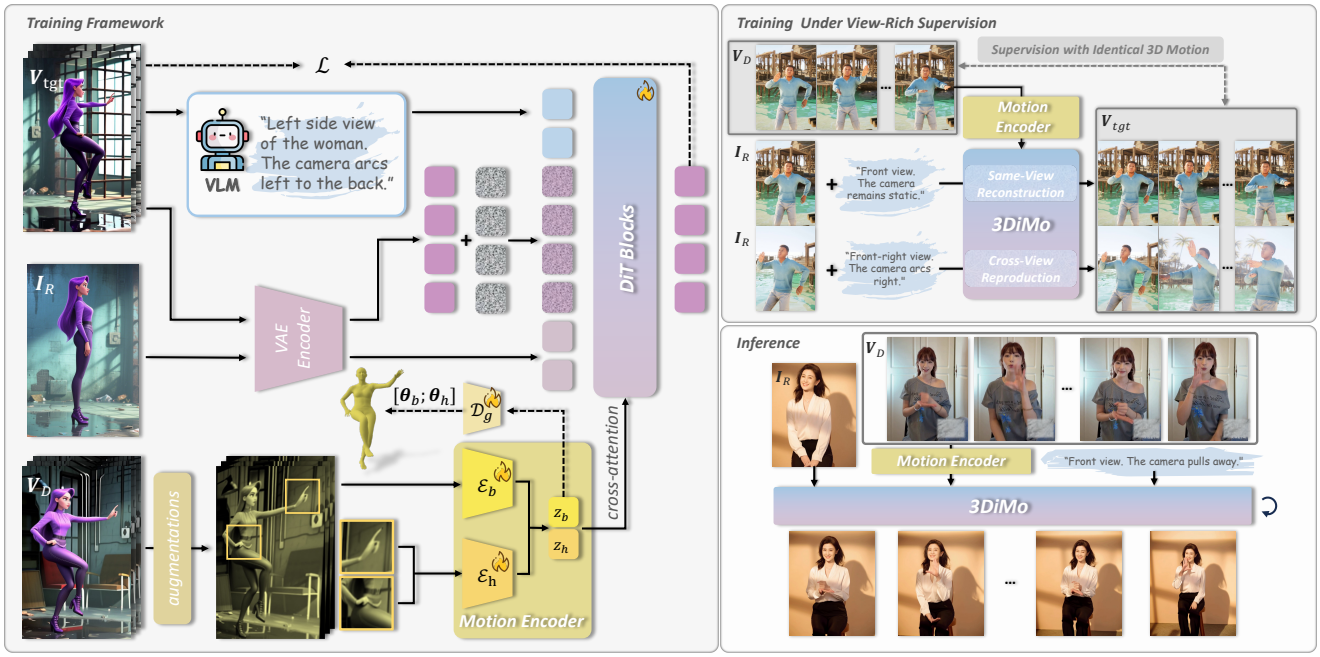

Figure 2. **Overview of 3DiMo.** Our framework consists of end-to-end trained motion encoders—$\mathcal{E}_b$ for the body and $\mathcal{E}_h$ for hands—and an DiT-based video generator. Given a reference frame $\boldsymbol{I}_R$ and a driving video $\boldsymbol{V}_D$, driving frames are first augmented with random perspective transformations before being encoded by the motion encoder to extract view-agnostic motion representations. These resulting features are then injected into the generator through cross-attention, enabling the model to synthesize a target sequence $\boldsymbol{V}_{\text{tgt}}$ that reenacts the same underlying 3D motion while preserving flexible text-driven camera control. To facilitate 3D-aware learning, we introduce early-stage auxiliary geometric supervision by regressing the encoded motion to external parametric reconstruction results θ_b and θ_h. During training, view-rich data is used to jointly supervise same-view reconstruction and cross-view motion reproduction, driving the emergence of expressive and 3D-aware motion representations. At inference, motion tokens extracted directly from 2D driving frames provide rich 3D spatial cues that can animate any reference character, supporting high-fidelity and view-adaptive motion-controlled video generation.

incorporated by concatenating its latent tokens with noised video tokens, facilitating cross-modal interaction among reference, video, and text tokens through the full self-attention. During training, we adopt a flow-based diffusion process [14, 15] that progressively adds noise to the target video latents, and the model is optimized using a v-prediction objective.

Parametric 3D Human Model. SMPL [16] represents a full-body human mesh using a compact set of parameters, including shape coefficients $\boldsymbol{\beta}_b$ and pose parameters $\boldsymbol{\theta}_b$ that describe the articulated body configuration. Similarly, MANO [25] models the articulated hand using an analogous formulation, with hand-shape parameters $\boldsymbol{\beta}_h$ and hand-pose parameters $\boldsymbol{\theta}_h$. Despite their well-known limitations in expressiveness and depth ambiguity, these parametric models provide robust 3D geometric priors that we leverage for early-stage auxiliary supervision.

3.2. End-to-end Framework with Implicit View-Agnostic Motion Control

As illustrated in Fig. 2, our framework features a motion encoders that extracts motion representations $\boldsymbol{z}$ from the driving video $\boldsymbol{V}_D$, which condition a pretrained DiT-based video generator. The reference image $\boldsymbol{I}_R$ and accompanying text prompt T are also provided as token sequences to the video generator, which produces an output video depicting the reference subject in $\boldsymbol{I}_R$ reenacting the motion from $\boldsymbol{V}_D$ under camera trajectories guided by T. Unlike previous methods that rely on external pose estimation, our framework jointly optimizes the motion encoder and the video generator in an end-to-end manner, learning semantically rich motion representations naturally aligned with the generator's inherent spatial and motion priors.

Implicit Motion Encoder. The core insight of our motion encoder design is to encourage view-agnostic representation learning that captures the semantically rich dynamics of 3D human motion, going beyond the superficial patterns observable in 2D projections. Following [33], we design our motion encoder as a Transformer-based 1D tokenizer. Each driving frame is patchified into visual tokens and concatenated with $K(= 5$ in our work) learnable latent tokens, which interact through several attention layers. Only the output latent tokens are retained as the motion representation. By compressing into compact 1D motion tokens, we

enforce a semantic bottleneck that eliminates 2D structural information, including both appearance details and view-specific pose configurations, while focusing on the intrinsic semantics of spatial motion.

To encourage view-agnostic motion representation learning, we apply random perspective transformations to the driving frames before motion encoding to introduce motion-invariant augmentations, which to some extent decouples the spatial motion from its view-specific 2D projection. Additionally, similar to [28, 36], we employ appearance augmentations (*e.g.*, color jittering and lightweight spatial transforms) to prevent identity leakage from the driving frames. Through this design, the motion encoder is encouraged to focus purely on the intrinsic dynamics of 3D spatial motion while avoiding both appearance leakage and view-specific pose constraints.

View-Agnostic Cross Attention Conditioning. Instead of converting our motion representations into view-dependent 2D spatially-aligned control using explicit camera parameters as in most existing works, we simply employ cross-attention to inject motion representations directly into the generator. This achieves flexible semantic-level interaction between visual and motion modalities without rigid spatial constraints. To be specific, we append a cross-attention layer after each full self-attention in the DiT generator, where only video tokens attend to the motion tokens, while text tokens remain unchanged.

Text-Guided Camera Control. Our view-agnostic motion representations and semantic-level conditioning naturally coexist with the generator's native text-driven camera control capability. Consequently, beyond motion control, our framework readily supports flexible viewpoint manipulation by simply augmenting the text prompt T with camera-movement descriptions, which interact with visual tokens through the same native mechanism as in the original DiT-based generator.

Dual-Scale Motion Encoding. Considering that a single compact representation struggles to capture both global body movements and fine-grained hand dynamics, we follow [28] and employ two motion encoders: a body encoder $\mathcal{E}_b$ for coarse body motion and a hand encoder $\mathcal{E}_h$ for detailed gestures. The resulting motion tokens are concatenated and jointly injected into the generator via cross-attention, enabling unified motion control over both full-body articulation and fine-grained hand movements.

3.3. 3D-Aware Training with View-Rich Data

While our motion encoder with distilled 1D tokens effectively filters redundant 2D spatial structures and captures semantically rich motion dynamics for expressive motion generation, this design alone does not guarantee genuine 3D motion understanding. When trained solely through same-view reconstruction, the model can achieve satisfactory results by merely learning view-dependent 2D motion patterns, as reproducing motion under identical viewpoints requires no true spatial reasoning. This reveals a fundamental limitation: *without more challenging supervision, the model lacks incentive to develop 3D awareness beyond the 2D projection domain.* To overcome this issue, we introduce view-rich data supervision that imposes a more demanding learning objective, compelling the model to reason about motion as it truly occurs in 3D space, invariant to viewpoint changes rather than as isolated 2D observations.

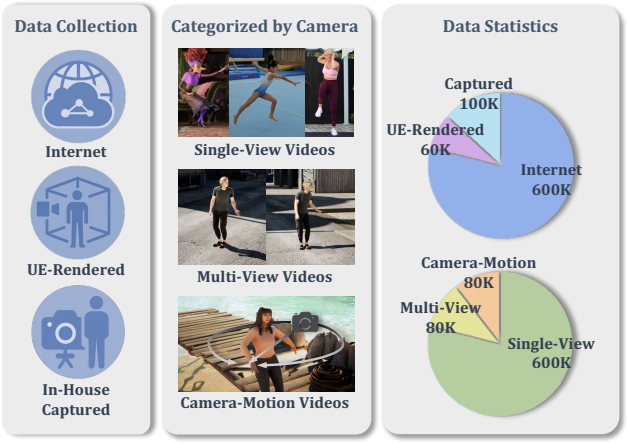

Figure 3. **Our collected view-rich dataset** combines internet videos, UE renderings, and in-house captures, covering camera categories including single-view, multi-view, and camera-motion sequences. High-quality large-scale single-view footage exposes the model to diverse human motions, while complementary multi-view data provides consistent cross-view observations that are crucial for learning genuine 3D-aware motion representations.

View-Rich Dataset Construction. To enable comprehensive supervision for learning both expressive and 3D-aware motion representations, we construct a large-scale dataset with diverse camera configurations. From the perspective of supervision objectives, our data serves three distinct purposes as shown in Fig. 3: 1) same-view reconstruction, where each motion-viewpoint pair is unique, enabling self-supervised learning of expressive motion dynamics; 2) multi-view motion reproduction, utilizing synchronized captures from fixed camera arrays for identical motions to enforce consist 3D motion learning across viewpoints; and 3) motion reproduction under moving cameras, featuring identical motions captured with different camera trajectories to decouple motion from viewpoint changes and support text-guided camera control.

To balance realism, diversity, and 3D consistency, we integrate three complementary data sources: 1) large-scale internet videos that provide diverse human motions for learning realistic dynamics, though limited to single-view supervision; 2) synthetic sequences rendered with Unreal Engine

5 (UE5) from [1, 18, 31], which offer precise motion captures under varied camera trajectories, despite potential domain gaps from real-world videos; 3) real-world multi-view captures, including both open-source datasets and our proprietary recordings, combining fixed multi-camera arrays and dynamic camera trajectories, which provide authentic 3D supervision within the real-world video domain. The detailed composition of our dataset is illustrated in Fig. 3. In practice, internet videos dominate in scale and drive the learning of natural, expressive motion patterns, whereas the multi-view and camera-trajectory data, though smaller in quantity, play a crucial role in fostering genuine 3D spatial understanding.

To obtain text descriptions for camera viewpoints and movements, we employ Qwen2.5-VL [2] to both annotate internet videos and convert predefined camera configurations from synthetic and real-world captured data into unified text prompts.

Training Strategies. With the constructed dataset, we enable 3D-aware training under view-rich supervision that includes both reconstruction and cross-view motion reproduction objectives. Specifically, given a driving video $\boldsymbol{V}_D$, we supervise the model output with either $\boldsymbol{V}_D$ itself (reconstruction) or corresponding videos $\boldsymbol{V}'_D$ of the same motion captured from a different viewpoint or camera trajectory (cross-view reproduction). The reference image is taken as the first frame of the supervision target, which automatically aligns the generated motion with the reference subject's facing direction—eliminating the need for explicit SMPL-to-image alignment or camera regression as in [4].

In practice, we adopt a progressive multi-stage training strategy with varying data mixtures based on supervision objectives. In the first stage, we exclusively use single-view data for self-reconstruction, exposing the model to diverse and expressive motion dynamics and enabling stable initialization of implicit motion learning. The second stage introduces a balanced mixture of reconstruction and cross-view motion reproduction, gradually transitioning the learned representations from 2D dynamics toward 3D spatial semantics. Finally, the third stage focuses entirely on multi-view and camera-motion data to strengthen the view-agnostic nature of the learned motion features and enhance compatibility with flexible text-guided camera control.

Geometric Supervision with Auxiliary Decoders. In our early experiments, we observed that direct end-to-end training often leads to slow and unstable convergence, especially after introducing cross-view supervision. This partly arises because the diffusion loss distributes uniformly across pixels, lacking targeted emphasis on motion-specific semantics. Moreover, the powerful DiT backbone tends to exploit its inherent motion priors to generate plausible videos from single images during early training, thereby reducing reliance on the encoded motion representations, which consequently receive weak gradient feedback.

To address this challenge, we introduce auxiliary geometric supervision to facilitate motion representation learning. Specifically, we employ a lightweight MLP-based geometric decoder $\mathcal{D}_g$ that processes the concatenated motion representations $\mathbf{z} = [\mathbf{z}_b; \mathbf{z}_h]$ to predict pose parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_b; \boldsymbol{\theta}_h]$, using pseudo ground-truth annotations derived from off-the-shelf SMPL and MANO estimators [20, 23]. Notably, we exclude the global root orientation during supervision to ensure view-agnostic learning.

Despite limitations in depth ambiguity and expressiveness, this parametric 3D geometric supervision effectively transfers robust spatial motion priors through the lightweight, easily optimized auxiliary decoder, providing a well-initialized motion distribution for subsequent learning. In practice, we apply auxiliary supervision during the first stage and the early part of the second stage, with its loss weight annealed progressively as training proceeds. The supervision is then completely removed for the remaining steps of the second stage and the entirety of the third stage. This schedule allows the model to evolve from geometry-guided initialization to learning motion representations that align with the DiT's perceptual and generative capabilities, ultimately achieving superior 3D-aware motion understanding supported by view-rich supervision.

4. Experiments

4.1. Experimental Setups

Implementation Details. We train our 3DiMo on the dataset described in Sec. 3.3, using 121-frame video clips resized to a target area of 480×854 pixels while preserving original aspect ratios. Training is performed with a total batch size of 64 using the Adam optimizer with a learning rate of 1e-5. The three training stages run for 10K, 15K, and 5K steps, completing in approximately three days.

Evaluation Data and Metrics. We evaluate our method on 50 videos from TikTok dataset [10] and 100 videos collected from the internet. Following [9], we use PSNR, SSIM, LPIPS, and FID to measure per-frame visual quality, and adopt FVD to evaluate the overall video fidelity.

4.2. Quantitative Evaluation

We compare 3DiMo with state-of-the-art human image animation methods, including both 2D pose-based approaches (AnimateAnyone [9], MimicMotion [35]) and 3D SMPL-based methods (Uni3C [4], MTVCrafter [5]), to validate the effectiveness of our implicit 3D-aware motion modeling and reenactment. Considering that most existing baselines do not support camera manipulation, we evaluate 3DiMo under prompts specifying a *static camera*.

As shown in Tab. 1 (Col. 2-6), our method surpasses all baselines on LPIPS, FID, and FVD, indicating superior vi-

Table 1. Quantitative evaluation and user study results of MOS with 95% confidence intervals. Top two are noted as first , second .

Method	Quantitative Evaluation					User Study			
	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	FVD ↓	Accuracy ↑	Naturalness ↑	3D Plausibility ↑	Overall ↑
AnimateAnyone	0.7325	17.21	0.2754	68.72	862.5	4.13±0.12	4.00±0.09	3.20±0.14	3.76±0.10
MimicMotion	0.7051	16.83	0.3286	62.45	628.2	3.84±0.14	4.15±0.06	3.01±0.12	3.84±0.10
MTVCrafter	0.7489	18.03	0.2542	57.21	379.6	3.69±0.09	3.98±0.10	3.69±0.12	4.19±0.09
Uni3C	0.7185	17.53	0.2639	41.28	321.9	3.72±0.06	4.03±0.06	3.97±0.10	4.24±0.06
Ours	0.7390	17.96	0.2206	36.92	297.4	4.28±0.08	4.18±0.06	4.05±0.09	4.38±0.08

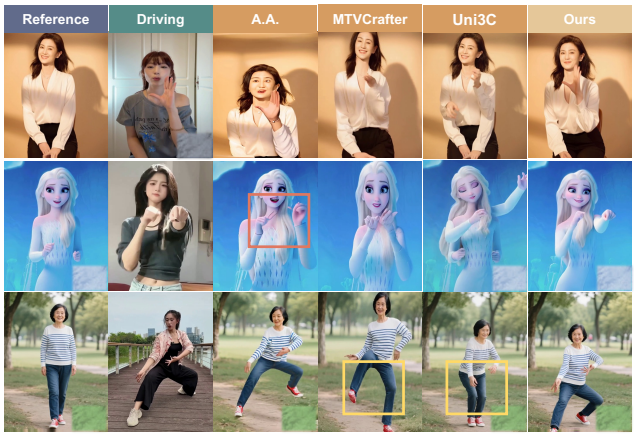

Figure 4. **Visualization comparisons with baselines.** Red and yellow bounding boxes highlight depth ambiguities and inaccurate poses, respectively. "A.A." denotes AnimateAnyone. Our method produces accurate and 3D-plausible motion reenactment videos.

sual quality and motion control. Although our SSIM and PSNR are slightly lower than MTVCrafter, this is expected: these pixel-wise metrics are sensitive to minor viewpoint deviations while some evaluation videos contain weak, unintended camera motions. Competing methods mainly operate within a 2D alignment paradigm and simply reproduce these motions, while our text-driven *static camera* prompt suppresses such drift to maintain geometric consistency. This yields perceptually better results but introduces small pixel-wise discrepancies from the ground truth.

4.3. Qualitative Evaluation

Comparison with SOTAs. Fig. 4 presents the quantitative comparisons, where we configure our method with a static camera as described in Sec. 4.2. Our implicit 3D-aware approach achieves precise motion control and expressive dynamics, producing high-fidelity and physically plausible human videos. In contrast, 2D pose-based methods often yield incorrect limb depth ordering due to their lack of geometric awareness, while SMPL-based approaches struggle to maintain accurate pose estimation and control under complex motions. More visualization results and extended comparisons are provided in the supplementary material.

Table 2. Ablation results. Top two are noted as first , second .

Method	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	FVD ↓
w/ SMPL ctrl.	0.724	17.1	0.238	39.7	348.2
w/ stage 1 only	0.745	18.3	0.220	40.5	305.4
w/ stage 1 & 2	0.723	17.9	0.221	38.2	314.5
w/ channel concat.	0.703	16.8	0.304	48.2	395.6
w/o geo. superv.	0.684	15.8	0.347	51.3	383.1
w/o hand enc.	0.726	17.5	0.234	38.1	298.7
Full Model	0.739	18.0	0.221	36.9	297.4

More Results of View-Adaptive Motion Control. To further demonstrate our model's capability in modeling 3D motion, Figure Fig. 1 showcases several motion control results under diverse text-guided camera configurations. Our approach inherits the DiT backbone's native ability for text-driven camera manipulation, while simultaneously achieving precise video-driven motion control that preserves physical plausibility and spatial consistency across dynamic camera trajectories and varying viewpoints.

User Study. We further conduct a user study involving 30 participants, where each participant evaluates 10 cross-identity animation videos generated by each method. We collect mean opinion scores (MOS) based on a 5-point Likert scale across four aspects: motion accuracy, motion naturalness, 3D physical plausibility, and overall visual quality. The results, summarized in Tab. 1 (Col. 7-10), show that our method consistently outperforms all existing baselines, especially in motion naturalness and physical plausibility, which emphasize spatial relationships and realistic dynamics. These strongly support that our learned motion representation, aligned with the large-scale pretrained video generator's spatial and motion priors, achieves more expressive and 3D-aware motion modeling and control compared to approaches that rely on externally predefined parameters.

4.4. Ablation Study and Analysis

SMPL *vs*. Our Implicit Motion Representations. The core of our approach lies in the end-to-end learning of an implicit 3D-aware motion representation. For comparison, we also consider a variant that directly uses SMPL

Figure 5. **Visualizations of ablation results.** Using SMPL poses as motion representation introduces typical depth ambiguity errors. Removing any view-rich data supervision impairs camera control. Removing auxiliary geometric supervision or using channel concatenation causes training instability and quality degradation. Without the hand encoder, fine-grained hand motions are lost.

pose coefficients θ_{body} as the motion representation, mapped through an MLP to match the token dimensionality before being injected into the generator via the same cross-attention mechanism.

As shown in Fig. 5, for a frontal one-hand-on-hip driving motion, the SMPL-based variant fails to maintain correct hand-hip contact from the side view. In contrast, our learned motion representation correctly preserves this physical relationship, effectively resolving the depth ambiguity that commonly occurs in parametric reconstructions. This demonstrates that our motion representation, distilled from the pretrained video generator, exhibits superior 3D spatial awareness aligned with real-world priors compared to off-the-shelf parametric reconstruction.

Multi-Stage View-Rich Learning. We further analyze the effect of our multi-stage training strategy under a leftward arcing camera trajectory. In the first stage with single-view reconstruction, the model learns diverse motion patterns but tends to collapse into 2D projections, often failing to follow text-guided camera motions. Introducing view-rich data in the second stage helps decouple motion and viewpoint, establishing initial 3D awareness; however, the camera motion sometimes only affects the background while the subject remains front-facing, indicating incomplete semantic interaction between motion and camera control. The third-stage refinement, trained solely on view-rich data, further strengthens this interaction, enabling the model to accurately follow camera trajectories while maintaining spatially consistent human motion and background rendering.

Conditioning Mechanism. We replace cross-attention with channel concatenation as an alternative conditioning approach. This substitution shows significantly degraded motion control capability, demonstrating that cross-attention is more suitable for semantic-rich interaction between our motion representations and the generator.

Auxiliary Geometric Supervision. Removing the auxiliary geometric supervision during early training leads to unstable convergence and collapsed motion control, demonstrating that this supervision provides crucial initialization for learning meaningful motion representations.

Dual-scale motion encoders. Omitting the hand motion encoder results in loss of fine-grained hand control, confirming its necessity for complete motion representation.

Tab. 2 further shows that removing most components consistently degrades visual quality, confirming their necessity for producing high-fidelity human motion videos. Discarding the last two view-rich training stages yields a slight but negligible improvement in visual metrics; however, these stages are essential for endowing the model with genuine 3D-aware motion understanding—precisely the core capability required to address our proposed 3D-aware motion control problem.

5. Conclusion

In this work, we present 3DiMo, an end-to-end framework for 3D-aware human motion control that learns to align with a video generator's intrinsic spatial priors rather than relying on explicit 3D parametric reconstruction. By jointly training a motion encoder with the pretrained video generation model and designing it to discard view-dependent layouts, 3DiMo aligns the two models and establishes the capacity for view-agnostic motion representation. When further trained under view-rich data supervision, the framework internalizes 3D spatial motion understanding directly from 2D observations. Combined with lightweight geometric initialization that is gradually annealed away, 3DiMo ultimately develops robust and expressive 3D-aware motion representations without relying on external estimates. Experiments demonstrate that 3DiMo faithfully reproduces driving motions under flexible text-driven camera control and consistently outperforms both 2D- and 3D-based baselines in motion fidelity and visual quality.

References

[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 3

[4] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025. 2, 3, 6

[5] Yanbo Ding, Xirui Hu, Zhizhi Guo, Chi Zhang, and Yali Wang. Mtvcrafter: 4d motion tokenization for open-world human image animation. *arXiv preprint arXiv:2505.10238*, 2025. 2, 3, 6

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[7] Xu He, Qiaochu Huang, Zhensong Zhang, Zhiwei Lin, Zhiyong Wu, Sicheng Yang, Minglei Li, Zhiyi Chen, Songcen Xu, and Xiaofei Wu. Co-speech gesture video generation via motion-decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2273, 2024. 2

[8] Xu He, Zhiyong Wu, Xiaoyu Li, Di Kang, Chaopeng Zhang, Jiangnan Ye, Liyang Chen, Xiangjun Gao, Han Zhang, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3437–3445, 2025. 2, 3

[9] 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*, pages 8153–8163, 2024. 2, 3, 6

[10] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 6

[11] 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. 2

[12] Kuaishou. Kling ai. https://klingai.kuaishou.com/, 2024. Accessed: 2025-05-19. 3

[13] Xuanyi Li, Daquan Zhou, Chenxu Zhang, Shaodong Wei, Qibin Hou, and Ming-Ming Cheng. Sora generates videos with stunning geometrical consistency. *arXiv preprint arXiv:2402.17403*, 2024. 2, 3

[14] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[15] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 4

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 4

[17] Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, and Tianshu Hu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2025. 3

[18] 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 *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–10, 2025. 6

[19] Midjourney. Midjourney. https://www.midjourney.com, 2024. Accessed: 2024. 3

[20] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025. 6

[21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[23] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 6

[24] 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*, pages 10684–10695, 2022. 3

[25] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3, 4

[26] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for

image animation. *Advances in neural information processing systems*, 32, 2019. 3

[27] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13653–13662, 2021. 3

[28] Guoxian Song, Hongyi Xu, Xiaochen Zhao, You Xie, Tianpei Gu, Zenan Li, Chenxu Zhang, and Linjie Luo. X-unimotion: Animating human images with expressive, unified and identity-agnostic motion latents. *arXiv preprint arXiv:2508.09383*, 2025. 3, 5

[29] Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *arXiv preprint arXiv:2511.04570*, 2025. 2, 3

[30] 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. 2, 3

[31] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 6

[32] 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. 2, 3

[33] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024. 4

[34] 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*, pages 3836–3847, 2023. 2

[35] 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. *arXiv preprint arXiv:2406.19680*, 2024. 2, 3, 6

[36] Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. *arXiv preprint arXiv:2507.23143*, 2025. 3, 5

[37] 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 *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 2, 3