

EgoVid-5M: A Large-Scale Video-Action Dataset for Egocentric Video Generation

Xiaofeng Wang^{1,2,4} Kang Zhao¹ Feng Liu⁴ Jiayu Wang¹ Guosheng Zhao^{2,4}
 Xiaoyi Bao^{2,4} Zheng Zhu³ Yingya Zhang¹ Xingang Wang²
¹Alibaba ²CASIA ³Tsinghua University ⁴UCAS

Project Page: <https://egovid.github.io>



Figure 1. *EgoVid-5M* is a meticulously curated high-quality action-video dataset designed specifically for egocentric video generation. It includes detailed action annotations, such as fine-grained kinematic control and high-level textual descriptions. Furthermore, it incorporates robust data cleaning strategies to ensure frame consistency, action coherence, and motion smoothness under egocentric conditions.

Abstract

Video generation has emerged as a promising tool for world simulation, leveraging visual data to replicate real-world environments. Within this context, egocentric video generation, which centers on the human perspective, holds significant potential for enhancing applications in virtual reality, augmented reality, and gaming. However, the generation of egocentric videos presents substantial challenges due to the dynamic nature of egocentric viewpoints, the intricate diversity of actions, and the complex variety of scenes encountered. Existing datasets are inadequate for addressing these challenges effectively. To bridge this gap, we present *EgoVid-5M*, the first high-quality dataset specifically curated for egocentric video genera-

tion. *EgoVid-5M* encompasses 5 million egocentric video clips and is enriched with detailed action annotations, including fine-grained kinematic control and high-level textual descriptions. To ensure the integrity and usability of the dataset, we implement a sophisticated data cleaning pipeline designed to maintain frame consistency, action coherence, and motion smoothness under egocentric conditions. Furthermore, we introduce *EgoDreamer*, which is capable of generating egocentric videos driven simultaneously by action descriptions and kinematic control signals. The *EgoVid-5M* dataset, associated action annotations, and all data cleansing metadata will be released for the advancement of research in egocentric video generation.

Dataset	Year	Domain	Gen.	Text	Kinematic	CM.	#Videos	#Frames	Res
HowTo100M [43]	2019	Open	✓	ASR	✗	✗	136M	~ 90	240p
WebVid-10M [2]	2021	Open	✓	Alt-Text	✗	✗	10M	~ 430	Diverse
HD-VILA-100M [68]	2022	Open	✓	ASR	✗	✗	103M	~ 320	720p
Panda-70M [8]	2024	Open	✓	Auto	✗	✗	70M	~ 200	Diverse
OpenVid-1M [44]	2024	Open	✓	Auto	✗	✗	1M	~ 200	Diverse
VIDGEN-1M [55]	2024	Open	✓	Auto	✗	✗	1M	~ 250	720p
LSMDC [50]	2015	Movie	✗	Human	✗	✗	118K	~ 120	1080p
UCF101 [53]	2015	Action	✗	Human	✗	✗	13K	~ 170	240p
Ego4D [16]	2022	Egocentric	✗	Human	IMU	✗	931	~ 417K	1080p
Ego-Exo4D [17]	2024	Egocentric	✗	Human	MVS	✗	740	~ 186K	1080p
EgoVid-5M (ours)	2024	Egocentric	✓	Auto	VIO	✓	5M	~ 120	1080p

Table 1. Comparison of *EgoVid-5M* and other video datasets, where *Gen.* denotes whether the dataset is designed for generative training, *CM.* denotes cleansing metadata, *#Videos* is the number of videos, and *#Frames* is the average number of frames in a video.

1. Introduction

One of the most promising avenues in video generation is the development of world simulators. These systems utilize visual simulations and interactions to deliver applications in the physical world. Contemporary research is increasingly validating the capabilities of video generation in this realm, including applications in autonomous driving [25, 58, 60, 71, 78, 79], autonomous agents [5, 10, 18, 19, 64, 72, 82], and even in general world [4, 13]. In the context of human-centric scenarios, leveraging behavioral actions to drive egocentric video generation has emerged as a pivotal strategy. This approach greatly enhances applications in Virtual Reality (VR), Augmented Reality (AR), and gaming, offering more immersive and interactive experiences and advancing the state of the art in these fields.

Video generation necessitates vast quantities of high-quality data for training. This requirement is even more stringent in egocentric video generation, which is inherently challenging due to the dynamic nature of egocentric perspectives, the richness of observed actions, and the diversity of encountered scenarios. Despite the critical need for specialized data, there is currently a scarcity of publicly available, large-scale datasets for training egocentric video generation models. To bridge this gap, we present the *EgoVid-5M* dataset, a pioneering high-quality dataset specifically curated for egocentric video generation (see Fig. 1). As shown in Tab. 1, *EgoVid-5M* is distinguished by several key features: (1) **High Quality**: This dataset offers 5 million egocentric videos at 1080p resolution. In contrast to Ego4D [16], which is intended for egocentric perception and includes excessive noisy camera motion that is unsuitable for generative training, *EgoVid-5M* undergoes a rigorous data cleaning process. The videos are curated based on stringent criteria, including the alignment between ac-

tion descriptions and video content, the magnitude of motion, and the consistency between frames. (2) **Comprehensive Scene Coverage**: *EgoVid-5M* boasts a comprehensive range of scenarios including household environments, outdoor settings, office activities, sports, and skilled operations. It encompasses hundreds of action categories, thus covering the majority of scenes encountered in egocentric perspectives. (3) **Detailed and Precise Annotations**: The dataset includes extensive behavioral annotations, which are categorized into fine-grained kinematic control and high-level action descriptions. For kinematic information, we employ Visual Inertial Odometry (VIO) to provide precise annotations, ensuring accurate alignment with video contents. For action descriptions, a multimodal large language model combined with a large language model is utilized to generate detailed text annotations.

Leveraging the proposed *EgoVid-5M* dataset, we train different video generation baselines to validate the dataset’s quality and efficacy. Various architectures, such as U-Net [3, 65] and DiT [30], are employed as baseline models, and the experimental results demonstrate that *EgoVid-5M* significantly bolsters the training of egocentric video generation. In addition, we propose *EgoDreamer*, which utilizes both action descriptions and kinematic control to drive the generation of egocentric videos. To provide a comprehensive assessment of action-driven egocentric video generation, we establish an extensive set of evaluation metrics. These metrics encompass multiple dimensions, including visual quality, frame coherence, semantic compliance with actions, and kinematic accuracy. Extensive experiments show that *EgoVid-5M* markedly enhances the capability of various video generation models to produce high-quality egocentric videos.

The main contributions of this paper can be summarized as follows: (1) We introduce *EgoVid-5M*, the first publicly

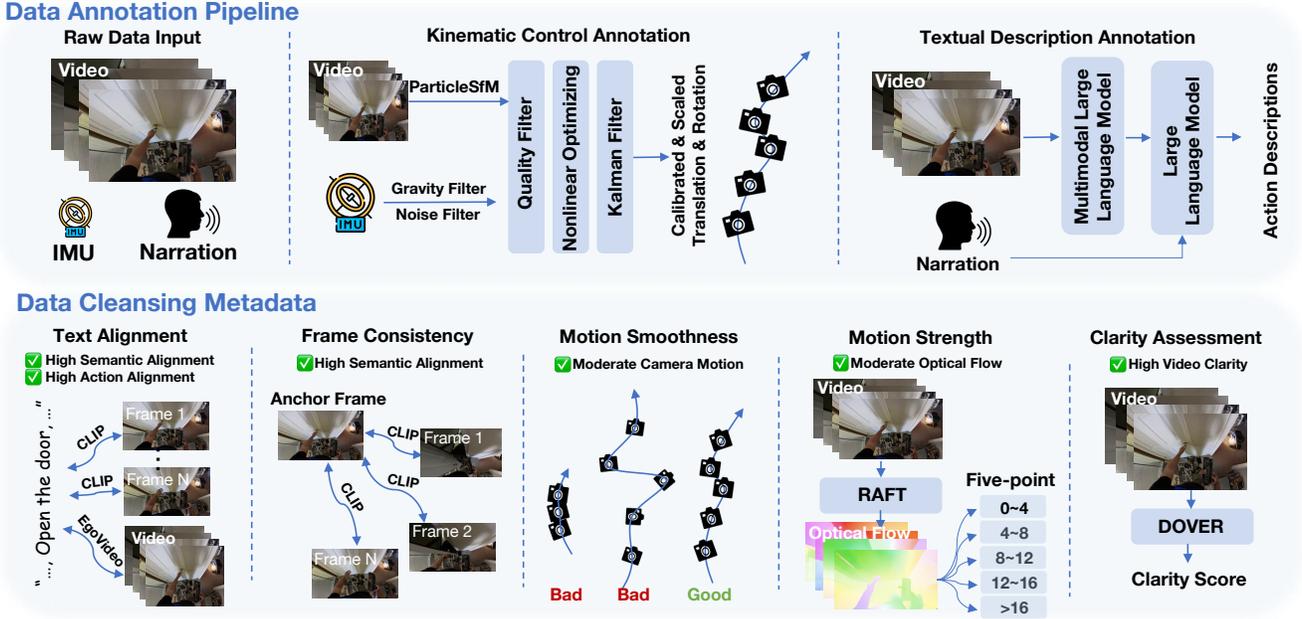


Figure 2. Data annotation pipeline and cleansing metadata of *EgoVid-5M*.

released, high-quality dataset tailored for egocentric video generation. This dataset is proposed to advance both research and applications in the domain of egocentric visual simulation. (2) Our dataset includes detailed and precise action annotations, incorporating both fine-grained kinematic control and high-level textual descriptions. In addition, we employ robust data cleaning strategies to ensure frame consistency, action coherence, and motion smoothness within *EgoVid-5M*. (3) Utilizing *EgoVid-5M*, we conducted extensive experiments on various video generation baselines to validate the dataset’s quality and efficacy. Furthermore, to support future advancements in action-driven egocentric video generation, we propose *EgoDreamer*, which leverages both action descriptions and kinematic control to drive egocentric video generation.

2. Related Work

2.1. Video Generation as World Simulators

Video generation technology has seen rapid advancements recently. Both diffusion-based [3, 6, 7, 21, 30, 41, 65, 81] and token-based [23, 54, 59, 69, 74, 75] video generation models have proven that the quality and controllability of video generation are steadily improving [83]. Notably, the introduction of the Sora model [4] attracts significant attention which convincingly shows that current video generation models are capable of understanding and adhering to physical laws, thereby substantiating the potential of these models to function as world simulators. This perspective is echoed by Runway, which posits that their

Gen-3 Alpha [14] is progressing along this promising trajectory. Additionally, video generation models, employed as simulators, have demonstrated significant utility in various real-world applications, including autonomous driving simulations [25, 58, 60, 71, 78, 79] and agent-based environments [5, 10, 18, 19, 64, 72, 82]. Within this context, action-driven egocentric video generation, which centers on the human perspective, holds significant potential for enhancing applications in VR, AR, and gaming. However, current research in the egocentric domain predominantly concentrates on understanding tasks [1, 15, 34, 37, 42, 45, 47, 48, 67], and generative tasks associated with egocentric scenarios are largely confined to exocentric-to-egocentric video synthesis [32, 36, 39]. This highlights a substantial gap in generating action-driven egocentric videos. While some methods have explored video generation driven by action interaction [20, 24, 26, 27, 40, 62, 66, 73], these approaches are mainly concerned with natural scenes featuring smooth camera transitions. This focus limits their ability to model intricate motion patterns inherent in egocentric videos.

2.2. Video Generation Datasets

In the realm of video generation, the quantity and quality of training data are pivotal for training effective models. Currently, the field of general video generation benefits from several pioneering open-source video datasets. WebVid-10M [2] consists of 52K hours of video, totaling 10.7M text-video pairs. Similarly, InternVid [61] contains over 7M videos spanning nearly 760K hours, resulting in 234M video clips and a comprehensive dataset with 4.1B words in descriptive texts. Panda70M [8] stands out with

its collection of 70M high-resolution and semantically coherent video samples. OpenVid-1M [44], offers a million-level, high-quality dataset encompassing diverse scenarios such as portraits, landscapes, cities, metamorphic elements, and animals. In contrast to these general-purpose datasets, specific-scenario datasets typically comprise a limited number of text-video pairs tailored for particular contexts. UCF-101 [53] is an action recognition dataset featuring 101 classes and 13,320 total videos. Taichi-HD [51], a more focused collection, includes 2,668 videos capturing a single person performing Taichi movements. In the domain of egocentric video generation, existing datasets such as Ego4D [16] and Ego-Exo4D [17] are primarily designed for egocentric scene understanding tasks and often include excessive noisy camera motion, rendering them unsuitable for generative training. Additionally, EgoGen [31], a synthetic dataset, can not fully encapsulate the complex variations inherent in real-world egocentric views. To address this gap, we introduce the *EgoVid-5M* dataset, a pioneering and meticulously curated collection designed explicitly for egocentric video generation. *EgoVid-5M* comprises 5M egocentric video clips with precise action annotations and cleansing metadata.

3. EgoVid-5M

The training of video generation relies on large-scale, high-quality video data. Therefore, we built *EgoVid-5M* based on the large-scale Ego4D dataset [16]. Notably, although Ego4D contains thousands of hours of egocentric videos, it is intended for egocentric perception and includes excessive noisy camera motion that is unsuitable for generative training. Additionally, the narration annotation in Ego4D is overly simplistic and lacks semantic consistency with frames. To address these issues, we propose a data annotation pipeline that provides detailed and accurate annotations of fine-grained kinematic control and high-level action descriptions. Furthermore, a data cleaning pipeline is developed to ensure alignment between action descriptions and video content, as well as the magnitude of motion and consistency between frames.

3.1. Data Annotation Pipeline

In order to simulate egocentric videos from actions, we construct detailed and accurate action annotations for each video segment, encompassing low-level kinematic control (e.g., ego-view translation and rotation), as well as high-level textual descriptions. The annotation pipeline is shown in the upper part of Fig. 2.

Kinematic Control Annotation In order to accurately describe complex egocentric movements, we utilize the Visual-Inertial Odometry (VIO) method to construct kinematic control signals. This involves using ParticleSfM [80] to obtain scale-ambiguous camera poses P_c from video, fol-

lowed by integrating IMU signals $\{I_t\}_{t=0}^{T-1}$ to obtain more accurate and scaled camera poses. However, there are several challenges to overcome. (1) The IMU signals are subject to noise. (2) The transformation matrix between the IMU and the camera is unknown. (3) The initial velocity of the IMU is unknown. (4) The scale factor of the P_c is unknown. To address the aforementioned problems, we first utilize high-pass Butterworth filters $\mathcal{F}_{IFFT}(\mathcal{H}_{\text{low}}(s) \cdot \mathcal{F}(s))$ and low-pass Butterworth filters $\mathcal{F}_{IFFT}(\mathcal{H}_{\text{high}}(s) \cdot \mathcal{F}(s))$ to filter out the gravity signal and high-frequency noise, where $\mathcal{F}(s) = \mathcal{F}_{FFT}(I)$ is the *Fast Fourier Transform* and \mathcal{F}_{IFFT} is the inverse operation. $\mathcal{H}_{\text{low}}(s) = \frac{1}{1+(\frac{s}{w_c})^{2n}}$ is the low-pass filter, $\mathcal{H}_{\text{high}}(s) = \frac{(\frac{s}{w_c})^{2n}}{1+(\frac{s}{w_c})^{2n}}$ is the high-pass filter, w_c represents the cutoff frequency while n represents the filter order. Next, we propose a quality filter to drop the low-quality P_c and I , where the motivation is that the number of reconstructed points N_p (generated from ParticleSfM) is a reflection of the accuracy of P_c [66], and the variance of IMU reflects the dynamic nature of the video. Therefore, the retained data needs to simultaneously satisfy $N_p \geq N_{\text{thres}}$ and $\frac{1}{T} \sum_{t=0}^{T-1} (I_t - \bar{I})^2 \leq V_{\text{thres}}$. Next, we perform the least squares minimization with P_c and the integrated IMU signal $\{I_t\}_{t=0}^{T-1}$ to calculate the initial velocity $v(0)$ of the IMU signal, the transformation matrix T_I from IMU to the camera, and the scale factor λ of the P_c :

$$\min_{v_0, T_I, \lambda} |T_I P_I (T - 1) - \lambda P_c|^2, \quad (1)$$

where $P_I(T - 1)$ can be derived from:

$$P_I(t + 1) = P_I(t) + v(t)\Delta t + \frac{1}{2}I(t)\Delta t^2, \quad (2)$$

$$v(t + 1) = v(t) + I(t)\Delta t, \quad (3)$$

with the initial condition $P(0) = \mathbf{0}$. Finally, we utilize the Kalman filter to fuse these two signals under the camera coordinate (see supplement for more details).

Textual Description Annotation In addition to kinematic control, another supplementary information of egocentric action is textual descriptions. In the Ego4D dataset, only human narrations serve as text annotations, but the narrations are relatively simple and lack semantic consistency with frames (see supplement). Therefore, we utilize a multimodal large language model (MLLM) to provide detailed action captions for the videos. Considering that existing open-source multimodal language models are not as proficient in following instructions as large language models (LLM), we first prompt LLaVA-NeXT-Video-32B-Qwen [77] to provide detailed captions for videos (including foreground, background, main subjects, and action information). Then, we prompt Qwen2 [70] to summarize egocentric action descriptions from the aforementioned captions, with human narrations as the supplementary prompt. Through the combination of MLLM

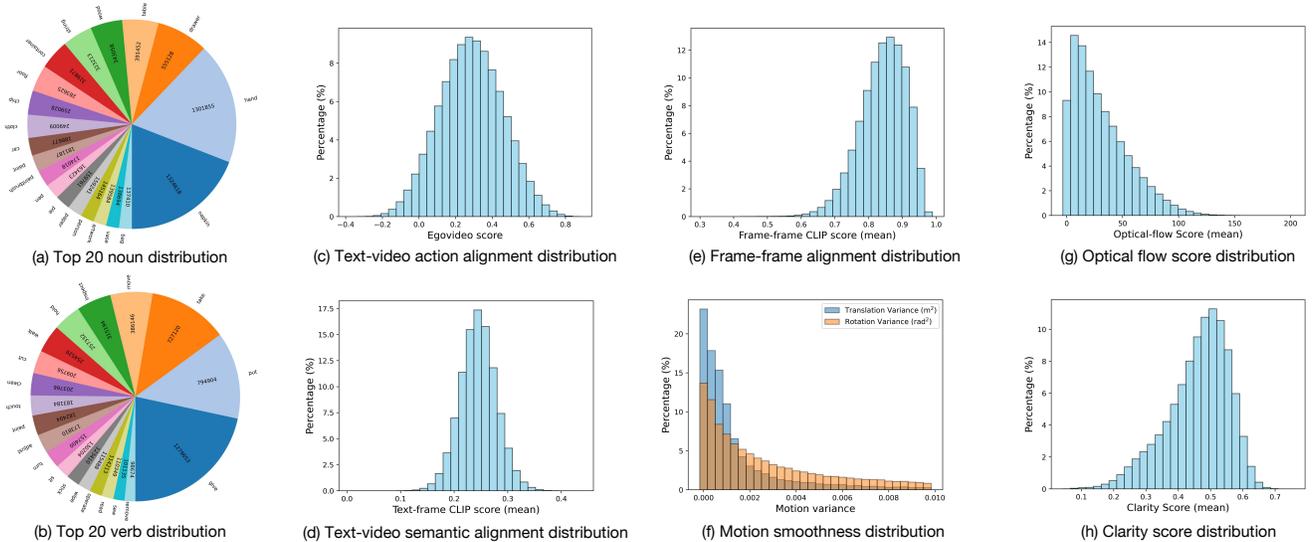


Figure 3. Data annotation distribution of *EgoVid-5M*. (a) and (b) describe the quantities of the top 20 verbs and nouns. (c) Text-video action alignment is assessed using the EgoVideo score. (d) and (e) measure the semantic similarity between text and frames and between frames and the first frame using the average CLIP score. (f) Motion smoothness is quantified by the variance of translation and rotation. (g) Motion strength is represented by the average global optical flow. (h) Video clarity is determined by the DOVER score.

and LLM, our textual descriptions can accurately describe egocentric action while ensuring semantic consistency. We also utilize LLM to analyze the *Nouns* and *Verbs* in each textual description, and classify them into hundreds of action categories (as shown in the Fig. 3(a)-(b)). The resulting textual descriptions include actions in household environments, outdoor settings, office activities, sports, and skilled operations, thus covering the majority of scenes encountered in egocentric perspectives.

3.2. Data Cleaning Pipeline

The data quality significantly influences the effectiveness of training generative models. Prior works [3, 44, 55] have delved into various cleaning strategies to improve video datasets, focusing on aesthetics, semantic coherence, and optical flow magnitude. Based on these cleaning strategies, this paper presents a specialized cleaning pipeline specifically designed for egocentric scenarios. The pipeline is illustrated in the lower part of Fig. 2.

Text-video Consistency We utilize CLIP EgoVideo [45] and [49] to evaluate the alignment between textual descriptions and video frames, leveraging EgoVideo’s focus on action alignment and CLIP’s emphasis on global semantic similarity. In particular, evenly-spaced frames are gathered to calculate the EgoVideo similarity with the text. (refer to Fig. 3(c) for EgoVideo score distribution). Subsequently, these four frames are separately extracted to calculate CLIP similarity with the corresponding text (see Fig. 3(d) for CLIP similarity score distribution).

Frame-frame Consistency The higher the semantic consistency

between video frames, the more conducive it is to generative training. To analyze this relationship, we uniformly extract three frames alongside the first frame to compute frame CLIP similarity. The distribution of semantic consistency is illustrated in Fig. 3(e).

Motion Smoothness Excessive egocentric motion can lead to video fluctuations, which is detrimental to training visual generation models. To address this issue, we propose measuring the degree of translation variation $\frac{1}{T} \sum_{t=0}^{T-1} (Tr_t - \overline{Tr})^2$ and rotation variation $\frac{1}{T} \sum_{t=0}^{T-1} (Ro_t - \overline{Ro})^2$ to quantify motion smoothness, where Tr and Ro are translation and rotation measured in Sec. 3.1 (see motion smoothness distribution in Fig. 3(f)).

Motion Strength A typical approach to describe video motion strength is optical flow [56]. Therefore, we first represent video motion by averaging global optical flow (see motion strength distribution in Fig. 3(g)), we additionally calculate the *five-point* optical flow, which includes the proportion of optical flow score across pixel intervals: 0–4, 4–8, 8–12, 12–16, and above 16 (more details see supplement). This method offers a multi-faceted perspective on motion strength, addressing both the movement of small foreground objects and the overall camera motion.

Clarity Assessment For egocentric scenes, clarity and realism are paramount. Therefore, instead of relying on CLIP for aesthetic scoring [9], we apply DOVER [63] to assess video clarity (refer to Fig. 3 for DOVER score distribution), prioritizing visual sharpness and detail in our dataset.

Based on the cleansing metadata, we vary thresholds to filter and obtain high-quality training data. Specifically,

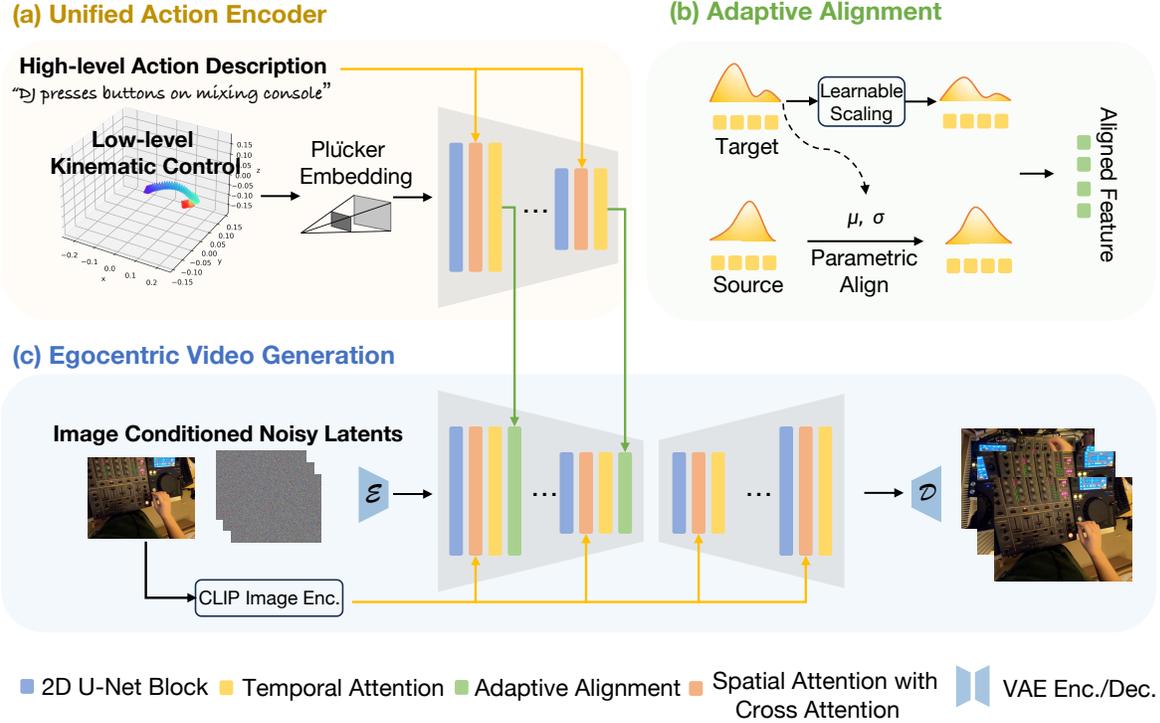


Figure 4. The overall framework of *EgoDreamer*. *EgoDreamer* introduces (a) the Unified Action Encoder to embed different action inputs simultaneously, and it utilizes (b) the Adaptive Alignment to integrate action conditions into the egocentric video generation branch (c).

experiments are conducted in Sec. 5.2 to explore the effects of three mainstream cleaning strategies on egocentric video generation training. Additionally, given the significance of data cleaning strategies in training video generation models [3, 44, 55], and the substantial computational cost—**thousands of GPU days**—to annotate and clean millions of videos, we release all annotation and cleansing metadata to encourage community research into the impact of various cleaning strategies on egocentric video training.

4. EgoDreamer

In the context of ego-centric world simulators, action-driven video generation is paramount. However, existing action-driven video generation approaches [20, 24, 26, 27, 40, 62, 66, 73] primarily focus on camera movements within static scenes, making it challenging to model complex egomotion. Therefore, we propose *EgoDreamer*, which can produce egocentric videos driven simultaneously by high-level action descriptions and low-level kinematic control. As illustrated in Fig. 4, *EgoDreamer* adopts a similar architecture of [65] to enable image-conditioned video generation. Besides, *EgoDreamer* features two key innovations: (1) It introduces a Unified Action Encoder (UAE) that embeds two distinct action inputs, allowing for a more nuanced representation of ego movements. (2) It leverages Adaptive Alignment (AA) that encapsulates multi-scale control sig-

nals in the parametric alignment perspective, enhancing the action control efficacy.

Unified Action Encoder. In this framework, the UAE simultaneously encodes both low-level and high-level actions. Specifically, it first utilizes Plücker embedding [20, 52] to encode kinematic signals:

$$\mathbf{p}_{u,v} = (\mathbf{t} \times \mathbf{d}_{u,v}, \mathbf{d}_{u,v}), \quad (4)$$

$$\mathbf{d}_{u,v} = \mathbf{R}\mathbf{K}^{-1}[u, v, 1]^T + \mathbf{t}, \quad (5)$$

where \mathbf{R} and \mathbf{t} is the rotation matrix and translation vector, \mathbf{K} is the intrinsic matrix, and $\mathbf{p}_{u,v}$ is the Plücker embedding at pixel (u, v) . Then, low-level signal \mathbf{p} is encoded through a series of U-Net blocks, while a high-level action description d is simultaneously embedded via CLIP [49] and cross-attention mechanisms. The action output A of one U-Net block can be formulated as:

$$A = \mathcal{F}_t(\mathcal{F}_c(\mathcal{F}_s(\mathcal{F}_{\text{conv}}(\mathbf{p})), \text{CLIP}(d))), \quad (6)$$

where \mathcal{F}_t is the temporal self-attention, \mathcal{F}_c is the cross-attention, \mathcal{F}_s is the spatial self-attention, $\mathcal{F}_{\text{conv}}$ is the 2D convolution block. Notably, previous methods [20, 66] encode text and kinematics separately, ignoring that low-level kinematics and high-level action descriptions are coupled. In contrast, the proposed UAE focuses on modeling the relationship between different action inputs, thus the generated action control signals capture both camera movements and complex egocentric dynamics (e.g., hand interactions).

open tool chest, places metal tool container in drawer

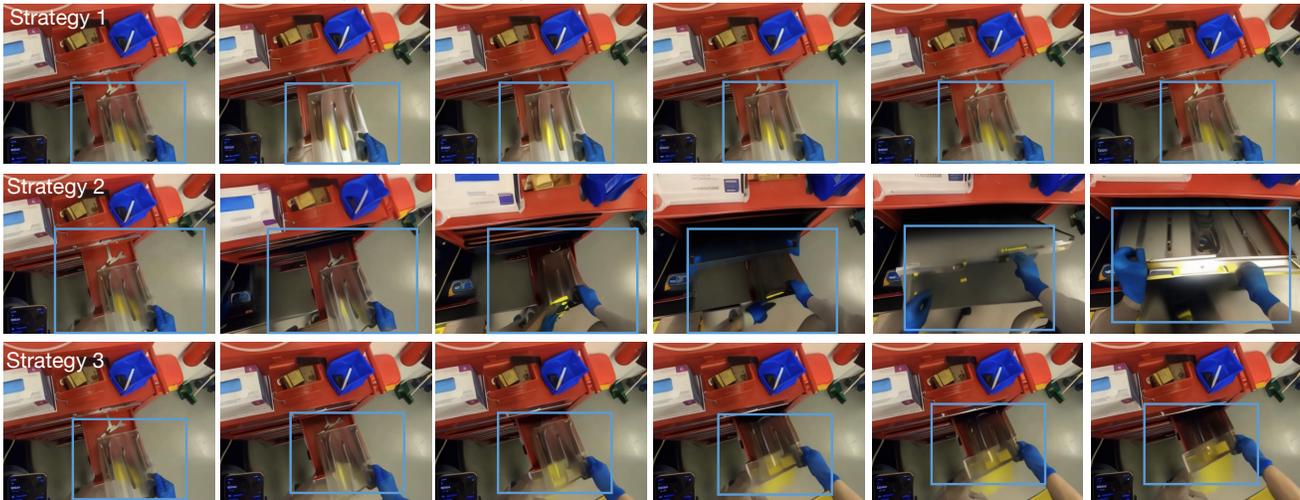


Figure 5. The video visualization comparison across different data cleaning strategies reveals distinct outcomes, where the blue box highlights the difference. Videos generated by strategy-1 fail to capture local motion and tend to be stationary. In contrast, videos produced by strategy-2 exhibit excessive motion, compromising semantic coherence. Meanwhile, videos generated by strategy-3 effectively model intricate hand movements, striking a balance between motion strength and semantic fidelity.

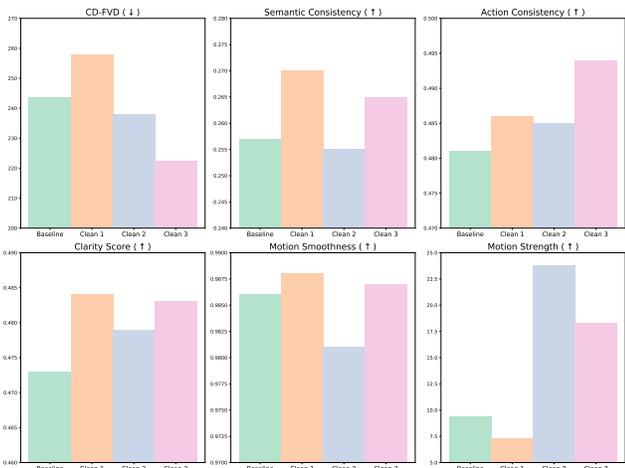


Figure 6. Video generation quantitative comparisons between different data cleaning strategies, where the baseline is DynamiCrafter [65] initialized with its original weights.

Adaptive Alignment. Based on the multi-scale U-Net architecture, the UAE outputs multi-scale $\{A_i\}_{i=0}^3$. Then *EgoDreamer* encapsulates control signals in the perspective of parametric alignment:

$$L_i = \alpha L_i + \frac{A_i - \mu_L}{\sigma_L}, \quad (7)$$

where L_i is the output of one U-Net block in the main Diffusion branch, α is a learnable parameter, μ_L, σ_L are the mean and standard deviation of L_i . The introduced AA module is inspired by cross normalization [46] and applies it to multi-

scale U-Net feature alignment. Compared to ControlNet’s zero-initialization [76], our method achieves better control effectiveness.

5. Experiment

5.1. Experiment Details

Dataset. The proposed *EgoVid-5M* dataset is partitioned as the training set and the validation set. For the validation set, we select samples with high text-video semantic consistency, moderate video motion, high video clarity, and diverse scene coverage including household environments, outdoor settings, office activities, sports, and skilled operations. This resulted in a final validation set *EgoVid-val* with 1.2K samples, with a training set *EgoVid-train* with 4.9M samples. Notably, due to the known issue in Ego4D IMU annotation¹, we annotate kinematic controls for 65K video samples with accurate IMU data. The annotated subset *EgoVid-65K* is $\sim 5\times$ larger than the current largest kinematic annotation dataset [66], which is utilized further to train the ability of kinematic control video generation.

Training We validate the effectiveness of our *EgoVid-5M* using video diffusion baselines with different architectures, including U-Net (SVD [3] and DynamiCrafter [65]), and DiT (OpenSora [81]). Building upon these pre-trained models, we employ a continuous training approach to train 480p videos for enhanced training efficiency. For *EgoDreamer*, we first initialize it with pre-trained weights [65],

¹<https://ego4d-data.org/docs/data/imu/>

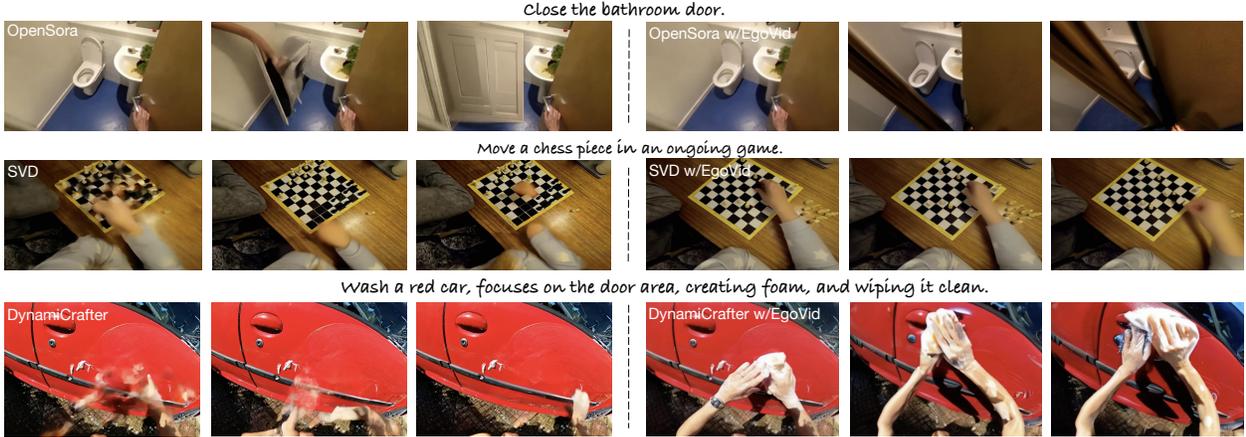


Figure 7. Visualizations demonstrate that *EgoVid*-fintuned baselines (OpenSora [81], SVD [3], DynamiCrafter [65]) generate egocentric videos with stronger frame-consistency and better semantic-alignment.

Method	w. <i>EgoVid</i>	CD-FVD ↓	Semantic Consistency ↑	Action Consistency ↑	Clarity Score ↑	Motion Smoothness ↑	Motion Strength ↑
SVD [3]	✗	591.61	0.258	0.465	0.479	0.971	18.897
SVD [3]	✓	548.32	0.266	0.471	0.485	0.974	21.032
DynamiCrafter [65]	✗	243.63	0.257	0.481	0.473	0.986	9.357
DynamiCrafter [65]	✓	236.82	0.265	0.494	0.483	0.987	18.329
OpenSora [81]	✗	809.46	0.260	0.489	0.520	0.983	7.608
OpenSora [81]	✓	718.32	0.266	0.494	0.528	0.986	15.871

Table 2. *EgoVid* significantly enhances egocentric video generation. Experimental results demonstrate that training with *EgoVid* improves performance across all three baselines on six metrics.

then *EgoDreamer* are further trained on *EgoVid* to adapt to egocentric scenes. Finally, we finetune the proposed UAE and AA using *EgoVid-65K*. All experiments are conducted on NVIDIA A800 GPUs. For additional training details, please refer to the supplementary materials.

Evaluation. We adopt a set of metrics from AIGCBench [11] and VBench [28] to assess the quality of the generated egocentric videos. Specifically, our evaluation metrics utilize the CD-FVD [12] for spatial and temporal quality, the CLIP [49] for semantic consistency, the EgoVideo [45] for action consistency, the DOVER [63] for clarity score, frame interpolation model [33] for motion smoothness, and RAFT [56] for motion strength. Additionally, following [20, 66], we assess kinematic control consistency using translation error and rotation error, which measures the difference between COLMAP poses and the ground truth poses in the canonical space [66]. The specific calculations for each metric are detailed in the supplement.

Next, we verify the impact of different data cleaning strategies on egocentric video generation. Subsequently, we substantiate, quantitatively and qualitatively, that the proposed *EgoVid* can enhance various baselines’ egocentric video generation capabilities. Finally, experiments are conducted to demonstrate that the proposed *EgoDreamer* can generate egocentric videos under the control of both action

descriptions and kinematic signals.

5.2. Data Cleaning Strategy Comparison

In this subsection, we employ the state-of-the-art video diffusion model DynamiCrafter [65] as the baseline, which is trained on the *Image+Text-to-Video* task to evaluate various data cleaning strategies.

Strategy-1. This strategy focuses on ensuring text-frame consistency (with $CLIP_{TF} \geq 0.275$) and frame-frame consistency ($CLIP_{FF} \geq 0.8$). Additionally, we retained videos with an average optical flow ≥ 3 and a DOVER score ≥ 0.3 . This process yielded a subset *EgoVid-1M-1*. DynamiCrafter is finetuned for one epoch using this subset. As illustrated in Fig. 6, this model achieved the highest semantic consistency metrics. However, the stringent criteria for both text-frame and frame-frame consistency favored the retention of videos with slow motion. Consequently, the motion strength of the generated videos falls below the baseline, which is not desirable for effective video generation.

Strategy-2. The thresholds for text-frame consistency and frame-frame consistency are relaxed ($CLIP_{TF} \geq 0.27$, $CLIP_{FF} \geq 0.75$). Besides, we retain videos with an average optical flow between 3 and 40, and those with a DOVER score ≥ 0.3 . This strategy results in a subset *EgoVid-1M-2*. Upon finetuning DynamiCrafter for one full epoch, as



Figure 8. Visualizations show that *EgoDreamer* can realize distinct action controls based on different text descriptions.

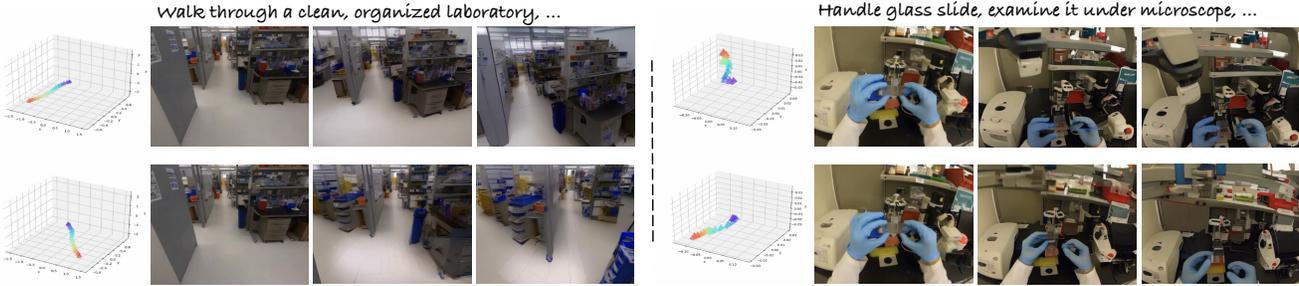


Figure 9. Visualizations demonstrate that *EgoDreamer* can generate various egocentric videos based on different low-level commands.

shown in Fig. 6, we observe a significant improvement in the motion strength. However, the accelerated motion introduces artifacts, leading to visual fragmentation. Consequently, this negatively impacts the text-frame semantic consistency, resulting in scores below the baseline.

Strategy-3. we further relax the thresholds for text-frame consistency ($CLIP_{TF} \geq 0.26$) and frame-frame consistency ($CLIP_{FF} \geq 0.7$), while introducing an action consistency constraint (EgoVideo score ≥ 0.22). Videos are retained with an average optical flow between 3 and 35, as well as those with a DOVER score ≥ 0.3 . Notably, as mentioned in Sec. 3.2, we also retain videos with average optical flow values below 3, provided that the proportion of optical flow (≥ 12 pixels) is greater than 3%. This resulted in the *EgoVid-IM-3* subset. Compared to the previous two strategies, the model finetuned on *EgoVid-IM-3* effectively enhances both semantic and action consistency while ensuring moderate motion strength, achieving the best CD-FVD score. Furthermore, the *5-point* optical flow filtering method allowed for a focus on local motion scenarios. As illustrated in Fig. 5, strategy-3 accurately models intricate hand movements, in contrast to the stationary visuals of strategy-1 and the exaggerated motion of strategy-2.

5.3. Enhancement in Egocentric Video Generation

In this subsection, experiments are conducted to verify that the proposed *EgoVid* enhances the egocentric video generation capabilities of various baselines. Specifically,

SVD [3], DynamiCrafter [65], and OpenSora [81] are selected as baselines, which are initialized with their original weights, and then we employ *EgoVid-IM-3* for finetuning. For training efficiency and fair comparison, we resize all input video to 480p and focus exclusively on the *Image+Text-to-Video* tasks. As shown in Tab. 2, the experiment results demonstrate that training with *EgoVid* improves performance across all three baselines on six different metrics. Specifically, the *EgoVid* finetuning significantly enhances the motion strength of egocentric videos while also improving consistency in text-video alignment, action-video alignment, and overall image clarity. Consequently, the CD-FVD metric shows a notable improvement. Additionally, we conduct a visualization comparison of different baselines before and after finetuning, as illustrated in Fig. 7. Prior to *EgoVid* finetuning, various baselines exhibit issues such as frame fragmentation and distortion in egocentric scenarios (e.g., appearance of incongruous objects and hand fragmentation). This underscores the inadequacy of most existing video generation models in egocentric contexts. However, after the *EgoVid* finetuning, the generated videos not only achieve superior alignment with text prompts, but also exhibit enhancement in visual quality.

5.4. EgoDreamer Experiments

In this subsection, we conduct experiments to demonstrate that *EgoDreamer* can generate egocentric videos under the control of both action descriptions and kinematic

w. <i>EgoVid</i>	ControlNet	ControlNeXt	AA	UAE	CD-FVD ↓	Semantic Consistency ↑	Action Consistency ↑	Rot Err ↓	Trans Err ↓
	✓				241.90	0.263	0.490	5.32	9.27
✓	✓				238.87	0.266	0.493	4.01	8.66
✓	✓			✓	239.01	0.268	0.494	3.58	8.41
✓		✓		✓	234.13	0.269	0.497	3.59	7.93
✓			✓	✓	229.82	0.268	0.498	3.28	7.62

Table 3. Ablation study on training strategy and different components of *EgoDreamer*.

signals. Additionally, the efficacy of the proposed UAE and AA modules will be validated. In our experiments, we initialize *EgoDreamer* using weights from [65]. The results are presented in Tab. 3. In Row-1, the low-level kinematic control signals are integrated via ControlNet [12], which resembles [20, 66]. Row-2 utilizes *EgoVid-1M-3* to pre-train the model. Compared with Row-1, results indicate significant improvements across five metrics after *EgoVid-1M-3* finetuning. In Row-3, we further introduce the UAE module to strengthen the association between low-level kinematic control and high-level action descriptions. The experimental results indicate that this enhancement further improves action alignment and reduces the deviation in low-level kinematic control compared to Row-2. In Row-4 and Row-5, we replace the ControlNet with ControlNext [46] and the AA module. The results reveal that the AA module exhibits superior performance compared to both ControlNet and ControlNext, as it facilitates learnable parameterized alignment from a multi-scale perspective. Finally, we visualize videos generated by *EgoDreamer*, as depicted in Fig. 8. Under initial frame conditions, varying the input text descriptions enables *EgoDreamer* to realize distinct action controls. Furthermore, as illustrated in Fig. 9, with the same initial frame, the model can generate videos that incorporate a composite of multiple low-level kinematic controls. Notably, *EgoDreamer* to produce videos with meter-level movements (e.g., walking) and centimeter-level nuanced movements (e.g., intricate hand actions in a laboratory environment). Additional visualizations can be found in the supplement.

6. Discussion and Conclusion

In this paper, we present *EgoVid-5M*, which is the first high-quality dataset meticulously curated for egocentric video generation, comprising 5 million video clips enriched with detailed action annotations. This dataset effectively addresses the challenges associated with the dynamic nature of egocentric perspectives, the intricate diversity of actions, and the complex variety of encountered scenes. The implementation of a sophisticated data cleaning pipeline further ensures the dataset’s integrity and usability, maintaining frame consistency, action coherence, and motion smoothness under egocentric conditions. Additionally, we propose

EgoDreamer, which showcases the ability to generate egocentric videos by simultaneously incorporating action descriptions and kinematic control signals, thereby enhancing the realism and applicability of generated content. We hope that the proposed *EgoVid-5M* dataset, along with the associated annotations and metadata, will serve as a valuable resource for the research community. We encourage researchers to leverage these innovations to propel further exploration and development in the realm of egocentric video generation, ultimately advancing applications in virtual reality, augmented reality, and gaming.

References

- [1] Peri Akiva, Jing Huang, Kevin J Liang, Rama Kovvuri, Xingyu Chen, Matt Feiszli, Kristin Dana, and Tal Hassner. Self-supervised object detection from egocentric videos. In *ICCV*, 2023. 3
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2, 3
- [3] 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, 3, 5, 6, 7, 8, 9, 14
- [4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2, 3
- [5] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. 2, 3
- [6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 3

- [7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 3
- [8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 2, 3
- [9] Schuhmann Christoph. Improved-aesthetic-predictor, 2022. 5
- [10] Danijar Hafner and Timothy P. Lillicrap and Mohammad Norouzi and Jimmy Ba. Mastering atari with discrete world models. In *ICLR*, 2021. 2, 3
- [11] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2023. 8, 17
- [12] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fr chet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8, 10, 17
- [13] Anastasis Germanidis. Introducing general world models. 2023. 2
- [14] Anastasis Germanidis. Introducing gen-3 alpha. 2024. 3
- [15] Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *CVPR*, 2023. 3
- [16] 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 *CVPR*, 2022. 2, 4
- [17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrahm Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatuminu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. 2, 4
- [18] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020. 2, 3
- [19] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2, 3
- [20] 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. 3, 6, 8, 10, 17
- [21] 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. 3
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 15
- [23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3
- [24] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3, 6
- [25] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2, 3
- [26] Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024. 3, 6
- [27] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 3, 6
- [28] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 8, 17
- [29] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 15
- [30] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, Apr. 2024. 2, 3
- [31] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mi-hai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang.

- EgoGen: An Egocentric Synthetic Data Generator. In *CVPR*, 2024. 4
- [32] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021. 3
- [33] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, 2023. 8
- [34] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wen-zhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 2022. 3
- [35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 15
- [36] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *ACMMM*, 2021. 3
- [37] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *CVPR*, 2022. 3
- [38] 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. 15
- [39] Hongchen Luo, Kai Zhu, Wei Zhai, and Yang Cao. Intention-driven ego-to-exo video generation. *arXiv preprint arXiv:2403.09194*, 2024. 3
- [40] Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*, 2023. 3, 6
- [41] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3
- [42] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *ICCV*, 2023. 3
- [43] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 2
- [44] Kepan Nan, Rui Xie, Penghao Zhou, Tiejhan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 2, 4, 5, 6
- [45] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 3, 5, 8, 17
- [46] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 7, 10
- [47] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In *ICCV*, 2023. 3
- [48] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *ICCV*, 2023. 3
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6, 8, 14, 17
- [50] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 2
- [51] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 4
- [52] Vincent Sitzmann, Semon Rezkikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *NeurIPS*, 2021. 6
- [53] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 4
- [54] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3
- [55] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024. 2, 5, 6
- [56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 5, 8, 14, 17
- [57] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 17
- [58] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 2, 3
- [59] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv: 2401.09985*, 2024. 3
- [60] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*, 2023. 2, 3
- [61] Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 3

- [62] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024. 3, 6
- [63] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 5, 8, 17
- [64] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *CoRL*, 2023. 2, 3
- [65] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023. 2, 3, 6, 7, 8, 9, 10, 14, 15
- [66] Dejie Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3, 4, 6, 7, 8, 10, 17
- [67] Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Egopca: A new framework for egocentric hand-object interaction understanding. In *ICCV*, 2023. 3
- [68] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 2
- [69] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3
- [70] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 4, 14
- [71] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *CVPR*, 2024. 2, 3
- [72] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024. 2, 3
- [73] 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*, 2024. 3, 6
- [74] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 3
- [75] Yu, Lijun, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 3
- [76] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 7
- [77] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llavavnext: A strong zero-shot video understanding model, 2024. 4, 14
- [78] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arXiv preprint arXiv:2410.13571*, 2024. 2, 3
- [79] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 2, 3
- [80] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *ECCV*, 2022. 4, 14
- [81] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. 3, 7, 8, 9, 15
- [82] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. 2, 3
- [83] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Ni-anchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 3

In the supplement materials, we first elaborate on the annotation and cleaning details of *EgoVid*. Then we present additional training details of different baselines and the proposed *EgoDreamer*. Subsequently, the evaluation details are elaborated. Finally, we present additional visualizations.

7. Annotation and Cleaning Details

7.1. Kinematic Annotation Details

To enhance kinematic annotation accuracy, we fuse camera poses from IMU and ParticleSfM [80], utilizing the Kalman filter. First, we filter the IMU data to remove gravitational components and noise. Next, we employ least squares estimation to determine the initial velocity and scale factor for the ParticleSfM poses. Finally, we align both the IMU poses and ParticleSfM results to the camera coordinate system (detailed explanations of these processes can be found in the main text). The Kalman filter implementation involves the following steps:

The state vector $\mathbf{x} = [x, y, z, q_1, q_2, q_3, q_4, v_x, v_y, v_z]$ is initialized from IMU pose to represent position, quaternion, and velocity. The error covariance matrix \mathbf{P} , process noise covariance \mathbf{Q} and observation noise covariance \mathbf{FR} are initialized as $0.1 \cdot \mathbf{I}_{10 \times 10}$, $0.01 \cdot \mathbf{I}_{10 \times 10}$ and $0.1 \cdot \mathbf{I}_{7 \times 7}$. In the prediction step, the state transition function \mathbf{f} is applied to predict the next state:

$$\mathbf{x}_{k|k-1} = \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_k), \quad (8)$$

where \mathbf{u}_k consists of IMU readings, and \mathbf{f} predicts the next state by updating the current state through integration, incorporating the linear acceleration and angular velocity measured by the IMU. The covariance of the predicted state is updated as:

$$\mathbf{P}_{k|k-1} = \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T + \mathbf{Q}, \quad (9)$$

where \mathbf{F} is the Jacobian of the transition matrix. In the update phase, we compute the measurement residual \mathbf{y}_k :

$$\mathbf{y}_k = \mathbf{x}'_k - \mathbf{H}\mathbf{x}_{k|k-1}, \quad (10)$$

where $\mathbf{x}' = [x', y', z', q'_1, q'_2, q'_3, q'_4]$ is the ParticleSfM pose, $\mathbf{H} = \begin{bmatrix} \mathbf{I}_{3 \times 3} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{4 \times 4} \end{bmatrix}$ is the Jacobian of the observation model.

The innovation covariance \mathbf{S}_k is given by:

$$\mathbf{S}_k = \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R}, \quad (11)$$

and the Kalman gain is calculated by:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1}. \quad (12)$$

The state estimate is then updated:

$$\mathbf{x}_k = \mathbf{x}_{k|k-1} + \mathbf{K}_k\mathbf{y}_k. \quad (13)$$

Finally, the error covariance matrix is updated:

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\mathbf{P}_{k|k-1}. \quad (14)$$

This iteration continues for each IMU reading, yielding a refined series of pose estimates.

7.2. Data Cleaning Details

Five-Point Optical Flow Filtering. A typical approach to describe video motion strength is optical flow [56]. Therefore, we first represent video motion by averaging global optical flow. Notably, this approach only encapsulates the average motion magnitude. However, in egocentric scenarios, where a substantial portion of the scene remains static and only foreground elements (e.g., hands) exhibit motion, applying a filtering strategy based solely on average optical flow may result in the inadvertent exclusion of valuable, fine-grained hand movement data. Therefore, as a supplement, we calculate the *five-point* optical flow, which involves the proportion $P_{m \sim n}$ of optical flow score across different pixel intervals:

$$P_{m \sim n} = \frac{\sum_{x,y} \delta(m \leq |F(x,y)| < n)}{N}, \quad (15)$$

where N is the total pixel number, F is the optical flow map, δ is the indicator function. Specifically, we calculate $P_{0 \sim 4}, P_{4 \sim 8}, P_{8 \sim 12}, P_{12 \sim 16}, P_{16 \sim}$, their distribution is shown in Fig 11. We performed data filtering based on the *five-point* optical flow, as illustrated in Fig. 10, where the average optical flow magnitude is less than 3 pixels, and over 3% of the pixels exhibit motion greater than 12 pixels. The figure shows that although most of the background elements remain static, the hand movements are dynamic and extensive. Such data are beneficial for training egocentric video generation with subtle hand motions.

Semantic Consistency Comparison.

In the Ego4D dataset, only human narrations are available as text annotations. However, these narrations are relatively simple and lack semantic alignment with the video frames. To address this, we first employ a multimodal large language model (MLLM) [77] to generate detailed captions for the videos. Then, a large language model (LLM) [70] is used to summarize egocentric action descriptions from these detailed captions. We calculate the semantic consistency between captions and the frames using CLIP [49]. As shown in Fig. 12, the semantic similarity of our generated captions is significantly higher than that of the original human narrations.

8. Training Details

We validated the effectiveness of our *EgoVid-5M* using video diffusion baselines with different architectures, including U-Net (SVD [3] and DynamiCrafter [65]), and DiT

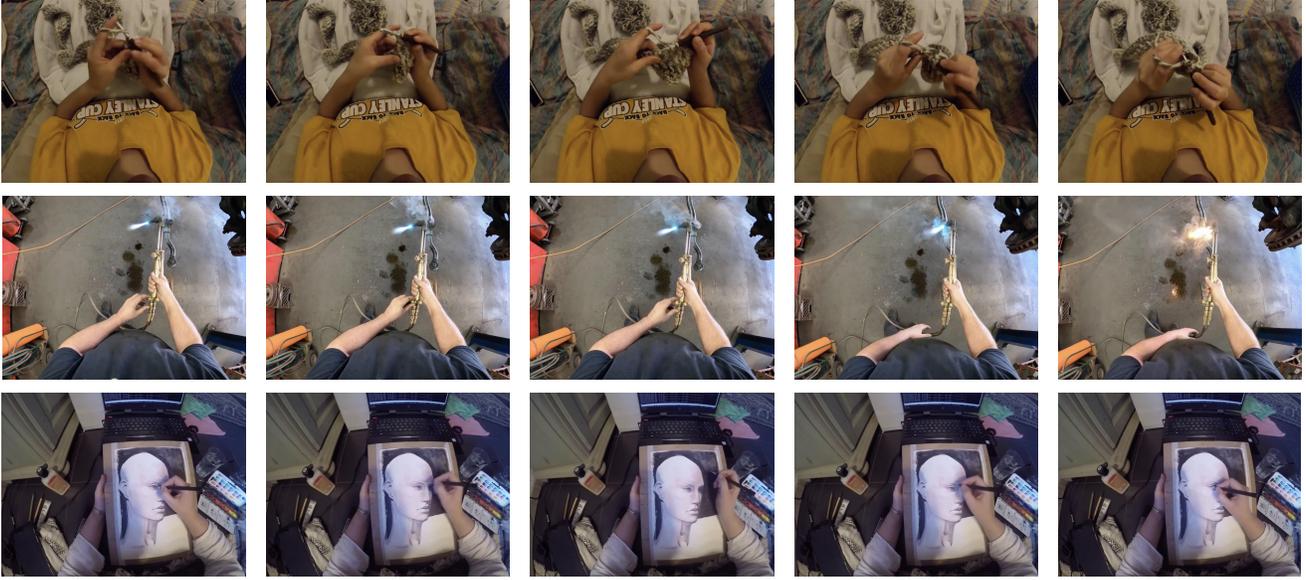


Figure 10. Videos cleaned from the *five-point* optical flow strategy (average optical flow below 3, and the proportion of optical flow (≥ 12 pixels) is greater than 3%). This strategy retains videos with a static background while capturing detailed and extensive motion in hands.

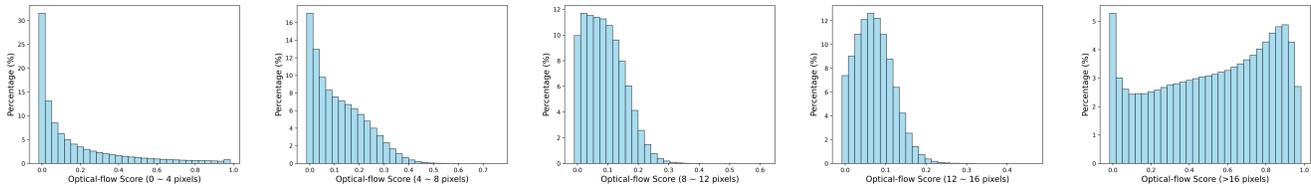


Figure 11. *Five-point* optical flow distribution.

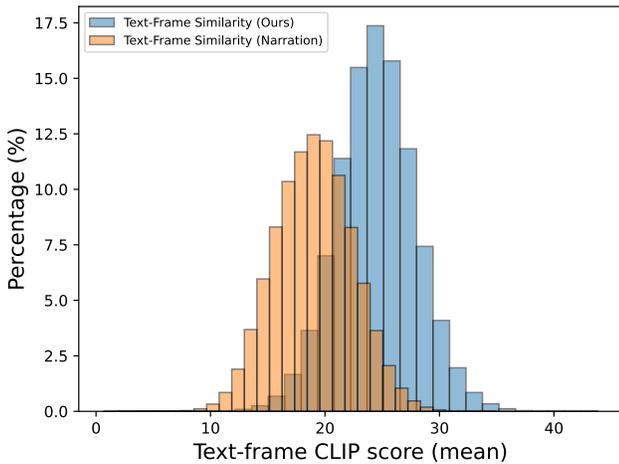


Figure 12. Semantic consistency comparison between our text annotation and the original human narration.

(OpenSora [81]). The training details are as follows: (1) For SVD, we employ the pre-trained 1.1 version² and extend its *img-to-video* architecture to an *Image+Text-to-Video* setup.

²huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt-1-1

Specifically, we replace the image CLIP branch with a text CLIP branch³, which is aligned with the image CLIP version used in SVD. During training, input videos are resized to 480p, and we employed the EDM scheduler [29] with a learning rate of $1e-4$ and a batch size of 64, finetuning on *EgoVid-1M-3* for one epoch. (2) For DynamiCrafter, we leverage the pre-trained model at 512 resolution⁴. Videos are resized to 480p during training, utilizing the DDPM scheduler [22] with a learning rate of $1e-5$ and a batch size of 64. The finetuning was conducted on *EgoVid-1M-3* for one epoch. (3) For OpenSora, we used the pre-trained version 1.2 model⁵, adjusting its data bucket strategy to train only on 480p inputs, and set mask ratios to mask only the first frame. The model was trained with the RF [35, 38] scheduler, a learning rate of $1e-4$, and a batch size of 64, using *EgoVid-1M-3* for one epoch.

For *EgoDreamer*, we first initialize it with the pre-trained model at 512 resolution [65], then *EgoDreamer* are further trained on *EgoVid-1M-3* to adapt to egocentric scenes, with batch size 64 and learning rate $1e-5$. Finally, we finetune

³huggingface.co/openai/clip-vit-base-patch32

⁴huggingface.co/Doubiiu/DynamiCrafter_512

⁵huggingface.co/hpcai-tech/OpenSora-STDiT-v3

Closing the bathroom door.



Walk into the bathroom.



Clean the wall, ...



Clean the stovetop, ...



Clean the white countertop, ...



Clean the window, ...



Sit on gray couch, place white mug on table.

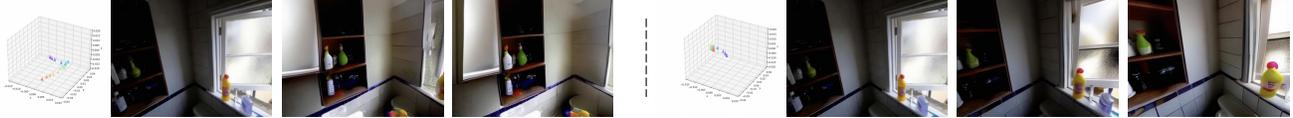


Sit on gray couch, hold white mug with both hands.



Figure 13. Visualizations showing that *EgoDreamer* can generate action-driven egocentric videos based on high-level text descriptions.

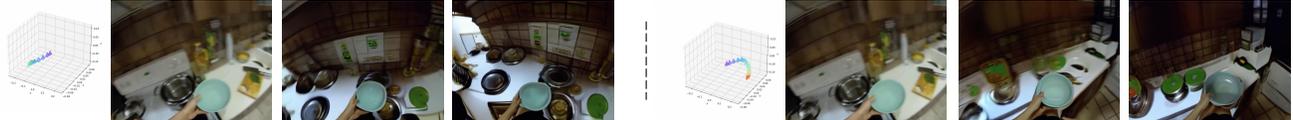
Focus on a bathroom shelf with cleaning supplies and toiletries.



Working on a motorcycle part with attention to detail.



Hold bowl, prepare to transfers noodles.



Push cart with potted plants through lab, past workstations and equipment.



Figure 14. Visualizations showing that *EgoDreamer* can generate action-driven egocentric videos based on low-level kinematic control.

the proposed Unified Action Encoder (UAE) and Adaptive

Alignment (AA) using *EgoVid-65K*, with batch size 32 and

learning rate 1e-5.

9. Evaluation Details

The evaluation metrics are mainly from AIGCBench [11] and VBench [28], along with other metrics such as CD-FVD [12], EgoVideo score [45] and kinematic consistency (Translation Error and Rotation Error) [20, 66]. These metrics are as follows:

Overall Quality. CD-FVD⁶ is utilized to measure spatial and temporal quality. Compared with traditional FVD [57], CD-FVD favors both quality and motion of video frames.

Semantic Consistency. CLIP⁷ [49] is employed to calculate the semantic consistency of text and frames. We uniformly sample four frames from each generated video, calculate the similarity between each frame and the text using CLIP, and then compute the average similarity score.

Action Consistency. EgoVideo⁸ [45] is utilized to calculate the action consistency of text and frames. In this metric, four frames are uniformly sampled from each video to calculate the action similarity between frames and text.

Motion Strength. We employed the optical flow score to quantify the motion strength in videos. Specifically, we utilized the RAFT model⁹ [56] to calculate the optical flow score. For each video, we sampled frames at 8-frame intervals as input to the model. The motion strength of the video segment was then determined by averaging the optical flow scores across all sampled frames.

Motion Smoothness. To assess the continuity of motion in the generated video, we utilize the AMT model¹⁰ [45]. Specifically, for a generated video with frames $[f_0, f_1, \dots, f_{2n-1}, f_{2n}]$, we remove the odd-numbered frames, resulting in $[f_0, f_2, \dots, f_{2n}]$. The AMT model is then employed to interpolate the omitted frames $[\hat{f}_1, \hat{f}_3, \dots, \hat{f}_{2n-1}]$. Finally, we compute the mean absolute error between the interpolated frames and the original ones.

Clarity. We leverage DOVER¹¹ [63] to calculate the video clarity, and we use the fused score that focuses on both aesthetic perspective and technical perspective.

Kinematic Consistency. Following [20, 66], we assess kinematic consistency using translation error and rotation error, which measures the difference between COLMAP poses and the ground truth poses in the canonical space:

$$\text{RotErr} = \sum_{i=1}^n \arccos \frac{\text{tr}(\mathbf{R}_{\text{gen}}^i \mathbf{R}_{\text{gt}}^{i\text{T}}) - 1}{2}, \quad (16)$$

$$\text{TransErr} = \sum_{i=1}^n \|\mathbf{T}_{\text{gt}}^i - \mathbf{T}_{\text{gen}}^i\|_2, \quad (17)$$

where $\mathbf{R}_{\text{gen}}^i, \mathbf{R}_{\text{gt}}^i$ are the generated and ground truth rotation matrix for the i -th frame. $\mathbf{T}_{\text{gen}}^i, \mathbf{T}_{\text{gt}}^i$ are translation vectors for the generated and ground truth camera translation in the i -th frame.

10. Visualizations

We conducted additional visualizations of the results generated by EgoDreamer. As shown in Fig. 15, *EgoDreamer* can leverage action descriptions to generate diverse egocentric videos, encompassing scenes such as householding, cooking, knitting, gardening, and music. These videos include both subtle hand movements and more extensive movements involving walking. Furthermore, as illustrated in Fig. 13, given the same initial frame, changing the high-level text descriptions can generate egocentric videos that comply with semantic control. Lastly, as depicted in Fig. 14, given the same initial frame, altering the low-level kinematic control can generate egocentric videos that conform to pose control.

⁶github.com/songweige/content-debiased-fvd

⁷huggingface.co/openai/clip-vit-large-patch14

⁸drive.google.com/file/d/1k6f1eRdcL17IvXtdX_J8WxNbj2Ms3AW/view

⁹github.com/princeton-vl/RAFT

¹⁰huggingface.co/lalala125/AMT/resolve/main/amt-s.pth

¹¹huggingface.co/teowu/DOVER/resolve/main/DOVER.pth

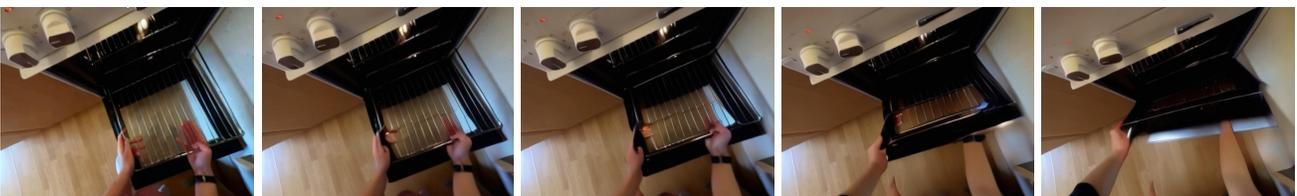
Clean shoes using a brush.



Cooking with a spatula, in the kitchen.



Close the oven door.



Knit on an armchair.



Transfer onions to a pot.



Pushing a lawnmower with both hands, outdoors.



Spinning knobs on the DJ deck.



Figure 15. Visualizations verifying that *EgoDreamer* can generate diverse egocentric videos based on action descriptions.