# EgoControl: Controllable Egocentric Video Generation via 3D Full-Body Poses

Enrico Pallotta[*1,2] Sina Mokhtarzadeh Azar[*1,2] Lars Doorenbos[1,2] Serdar Ozsoy[1,2]
Umar Iqbal[3] Juergen Gall[1,2]
[1]University of Bonn [2]Lamarr Institute for Machine Learning and Artificial Intelligence [3]NVIDIA
[*]Equal Contribution

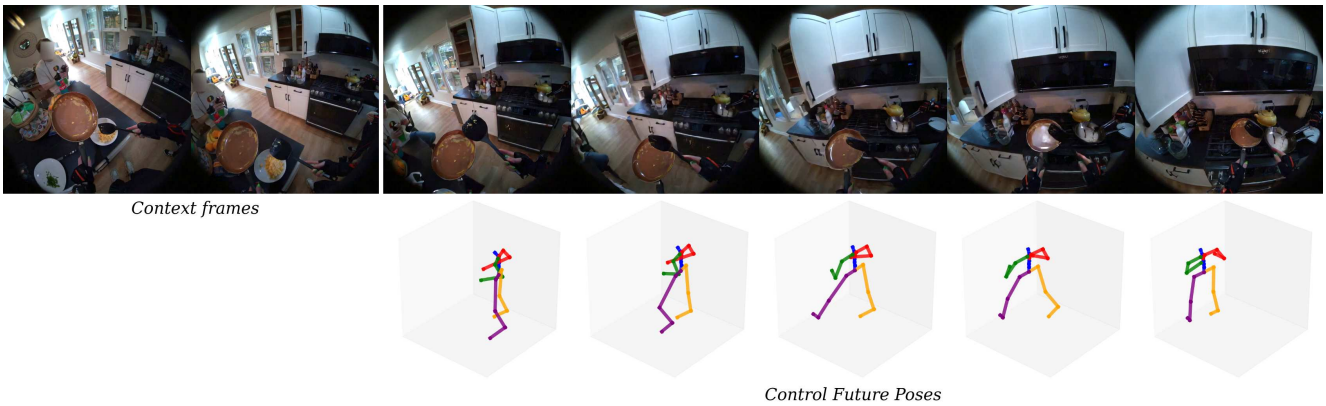



Figure 1. EgoControl generates highly realistic egocentric videos from observed context frames and allows detailed control over the body of the camera wearer. The camera view of the generated frames follows the head pose and the visible body parts of the human pose interact with the objects in the scene in a plausible manner.

## Abstract

*Egocentric video generation with fine-grained control through body motion is a key requirement towards embodied AI agents that can simulate, predict, and plan actions. In this work, we propose EgoControl, a pose-controllable video diffusion model trained on egocentric data. We train a video prediction model to condition future frame generation on explicit 3D body pose sequences. To achieve precise motion control, we introduce a novel pose representation that captures both global camera dynamics and articulated body movements, and integrate it through a dedicated control mechanism within the diffusion process. Given a short sequence of observed frames and a sequence of target poses, EgoControl generates temporally coherent and visually realistic future frames that align with the provided pose control. Experimental results demonstrate that EgoControl produces high-quality, pose-consistent egocentric videos, paving the way toward controllable embodied video simulation and understanding.*

## 1. Introduction

Egocentric (first-person) vision has recently become a central topic in computer vision [39, 48]. Unlike third-person imagery, egocentric video captures the sensorimotor experience of an agent, characterized by rapid camera motion, frequent occlusions from hands and objects, and a tight coupling between body pose and visual observations. This perspective is critical for problems where perception and action influence each other, such as embodied navigation, manipulation, augmented reality, and human–robot collaboration. Consequently, the community has invested heavily in large-scale egocentric datasets and benchmarks [8, 15, 31], as well as models that can understand [25, 28] and predict [14] first-person behavior. These efforts have made clear that modeling the interplay between body motion and visual appearance is both highly challenging and practically important.

For an embodied agent, the ability to *forecast* future egocentric frames is essential: it is a mechanism to internally simulate the visual consequences of planned actions. Predictive visual models support safer planning (by imagining risky outcomes), enable fine-grained action conditioning (by checking whether a planned motion will achieve a desired viewpoint or object configuration), and allow interactive systems (AR/VR, teleoperation, or games) to render expected views before they are executed. To be useful in these scenarios, generative models must not only be photorealistic but also controllable with fine-grained, physically

plausible body-level commands that reflect how a human or a robot would move.

Existing state-of-the-art generative video models [1, 51] are impressive at content synthesis and, more recently, at conditioning on high-level signals such as text prompts or camera trajectories [4, 18]. However, these conditioning modalities do not provide direct, explicit control over the camera wearer's articulated body. In the egocentric setting, this is a critical gap: camera motion is produced by the wearer's head global translation and rotation, while local articulated motions (arms, hands, legs) produce the occlusions and object interactions that define first-person scenes. Without the ability to specify full-body pose sequences, a model cannot reliably simulate the visual consequences of specific actions (e.g., reaching, turning while walking, or performing a hand-object manipulation).

In this work, we introduce EgoControl, the first egocentric video generative model that enables explicit control over the camera wearer's body through 3D full-body poses. EgoControl is a latent diffusion model trained to synthesize future egocentric frames conditioned on a short sequence of past observations and a sequence of target body poses. We design an informative pose representation that jointly captures global camera motion and articulated body dynamics, and integrate it via a dedicated control pathway within the diffusion process. This allows EgoControl to accurately translate body motion into realistic egocentric visual outcomes, as shown in Fig. 1.

To rigorously assess pose controllability, we establish a comprehensive evaluation protocol that measures visual fidelity, global motion accuracy, and body alignment. EgoControl achieves pose-consistent, high-resolution video generation. We further showcase qualitative examples demonstrating fine-grained control across diverse scenarios, including walking, turning, and hand-object interactions.

In summary, our main contributions are:

- We introduce EgoControl, a diffusion-based egocentric video generator conditioned on explicit 3D full-body pose sequences of the camera wearer.
- We propose a compact and informative pose representation, along with an integrated control mechanism that jointly encodes global (camera) and articulated (body) motion to condition the diffusion process.
- We present a thorough quantitative and qualitative evaluation protocol designed to measure visual fidelity, camera motion accuracy, and body alignment consistency.

## 2. Related Work

**Egocentric vision.** Egocentric vision has seen growing interest because first-person data provides a natural view of embodied behaviors and interactions. This surge motivated the creation of large, curated egocentric datasets and benchmarks for understanding and generation [8, 15, 27, 31]. Work in egocentric perception covers a range of tasks, including action recognition [25], representation learning [28], dense understanding [47], and pretraining for downstream tasks [26]. Another important thread focuses on action anticipation [14, 61] and early object interaction prediction from first-person streams [11]. Recent works address ego-pose estimation [37] and ego↔exo generation [29, 30, 58], which provide useful building blocks for egocentric generation pipelines.

**Video generation and prediction.** Recent years have seen rapid progress in video generative models, driven by the rise of powerful image [42] and video diffusion [20] approaches, as well as large-scale pretraining. Early efforts primarily focused on generating visually appealing videos from text prompts [5, 19, 45], while more recent approaches aim for higher temporal coherence and longer horizons. State-of-the-art models include large-scale diffusion [1, 60] and flow-matching frameworks [40, 51] built upon DiT architectures [38], which currently set the benchmark for quality and realism. In parallel, the task of video prediction, which involves forecasting future frames based on past observations, has been widely explored for modeling temporal dependencies and motion dynamics. Earlier approaches employed recurrent and convolutional architectures for short- and mid-term forecasting [13, 54], while more recent methods leverage generative diffusion models to capture uncertainty in stochastic motion [35, 50]. Motivated by similar principles as text-to-video generation, this line of work has inspired a growing family of image- and video-to-video generative models that condition on past frames, enabling more detailed and context-aware guidance during generation [7, 21, 34, 57].

**Controllable video generation.** Recent advances in video synthesis increasingly emphasize controllability, enabling explicit conditioning on signals such as camera motion, planned trajectories, object motion, and human pose. Several methods generate videos that follow high-level camera movements [16, 59] or specified camera trajectories to emulate first-person perspectives [4, 9, 18, 55, 63]. In autonomous-driving and navigation settings, trajectory-conditioned generators synthesize scene evolution from planned routes, demonstrating how long-range spatial cues can be integrated into diffusion pipelines [12, 17, 43]. Other approaches condition on object-centric trajectories [2, 44, 55] or motion sketches [46, 56] to achieve fine-grained control of dynamic elements within a scene. Human-body pose has also emerged as a natural and effective control modality for exocentric generation: prior work guides actor motion with 2D skeletal poses using cross-attention [32], concatenation with the input [62], or ControlNet-like mechanisms [6, 53]. Importantly, these prior methods operate

under a third-person regime where the subject is largely visible and camera motion is limited. Furthermore, they only consider 2D poses, which cannot be applied to egocentric videos. In contrast, controlling generation through the *egocentric* full body pose of the camera wearer presents a substantially more challenging scenario, characterized by strong viewpoint changes, frequent self-occlusions, and rich hand–object interactions. A first step in this direction is the concurrent work PEVA [3]. It is an auto-regressive approach that generates one frame conditioned on the 3D upper-body pose per step. The approach is very limited in terms of temporal and spatial resolution. Our work not only enables 3D full-body pose controlled egocentric video generation at higher resolution, but it also generates videos instead of single frames, achieving temporal coherence and pose fidelity under this challenging scenario.

# 3. EgoControl

As illustrated in Fig. 1, EgoControl is a video diffusion model that synthesizes egocentric videos conditioned on two factors: the past visual context $\mathbf{x}=(x_1, x_2, \ldots, x_N)$, i.e., the current observed state, and the future motion intent expressed by the human pose sequence $\mathbf{P}=(p_1, p_2, \ldots, p_M)$. The human pose sequence not only controls the camera view, which depends on the head pose in egocentric videos, but also the body parts that are visible in the synthesized frames $\mathbf{y}=(y_1, y_2, \ldots, y_M)$. The goal is thus to generate egocentric video frames $y_i \in \mathbf{y}$ that not only look realistic, but that are also visually coherent with their corresponding control pose $p_i \in \mathbf{P}$.

Our method is based on a latent conditional video diffusion model [1]. Initially, a tokenizer $\mathcal{E}$ maps frames to compact, continuous latents $z_0 = \mathcal{E}(x)$. While $q(z_t \mid z_0)$ denotes the forward perturbation process in latent space, $p_\theta(z_{t-1} \mid z_t, c)$ denotes the learned denoiser conditioned on some context $c$, e.g., $c = \mathbf{x}$. Following the EDM formulation [24], the model perturbs the clean latent using a continuous noise level $\sigma$:

$$z = z_0 + \sigma\,\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

and trains the denoiser to predict the original clean latent $z_0$ directly:

$$\mathcal{L}(\theta) = \mathbb{E}_{z_0,\varepsilon,\sigma,c}\big[\, w(\sigma)\, \|\, z_\theta(z, \sigma, c) - z_0 \,\|_2^2 \,\big], \tag{2}$$

where $w(\sigma)$ is a weighting function controlling the relative importance of each noise level and $z_\theta$ is the denoising network, implemented as a DiT [38].
In our work, the context $c$ not only includes the past visual context $\mathbf{x}$ but also the future motion intent expressed by the human pose sequence $\mathbf{P}$. We discuss how we encode the human pose sequences $\mathbf{P}$ in Sec. 3.1 and describe how we effectively control the diffusion model in the challenging setting of egocentric video generation in Sec. 3.2.

## 3.1. Pose Representation

Egocentric datasets typically provide frame-synchronized 3D body poses defined in a global reference frame. However, since EgoControl is designed to operate from an embodied perspective, we require a pose representation that reflects the agent's own motion rather than its absolute position in space. To this end, we represent each pose sequence $\mathbf{P}$ in terms of relative transformations that capture frame-to-frame motion dynamics.

Each 3D pose is given as a set of $J = 23$ transformation matrices describing the position and orientation of each joint with respect to the global frame. From an egocentric viewpoint, most perceptual changes in the video arise from the movement of the camera, which corresponds to the head motion. Therefore, given the sequence of global head poses $\mathbf{H} = (\mathbf{H}_0, \mathbf{H}_1, \ldots, \mathbf{H}_M)$, where $\mathbf{H}_i \in \mathbb{R}^{4\times4}$, we compute the relative head transformation between consecutive frames as

$$\Delta\mathbf{H}_i = \mathbf{H}_i \mathbf{H}_{i-1}^{-1}, \quad \forall i \in [1, M]. \tag{3}$$

Each $\Delta\mathbf{H}_i$ is then converted into a 6D vector $\Delta\mathbf{h}_i \in \mathbb{R}^{1\times6}$ representing translation and rotation in Euler format, yielding the sequence of head movements $\mathbf{\Delta h} = (\Delta\mathbf{h}_1, \ldots, \Delta\mathbf{h}_M) \in \mathbb{R}^{M\times1\times6}$.

To incorporate full-body state, we also compute the relative transformations of all joints with respect to the pelvis at each frame, denoted as $\mathbf{J} \in \mathbb{R}^{M\times21\times6}$, where 21 is the number of joints excluding head and pelvis. Furthermore, we model the pelvis as the root joint and compute its frame-to-frame relative motion $\mathbf{\Delta r} = (\Delta\mathbf{r}_1, \ldots, \Delta\mathbf{r}_M) \in \mathbb{R}^{M\times1\times6}$ as we did for the head.

Finally, we concatenate these components to obtain the full pose representation that will control the generation of the $M$ future frames:

$$\mathbf{P} = [\mathbf{\Delta h}, \mathbf{\Delta r}, \mathbf{J}] \in \mathbb{R}^{M\times23\times6}. \tag{4}$$

This representation allows EgoControl to capture both egocentric camera dynamics and articulated body motion in a unified form.

## 3.2. Control Mechanism

We propose to use two control mechanisms to condition the video generation on the full human body pose representation $\mathbf{P}$ as illustrated in Fig. 2. The first approach via modulation is described in Sec. 3.2.1 and the second approach via cross-attention is described in Sec. 3.2.2. We show that the combination of both gives the best results.

### 3.2.1. Pose control via modulation

Inspired by [1, 23, 36], we use a modulation and gating approach to update the features in the DiT [38] blocks, where

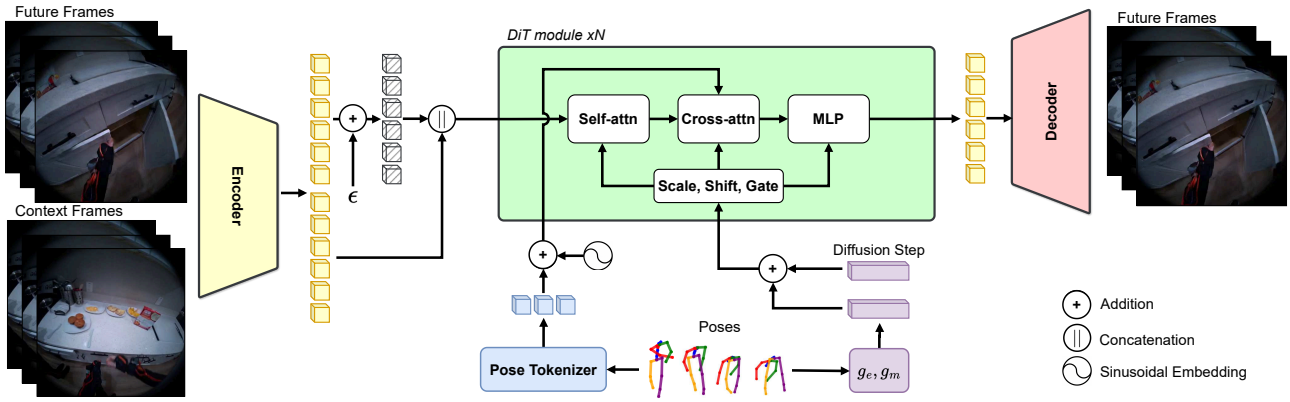



Figure 2. EgoControl generates future egocentric frames conditioned on the past frames and sequence of human poses. We condition the model on the human poses in two ways. The purple branch uses the networks $g_e$ and $g_m$ to obtain modulation vectors that are combined with the ones obtained from the diffusion step. These vectors then impact the modulation by scaling, shifting, and gating the self-attention, cross-attention, and MLP blocks of DiT. The blue branch tokenizes the poses and computes the cross-attention with the encoded visual tokens. Both branches are necessary to generate realistic egocentric video frames that are well-aligned with the human pose sequence.

the modulation and gating parameters depend on the full-body poses. To this end, we first flatten the entire pose tensor $\mathbf{P} \in \mathbb{R}^{M\times 23\times 6}$ across time and joints, yielding a single global vector $\mathbf{P}_{\text{flat}} \in \mathbb{R}^{(M\cdot 23\cdot 6)}$. A pair of two-layer MLPs $g_e$ and $g_m$ then map this vector to the embedding and modulation space:

$$\mathbf{e}_P = g_e(\mathbf{P}_{\text{flat}}), \qquad \mathbf{m}_P^{\beta\gamma g} = [\mathbf{m}_P^{\beta}, \mathbf{m}_P^{\gamma}, \mathbf{m}_P^{g}] = g_m(\mathbf{P}_{\text{flat}}). \tag{5}$$

Given the pose embedding $\mathbf{e}_P$ and its modulation vector $\mathbf{m}_P^{\beta\gamma g}$, different shift, scale and gate parameters are predicted for the self-attention, cross-attention, and MLP components in each block:

$$[\boldsymbol{\beta}_P^{(k)}, \boldsymbol{\gamma}_P^{(k)}, \mathbf{g}_P^{(k)}] = \mathbf{W}_{m1}^{k}\mathbf{W}_{m2}^{k}\text{SiLU}(\mathbf{e}_P) + \mathbf{m}_P^{\beta\gamma g}, \tag{6}$$

where $k$ denotes self-attention, cross-attention, or MLP. $\mathbf{W}_{m1}^{k} \in \mathbb{R}^{3\cdot d\times r}$ and $\mathbf{W}_{m2}^{k} \in \mathbb{R}^{r\times d}$ are learnable projection matrices with $r < d$. Each component input $\mathbf{u}$ is then normalized and modulated via adaptive layer normalization:

$$\text{AdaLN}^{(k)}(\mathbf{u}; \boldsymbol{\beta}_P^{(k)}, \boldsymbol{\gamma}_P^{(k)}) = \text{LN}(\mathbf{u}) \odot (1 + \boldsymbol{\gamma}_P^{(k)}) + \boldsymbol{\beta}_P^{(k)}, \tag{7}$$

where LN denotes layer normalization. The resulting outputs of the self-attention, cross-attention, and MLP components are gated and combined through residual connections:

$$\mathbf{u} \leftarrow \mathbf{u} + \mathbf{g}_P^{(\text{self})} \odot \text{SelfAttn}\Big(\text{AdaLN}^{(\text{self})}(\mathbf{u})\Big), \tag{8}$$

$$\mathbf{u} \leftarrow \mathbf{u} + \mathbf{g}_P^{(\text{cross})} \odot \text{CrossAttn}\Big(\text{AdaLN}^{(\text{cross})}(\mathbf{u}), \mathbf{c}\Big), \tag{9}$$

$$\mathbf{u} \leftarrow \mathbf{u} + \mathbf{g}_P^{(\text{mlp})} \odot \text{MLP}\Big(\text{AdaLN}^{(\text{mlp})}(\mathbf{u})\Big), \tag{10}$$

where $\mathbf{c}$ is the context for the cross-attention, which we will describe in Sec. 3.2.2. This mechanism allows $\mathbf{P}$ to modulate normalization and residual pathways across all layers.

As illustrated in Fig. 2, this mechanism is also used to integrate the diffusion step $t$. To this end, $t$ is mapped to a continuous embedding space using a sinusoidal embedding function

$$\mathbf{e}_t = \text{sinusoidal}(t) \in \mathbb{R}^{D}, \tag{11}$$

and we compute the modulation parameters for the diffusion step in the same way, i.e., $\mathbf{m}_t^{\beta\gamma g} = f(\mathbf{e}_t)$ where $f$ is a two-layer MLP. We then combine it with Eq. (5) by

$$\mathbf{e}_P = \mathbf{e}_t + \mathbf{e}_P, \qquad \mathbf{m}_P^{\beta\gamma g} = \mathbf{m}_t^{\beta\gamma g} + \mathbf{m}_P^{\beta\gamma g}. \tag{12}$$

This additive conditioning enables both diffusion step and body pose to jointly modulate the layer normalization and residual dynamics throughout the transformer blocks.

#### 3.2.2. Pose control via cross-attention

In addition to the AdaLN-based modulation, we introduce a fine-grained control pathway by further conditioning the transformer blocks through cross-attention with body pose tokens as illustrated in Fig. 2. Unlike the AdaLN branch, here we preserve the temporal structure of $\mathbf{P}$ to maintain frame-level alignment. Each frame's pose vector $\mathbf{P}_m \in \mathbb{R}^{23\times 6}$ is projected to the model's feature space as:

$$\mathbf{P}_m^{'} = \text{LayerNorm}(\text{GELU}(\mathbf{W}_p\mathbf{P}_m)) \in \mathbb{R}^{D_p}, \tag{13}$$

yielding a sequence of $M$ pose tokens. We then apply a sinusoidal positional encoding $\text{PE}(\cdot)$ to preserve the temporal information:

$$\mathbf{c}_m = \text{PE}(\mathbf{P}_m^{'}). \tag{14}$$

These pose tokens are then concatenated and used as context $\mathbf{c}$ for the cross-attention in Eq. (9). This cross-attention pathway provides temporally localized control signals that complement the global modulation from AdaLN, together enabling coherent and controllable egocentric video synthesis.

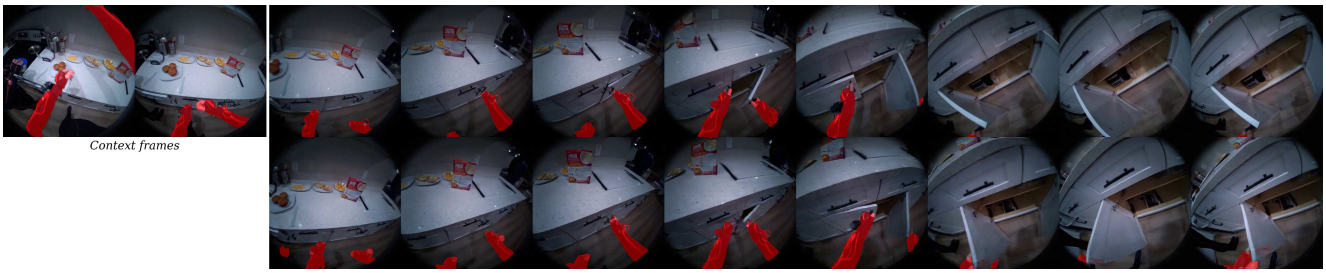



Figure 3. SAM2 [41] is used to segment and track the visible arms in both ground truth (first row) and the generated frames (second row). The extracted segmentation masks (highlighted in **red**) are then used to assess the quality of body pose control with mIoU and Acc%.

# 4. Experiments

We use Nymeria [31] for our experiments, which is the largest existing dataset of human motion captured 'in the wild'. It features diverse individuals performing everyday activities across a wide range of real-world environments. Nymeria contains over 1,100 high-resolution egocentric video sequences ($1408 \times 1408$ at 30 FPS), each temporally synchronized with 3D body poses obtained from a full-body XSens motion capture system [33]. We resample the videos to 16 frames per second and resize them to $480 \times 480$ resolution. From the full dataset, we selected a subset of 186 videos (approximately 50 hours in total) containing multiple participants and complex interactions. These videos were further divided into non-overlapping clips of 45 frames.

## 4.1. Evaluation Protocol

A comprehensive evaluation protocol is essential to assess whether an egocentric, pose-controllable video generation model can function as a reliable simulator of an embodied agent. To this end, we design an evaluation framework along three complementary dimensions that jointly capture visual quality, motion accuracy, and body control.

**Visual quality and fidelity.** We evaluate the overall realism and perceptual quality of the generated videos using standard metrics for video generation. Specifically, we report Fréchet Video Distance (FVD) [49] for sequence-level evaluation, and the frame-level metrics SSIM, LPIPS, and DreamSim [10] to quantify perceptual similarity to the ground truth. We scale SSIM, LPIPS, and DreamSim by a factor of 100 for easier comparison. These measures also provide an implicit indication of temporal consistency and adherence to pose guidance.

**Global motion control.** From an egocentric perspective, most visual changes correspond to the agent's motion through the environment, which is effectively represented by the camera's movement. Following prior work on camera-controllable video generation [18], we evaluate global motion control by measuring the average translation error (`TransError`) and rotation error (`RotError`). We used Vipe [22] to extract camera poses from both the ground-truth and generated videos. The full sequences, comprising both conditioning and predicted frames, were provided to Vipe, ensuring consistent scale alignment between the two pose trajectories.

**Agent body control.** While the previous metrics implicitly capture alignment between the generated video and the conditioning poses, they do not explicitly quantify the accuracy of full-body motion reproduction. To address this, we focus on evaluating visible body parts, primarily the arms, within the agent's field of view. We employ SAM2 [41] for video segmentation and tracking, applied to a random subset of sequences where the arms are visible. Keypoints were manually annotated in the first frame, and SAM2 was used to propagate the corresponding segmentation masks throughout the sequence as shown in Fig. 3. We then compute the mean Intersection-over-Union (IoU) between the ground-truth and generated arm masks, as well as the percentage of frames correctly detecting arm presence as a binary measure of body-control consistency.

## 4.2. Implementation Details

We used Cosmos [1] as latent conditional video diffusion model and baseline. Cosmos is pretrained at scale on large video corpora (including $\approx 8\%$ first-person view videos), making it a suitable backbone for our conditional egocentric video generation. We ran all our experiments using the `cosmos-predict2` (2B, 480p, 16 FPS) pre-trained model. We trained the model on videos of 45 frames, using 13 past frames as conditioning and generate 32 future frames (2 seconds) controlled by 3D full-body poses as described in the previous section. This model operates in a spatio-temporal compressed latent space using the Cosmos-Tokenizer, reducing the input to 12 latent frames. More details are in the supp. material and available at `cvg-bonn.github.io/EgoControl`.

| Experiments | Frame-level fidelity | | | Video quality | Motion control | | Body control | |
|---|---|---|---|---|---|---|---|---|
| | SSIM↑ | LPIPS↓ | DreamSim↓ | FVD↓ | TransError | RotError | mIoU | Acc% |
| Base Cosmos | 42.29 | 50.62 | 23.00 | 71.00 | 16.53 | 15.60 | 20.03 | 85.36 |
| Finetuned | 47.47 | 45.74 | 18.14 | 40.70 | 9.93 | 13.65 | 25.13 | 85.20 |
| Head control | 56.94 | 29.71 | 10.22 | 22.68 | 5.16 | 3.29 | 33.70 | 91.14 |
| Body control | **58.60** | **26.71** | **8.54** | **20.18** | **4.90** | **2.96** | **52.13** | **96.33** |

Table 1. EgoControl (row 4) shows notable improvements across all dimensions over the baseline and the head only control.

### 4.3. Results

**Quantitative results.** Tab. 1 summarizes our main quantitative findings and performances of EgoControl. We report results for the pre-trained Cosmos backbone (row 1), then we establish a baseline by fine-tuning Cosmos using only past frames for context, with no pose guidance (row 2). Starting from this video prediction baseline, we introduce two variants: one conditioned solely on head motion ($\mathbf{\Delta h}$) and our final model (rows 3-4), which incorporates the complete full-body pose representation ($\mathbf{P}$) as defined in Sec. 3.1. Results show that, across all standard image- and video-level metrics, fine-tuning the pretrained backbone on Nymeria yields substantial gains over the off-the-shelf model. Using our proposed approach conditioned solely on head motion (row 3) substantially improves all metrics including body control further. The best results, however, are achieved when we control the video generation by the 3D full-body pose (row 4). The mIoU metric, which measures the alignment of the generated arms with the controlling arm poses, increases by nearly 55% relative to the head-only control. In over 96% of the cases, the visibility of the arms in the generated frames is consistent with the ground-truth. It is very interesting to note that the full-body pose also improves the consistency of the camera view with the head pose compared to the head-only control. These results demonstrate that full-body information is critical to have full control over egocentric videos. Furthermore, the visual quality metrics (SSIM, LPIPS, DreamSim, FVD) are also improved by our model using full-body control.

In Tab. 2, we provide a comparison with the results reported by PEVA [3]. Because PEVA generates instead of a video a single frame 2 s in the future at $224 \times 224$ resolution given 15 past frames, we take the 32nd generated frame of EgoControl and downsample it to match this horizon and image resolution. Under this protocol, EgoControl outperforms PEVA despite using a shorter context window. In Tab. 3, we also compare EgoControl with full-body control to CAMI2V [63], a camera control model finetuned on Nymeria.

**Qualitative results.** Qualitatively, EgoControl produces temporally coherent egocentric sequences that closely follow the controlling full-body pose. Fig. 5 shows how different pose sequences, representing distinct actions like walking, rotating or moving an arm, can be applied to the same context frames, resulting in generated futures that follow the intended motions while adapting to the scene and objects. Fig. 6 further demonstrates the control abilities of EgoControl by applying the same pose sequence to different initial contexts. In this case, the original pose of the video in the first row, representing the action of bending down to pick an object (a badminton ball), is applied to two different context frames in addition to the original one. In all three cases, the agent clearly reproduces the motion by bending down. We provide more qualitative examples including object interactions in the supp. material and at cvg-bonn.github.io/EgoControl. This also includes results at higher resolution ($960 \times 960$) and on Ego4D [15].

| Model | LPIPS ↓ | DreamSim ↓ | FID ↓ |
|---|---|---|---|
| PEVA XXL [3] | 29.8 | 18.6 | 61.10 |
| EgoControl | **24.3** | **11.3** | **50.68** |

Table 2. Comparison with PEVA XXL. Note that PEVA uses 15 past frames at 4 FPS ($224 \times 224$) as conditioning and generates a single frame in the future. The comparison is for a generated frame 2 seconds in the future.

| Experiments | SSIM | TransError | RotError | mIoU |
|---|---|---|---|---|
| CAMI2V [63] | 43.28 | 4.60 | 3.79 | 24.72 |
| EgoControl | **58.87** | **2.22** | **1.33** | **56.85** |

Table 3. Comparison with the camera control method CAMI2V. Since CAMI2V generates fewer frames, we evaluate only the first 16 frames.

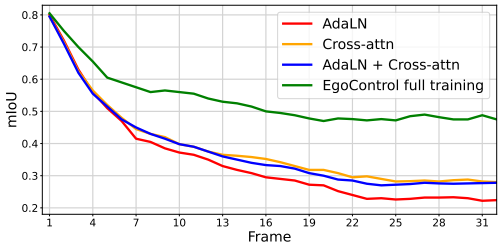
### 4.4. Ablation Studies

We conducted a series of ablation studies to validate our core design choices. We systematically investigated the efficacy of our proposed control mechanism and pose representation.

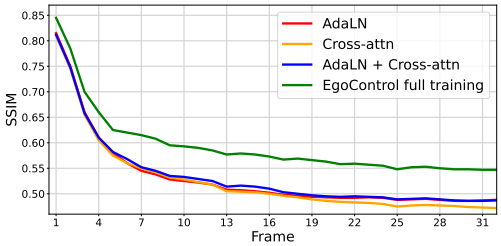
**Control method.** To justify our hybrid design, which integrates pose control via both modulation (AdaLN) and

| Control type | SSIM | LPIPS | DreamSim | FVD | mIoU | `TransError` | `RotError` |
|---|---|---|---|---|---|---|---|
| AdaLN | 52.16 | 38.38 | 11.48 | 29.14 | 33.44 | 6.07 | 5.99 |
| Cross-attn (CA) | 51.65 | 38.89 | 12.20 | 29.28 | **37.84** | 6.85 | 7.19 |
| AdaLN + CA | **52.60** | **36.79** | **10.94** | **27.51** | 37.40 | **5.59** | **5.23** |

Table 4. Ablation on the control mechanism.




(a) Average mIoU per frame.




(b) Average SSIM per frame.

Figure 4. Pose alignment (mIoU) and SSIM for each of the 32 generated frames. The green line denotes the fully trained model.

cross-attention, we evaluated three distinct variants in Tab. 4. For the ablation studies, we trained the models with 7.000 iterations, which is less than in Tab. 1. When we compare conditioning with only AdaLN (row 1) and only cross-attention (row 2), we observe that AdaLN captures appearance and head motion better but struggles to enforce fine-grained articulated body control, i.e., it is better for the visual quality and global motion metrics (`TransError`, `RotError`) but it is worse in mIoU. The advantage of cross-attention for body alignment is expected since it utilizes the information of the controlling human pose sequence in a more granular way than AdaLN. The combination of both methods yields superior performance in both visual quality and motion accuracy, while being similar to using only cross-attention for body control (mIoU). Figs. 4a and 4b furthermore plot mIoU and SSIM for each generated frame separately, confirming the advantage of combining both control mechanisms. The plots also show the results for the fully trained model.

**Pose representation.** The manner in which motion is represented is critical for effective conditioning. We explored several alternatives to our proposed formulation. First, for head-only control, we compared our differential representation (encoding relative transformations between consecutive frames) against a cumulative representation (encoding transformations relative to the initial pose $\mathbf{H}_0$). As shown in Tab. 5, the `TransError` of the cumulative representation is nearly twice as high compared to our differential strategy. While conceptually simpler, the cumulative approach suffers from numerical instability as motion values grow over long sequences. The differential formulation provides a more stable and temporally accurate signal, leading to superior performance in both visual (e.g. SSIM and FVD) and motion metrics. Second, for full-body control, we validated our choice of representing joint locations relative to the pelvis. We compared this approach with an alternative that encodes the per-frame relative motion of each joint ($\mathbf{\Delta j}$), similar to our representation of head motion. The results, also in Tab. 5, show that this frame-to-frame differential encoding for joints is suboptimal. The model struggles to maintain body coherence, as evidenced by a significant drop in pose control metrics. An improvement of 5.55 in mIoU confirms that a pelvis-centric coordinate system provides a more robust and easier-to-learn representation for articulated body dynamics.

| | Head motion | | | | | |
|---|---|---|---|---|---|---|
| Head only variants | SSIM | LPIPS | DreamSim | FVD | `TransError` | `RotError` |
| Cumulative | 54.51 | **32.69** | 11.08 | 27.23 | 12.31 | **4.11** |
| Differential | **55.16** | 32.99 | **10.86** | **24.01** | **6.75** | 4.67 |
| | Full-body pose representation | | | | | |
| Fullbody variants | SSIM | LPIPS | DreamSim | FVD | mIoU | Acc% |
| $\mathbf{P} = [\mathbf{\Delta j}] \times 23$ | **52.90** | **36.07** | 11.26 | 30.12 | 31.85 | 91.17 |
| $\mathbf{P} = [\mathbf{\Delta h}, \mathbf{\Delta r}, \mathbf{J}]$ | 52.60 | 36.79 | **10.94** | **27.51** | **37.40** | **93.03** |

Table 5. Ablation on the pose representation.

## 5. Conclusion

We presented EgoControl, a diffusion-based egocentric video generator that conditions future-frame synthesis on 3D full-body pose sequences of the camera wearer. By introducing a compact pose representation and a dedicated control pathway, EgoControl translates global camera dynamics and articulated body motion into temporally coherent, pose-consistent egocentric videos, improving substantially both visual fidelity and motion alignment compared to baselines. At the same time, our study reveals clear limitations that point to promising future work. First, Nymeria does not provide explicit hand-pose annotations, so our current model does not explicitly model fine-grained hand articulation. While full-body conditioning substantially improves arm and coarse-hand motion, this gap limits accurate hand control. Second, our training data is biased toward a specific acquisition setup with fisheye-mounted cameras and subjects wearing a motion-capture suit, which may reduce out-of-distribution generalization to everyday viewpoints, clothing, and sensor rigs. Future work should therefore extend pose conditioning to include explicit hand rep-

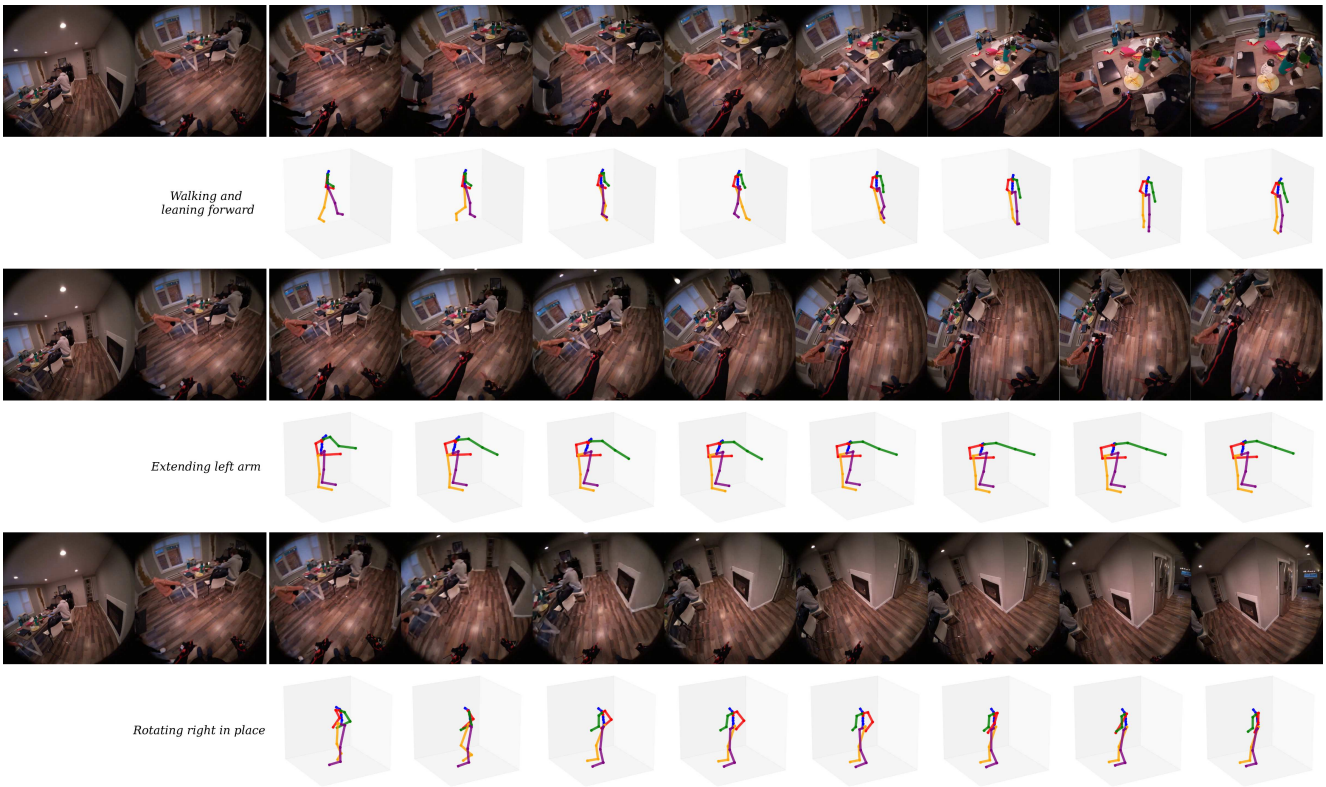



Figure 5. Given the same initial context, EgoControl is capable of generating videos following different body movements. See videos at cvg-bonn.github.io/EgoControl for better visualization.

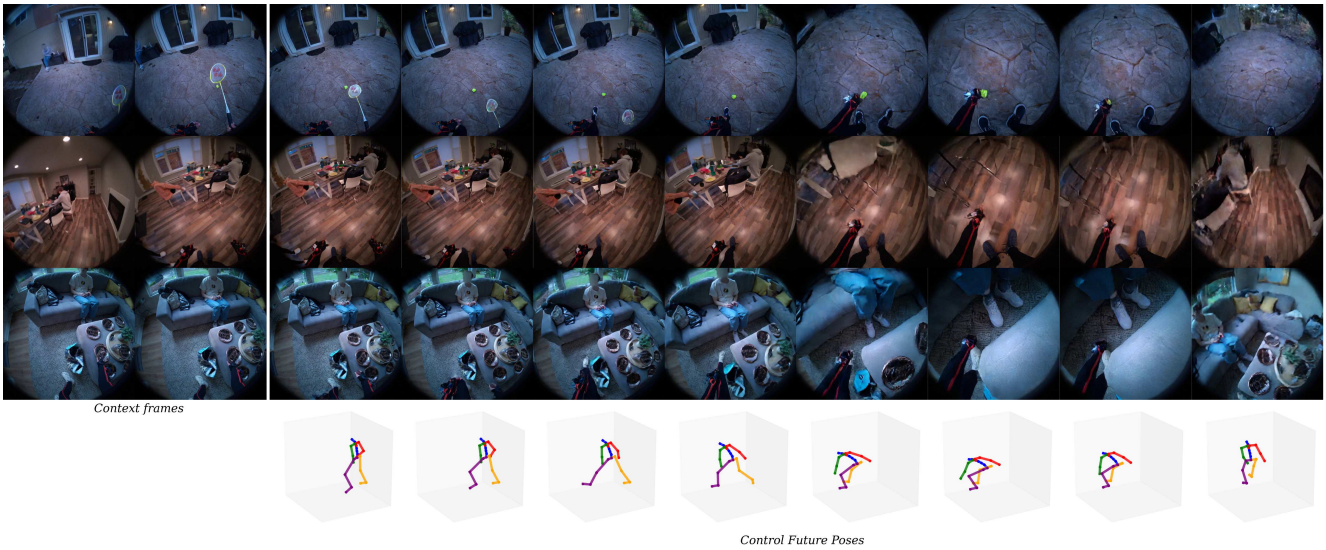



Figure 6. EgoControl shows accurate pose alignment across three scenarios through different context frames. See videos at cvg-bonn.github.io/EgoControl for better visualization.

resentations and utilize additional training data to broaden camera and clothing variability. We believe these directions will further close the gap between controllable egocentric synthesis and realistic embodied simulation for downstream planning and interaction tasks.

## Acknowledgements

This work was supported by the ERC Consolidator Grant FORHUE (101044724). We gratefully acknowledge the Federal Ministry of Research, Technology and Space, the Ministry of Culture and Science of the State of North Rhine-Westphalia, the Ministry of Science, Research and Arts of the State of Baden-Württemberg, the Bavarian State Ministry of Science and the Arts and the Gauss Centre for Supercomputing e.V. (GCS) for funding this project by providing computing time on the Supercomputer JUPITER at Jülich Supercomputing Centre (JSC) of Forschungszentrum Jülich through the Gauss AI Compute Competition. We also thank the German AI Service Center WestAI (01IS22094A) for providing additional computational resources.

## References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2, 3, 5

[2] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12467–12479, 2025. 2

[3] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. In *Annual Conference on Neural Information Processing Systems*, 2025. 3, 6

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15791–15801, 2025. 2

[5] 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. 2

[6] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *International Conference on Machine Learning*, pages 6263–6285. PMLR, 2024. 2

[7] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 2

[9] Luis Denninger, Sina Mokhtarzadeh Azar, and Juergen Gall. CamC2V: Context-aware controllable video generation. In *International Conference on 3D Vision*, 2026. 2

[10] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36:50742–50768, 2023. 5

[11] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 2

[12] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024. 2

[13] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3170–3180, 2022. 2

[14] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 1, 2

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 1, 2, 6

[16] 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. In *International Conference on Learning Representations*, 2024. 2

[17] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22404–22415, 2025. 2

[18] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 5

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[21] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18219–18228, 2022. 2

[22] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025. 5

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 3

[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 3

[25] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5492–5501, 2019. 1, 2

[26] Gen Li, Yutong Chen, Yiqian Wu, Kaifeng Zhao, Marc Pollefeys, and Siyu Tang. Egom2p: Egocentric multimodal multitask pretraining. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10830–10843, 2025. 2

[27] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6943–6953, 2021. 2

[28] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 1, 2

[29] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *ACM International Conference on Multimedia*, pages 974–982, 2021. 2

[30] Hongchen Luo, Kai Zhu, Wei Zhai, and Yang Cao. Intention-driven ego-to-exo video generation. *arXiv preprint arXiv:2403.09194*, 2024. 2

[31] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision (ECCV)*, pages 445–465. Springer, 2024. 1, 2, 5

[32] 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 *AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[33] Movella. *MVN User Manual*. Movella, 2021. https://www.movella.com/hubfs/MVN_User_Manual.pdf. 5

[34] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18444–18455, 2023. 2

[35] Enrico Pallotta, Sina Mokhtarzadeh Azar, Shuai Li, Olga Zatsarynna, and Juergen Gall. Syncvp: Joint diffusion for synchronous multi-modal video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13787–13797, 2025. 2

[36] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2332–2341, 2019. 3

[37] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10318–10329, 2025. 2

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 2, 3

[39] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *Int. J. Comput. Vision*, 132(11): 4880–4936, 2024. 1

[40] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*. 5

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[43] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. 2

[44] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH*, pages 1–11, 2024. 2

[45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual,

Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023. 2

[46] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *European Conference on Computer Vision (ECCV)*, pages 414–430. Springer, 2024. 2

[47] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *Advances in Neural Information Processing Systems*, 36:33485–33498, 2023. 2

[48] Anirudh Thatipelli, Shao-Yuan Lo, and Amit K. Roy-Chowdhury. Egocentric and exocentric methods: A short survey. *Computer Vision and Image Understanding*, 257, 2025. 1

[49] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *arXiv preprint*, 2019. 5

[50] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022. 2

[51] 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

[52] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 1

[53] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9336, 2024. 2

[54] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[55] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, pages 1–11, 2024. 2

[56] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2024. 2

[57] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision (ECCV)*, pages 399–417. Springer, 2024. 2

[58] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. X-gen: Ego-centric video prediction by watching exo-centric videos. In *International Conference on Learning Representations*, 2025. 2

[59] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH*, pages 1–12, 2024. 2

[60] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. 2

[61] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Multi-modal temporal convolutional network for anticipating actions in egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2249–2258, 2021. 2

[62] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, pages 74896–74910. PMLR, 2025. 2

[63] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 2, 6