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

Xiaofeng Wang[1,2] **Kang Zhao**[1] **Feng Liu**[4] **Jiayu Wang**[1]
Guosheng Zhao[2,3] **Xiaoyi Bao**[2,3] **Yingya Zhang**[1] **Xingang Wang**[2] **Zheng Zhu**[2†]
[1] Tongyi Lab, [2] Tsinghua University, [3] CASIA, [4] UCAS
https://egovid.github.io/

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 generation. *EgoVid-5M* encompasses 5 million egocentric video clips and is enriched with detailed action annotations, including 5M high-level textual descriptions and 67K fine-grained kinematic control annotation. 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.

1 Introduction

One of the most promising avenues in video generation is the development of world simulators, which utilize visual simulations and interactions to deliver applications in the physical world. Contemporary research is increasingly validating such capabilities, including applications in autonomous driving [1, 2, 3, 4, 5, 6], autonomous agents [7, 8, 9, 10, 11, 12, 13], and even in general world [14, 15]. In human-centric scenarios, leveraging behavioral actions to drive egocentric video generation has emerged as a pivotal strategy. This approach enhances applications in Virtual Reality (VR), Augmented Reality (AR), and gaming, offering immersive and interactive experiences and advancing the state of the art in these fields.

Video generation requires large amounts of high-quality data, especially for egocentric videos, which are challenging due to their dynamic nature, rich actions, and diverse scenarios. Currently, there is a significant lack of large-scale, specialized datasets for training egocentric video generation models. To address this, we introduce the *EgoVid-5M* dataset, a high-quality resource specifically curated for egocentric video generation. As shown in Tab. 1, *EgoVid-5M* is distinguished by several key features: (1) **High Quality**: It consists of 5M 1080p egocentric videos, rigorously cleaned to ensure alignment

[†]Corresponding author. zhengzhu@ieee.org

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Dataset	Year	Domain	*Gen.*	Text	Kinematic	*CM.*	#Videos	#Frames	Res
HowTo100M [19]	2019	Open	✓	ASR	✗	✗	136M	~ 90	240p
WebVid-10M [20]	2021	Open	✓	Alt-Text	✗	✗	10M	~ 430	Diverse
HD-VILA-100M [21]	2022	Open	✓	ASR	✗	✗	103M	~ 320	720p
Panda-70M [22]	2024	Open	✓	Auto	✗	✗	70M	~ 200	Diverse
OpenVid-1M [23]	2024	Open	✓	Auto	✗	✗	1M	~ 200	Diverse
VIDGEN-1M [24]	2024	Open	✓	Auto	✗	✗	1M	~ 250	720p
LSMDC [25]	2015	Movie	✗	Human	✗	✗	118K	~ 120	1080p
UCF101 [26]	2015	Action	✗	Human	✗	✗	13K	~ 170	240p
Ego4D [27]	2022	Egocentric	✗	Human	IMU	✗	931	$\sim 417K$	1080p
Ego-Exo4D [28]	2024	Egocentric	✗	Human	MVS	✗	740	$\sim 186K$	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.

between action descriptions and video content, and consistent frame quality. (2) **Comprehensive Scene Coverage**: The dataset covers a wide range of scenarios, including household environments, outdoor settings, office activities, sports, and skilled operations, with hundreds of action categories. (3) **Detailed and Precise Annotations**: It includes fine-grained kinematic control and high-level action descriptions. Kinematic information is annotated using Visual Inertial Odometry (VIO), while action descriptions are generated by a multimodal large language model.

Leveraging the proposed *EgoVid-5M*, we train different video generation baselines to validate the dataset's quality and efficacy, including U-Net [16, 17] and DiT [18]) models. Experimental results demonstrate that *EgoVid-5M* 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 a 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* 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 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 [29, 16, 30, 31, 17, 32, 33, 18] and token-based [34, 35, 36, 37, 38, 39] video generation models have proven that the quality and controllability of video generation are steadily improving [40]. Notably, the introduction of the Sora model [14] 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 [41] 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 [1, 3, 5, 4, 6, 2] and agent-based environments [7, 8, 9, 10, 11, 12, 13]. Within this context, action-

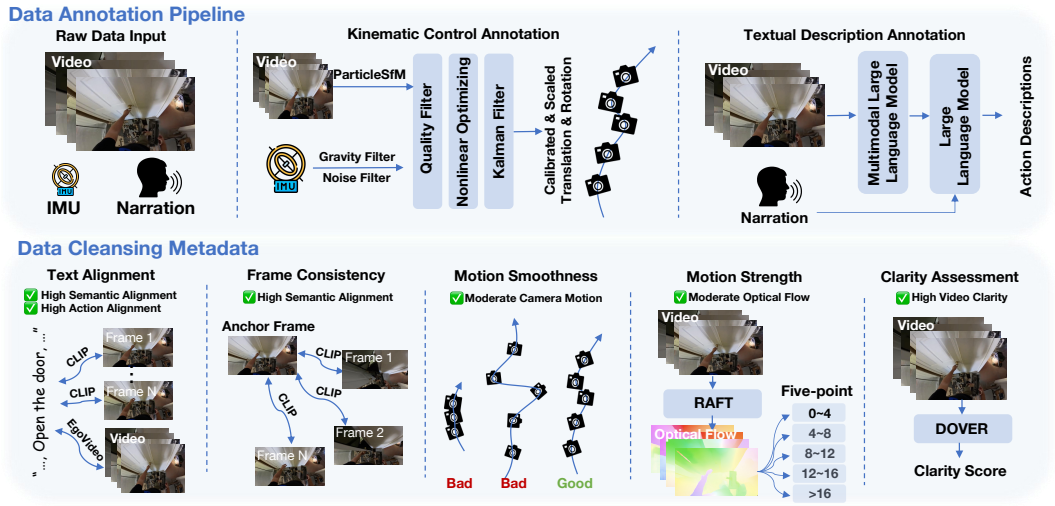

Figure 1: Data annotation pipeline and cleansing metadata of *EgoVid-5M*.

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 [42, 43, 44, 45, 46, 47, 48, 49, 50], and generative tasks associated with egocentric scenarios are largely confined to exocentric-to-egocentric video synthesis [51, 52, 53]. This highlights a substantial gap in generating action-driven egocentric videos. While some methods have explored video generation driven by action interaction [54, 55, 56, 57, 58, 59, 60, 61], 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 [20] consists of 52K hours of video, totaling 10.7M text-video pairs. Similarly, InternVid [62] contains over 7M videos spanning nearly 760K hours, with 4.1B words in descriptive texts. Panda70M [22] stands out with its collection of 70M high-resolution and semantically coherent video samples. OpenVid-1M [23], 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. UCF-101 [26] is an action recognition dataset featuring 101 classes and 13,320 total videos. Taichi-HD [63], a more focused collection, includes 2,668 videos capturing a single person performing Taichi. In egocentric video generation, existing datasets such as Ego4D [27] and Ego-Exo4D [28] are primarily designed for egocentric understanding tasks and often include excessive noisy camera motion, making them unsuitable for generative training. Additionally, EgoGen [64], a synthetic dataset, can not fully encapsulate the complex variations 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 [27]. 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

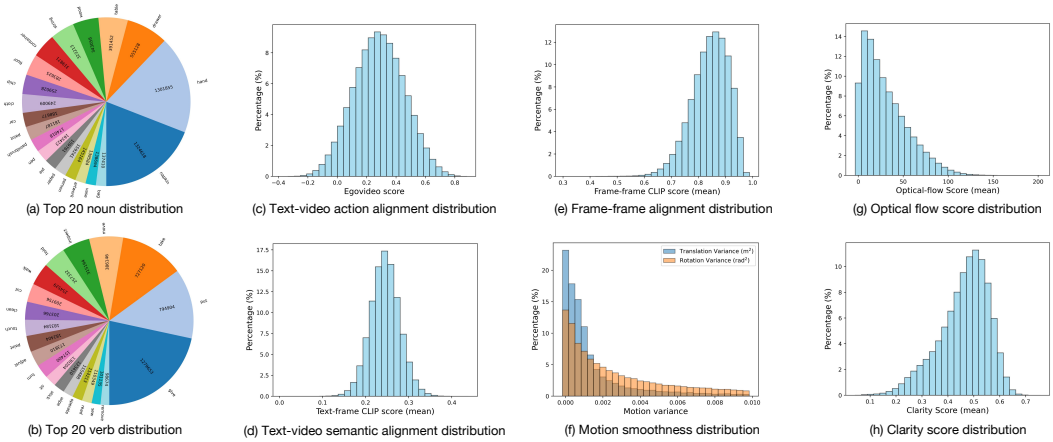

Figure 2: Data annotation distribution of *EgoVid-5M*. (a) and (b) describe the quantities of the top 20 verbs and nouns. (c) Text-video action alignment is assessed using the EgoVideo score. (d) and (e) measure the semantic similarity between text and frames and between frames and the first frame using 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.

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 Figure 1.

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 [65] to obtain scale-ambiguous camera poses P_c from video, followed 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 [56], 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 $\frac{1}{T}\sum_{t=0}^{T-1}(I_t - \overline{I})^2 \leq V_{\text{thres}}$. 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:

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

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

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

$$v(t+1) = v(t) + I(t)\Delta t, \tag{3}$$

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 [66] to provide detailed captions for videos (including foreground, background, main subjects, and action information). Then, we prompt Qwen2 [67] to summarize egocentric action descriptions from the aforementioned captions, with human narrations as the supplementary prompt. Through the combination of MLLM 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 Figure 2(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 [23, 16, 24] 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 Figure 1.

Text-video Consistency We utilize CLIP EgoVideo [47] and [68] 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 Figure 2(c) for Egovideo score distribution). Subsequently, these four frames are separately extracted to calculate CLIP similarity with the corresponding text (see Figure 2(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 Figure 2(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 Figure 2(f)).

Motion Strength A typical approach to describe video motion strength is optical flow [69]. Therefore, we first represent video motion by averaging global optical flow (see motion strength distribution in Figure 2(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 [70], we apply DOVER [71] to assess video clarity (refer to Figure 2 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, experiments are conducted in Sec. 5.2 to explore the effects of three mainstream

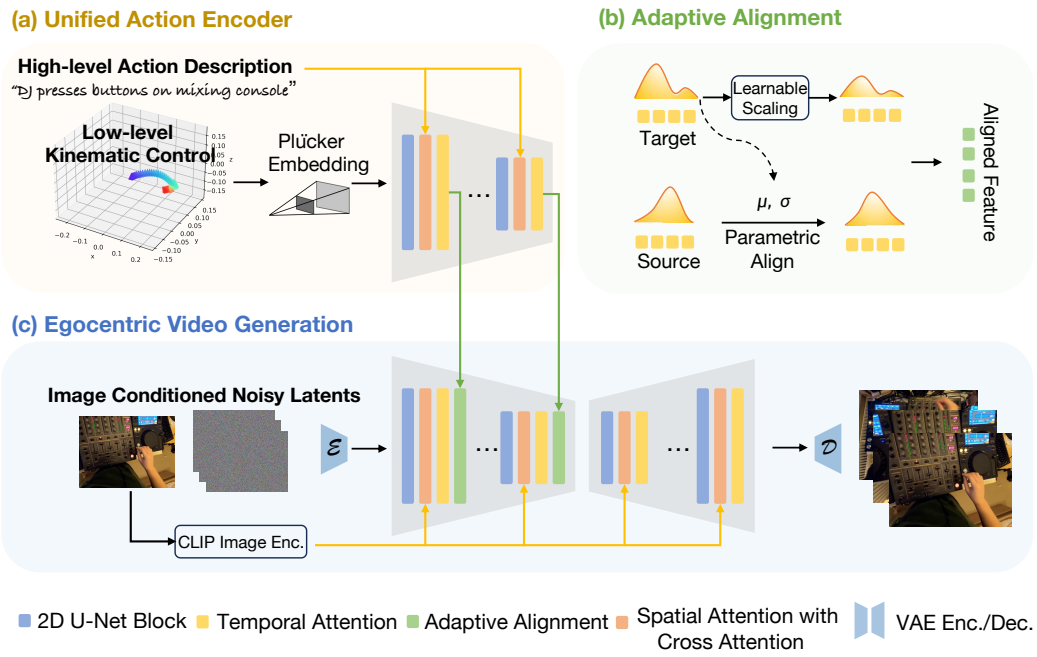

Figure 3: The overall framework of *EgoDreamer*. *EgoDreamer* introduces (a) the Unified Action Encoder to embed different action inputs, and it utilizes (b) the Adaptive Alignment to integrate action conditions into egocentric video generation branch (c).

cleaning strategies on egocentric video generation training. Additionally, given the significance of data cleaning strategies in training video generation models [23, 16, 24], 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 [54, 55, 56, 57, 58, 59, 60, 61] primarily focus on camera movements within static scenes, making it challenging to model complex ego-motion. Therefore, we propose *EgoDreamer*, which can produce egocentric videos driven simultaneously by high-level action descriptions and low-level kinematic control. As illustrated in Figure 3, *EgoDreamer* adopts a similar architecture of [17] 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 signals 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 [57, 72] to encode kinematic signals:

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

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

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 [68] and cross-attention mechanisms. The action output A of one U-Net block can be formulated as:

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

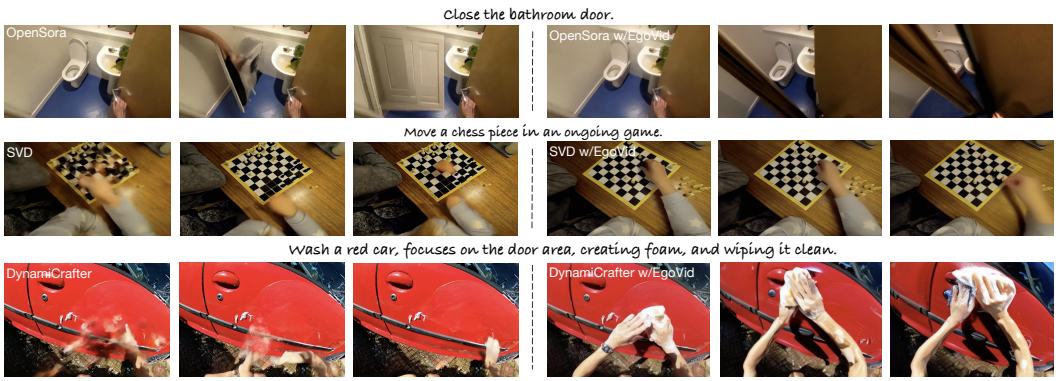

Figure 4: Visualizations demonstrate that *EgoVid*-fintuned baselines (OpenSora [33], SVD [16], DynamiCrafter [17]) generate egocentric videos with stronger frame-consistency and better semantic-alignment.

Method	w. *EgoVid*	CD-FVD ↓	Semantic Consistency ↑	Action Consistency ↑	Clarity Score ↑	Motion Smoothness ↑	Motion Strength ↑
SVD [16]	✗	591.61	0.258	0.465	0.479	0.971	18.897
SVD [16]	✓	**548.32**	**0.266**	**0.471**	**0.485**	**0.974**	**21.032**
DynamiCrafter [17]	✗	243.63	0.257	0.481	0.473	0.986	9.357
DynamiCrafter [17]	✓	**236.82**	**0.265**	**0.494**	**0.483**	**0.987**	**18.329**
OpenSora [33]	✗	809.46	0.260	0.489	0.520	0.983	7.608
OpenSora [33]	✓	**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.

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 [56, 57] 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).

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:

$$L_i = \alpha L_i + \frac{A_i - \mu_L}{\sigma_L}, \tag{7}$$

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 [73] and applies it to multi-scale U-Net feature alignment. Compared to ControlNet's zero-initialization [74], 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 annotation*, 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 [56], which is utilized further to train the ability of kinematic control video generation.

*https://ego4d-data.org/docs/data/imu/

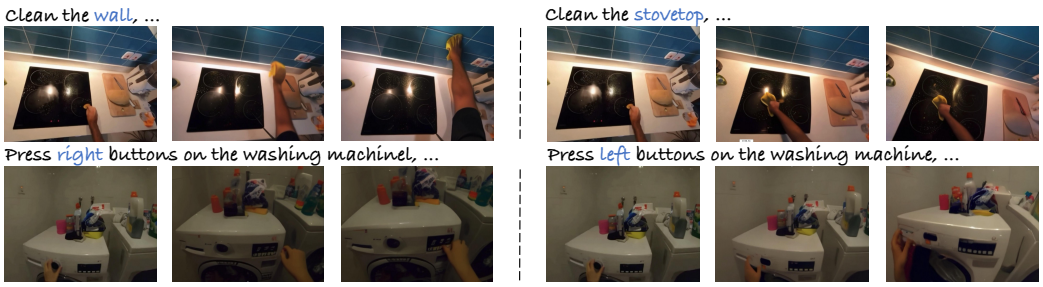

Figure 5: Visualizations show that *EgoDreamer* can realize distinct action controls based on different text descriptions.

Training We validate the effectiveness of our *EgoVid-5M* using video diffusion baselines with different architectures, including U-Net (SVD [16] and DynamiCrafter [17]), and DiT (OpenSora [33]). Building upon these pre-trained models, we employe a continuous training approach to train 480p videos for enhanced training efficiency. For *EgoDreamer*, we first initialize it with pre-trained weights [17], 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 [75] and VBench [76] to assess the quality of the generated egocentric videos. Specifically, our evaluation metrics utilize the CD-FVD [77] for spatial and temporal quality, the CLIP [68] for semantic consistency, the EgoVideo [47] for action consistency, the DOVER [71] for clarity score, frame interpolation model [78] for motion smoothness, and RAFT [69] for motion strength. Additionally, following [56, 57], 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 [56]. The specific calculations for each metric are detailed in the supplement.

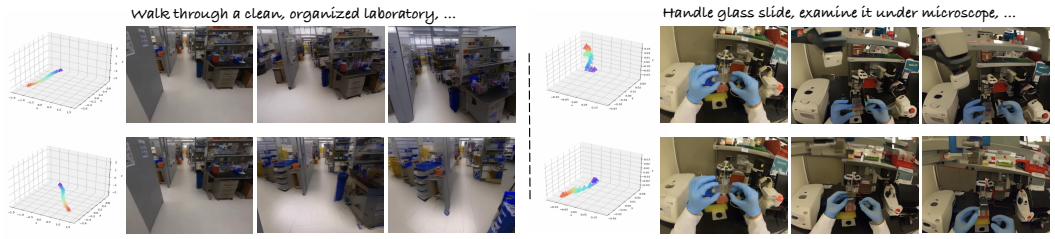

Figure 6: Visualizations demonstrate that *EgoDreamer* can generate various egocentric videos based on different low-level commands, where the left-side poses denotes the kinematic movements in 3D space (from blue to red).

w. EgoVid	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*.

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 [17] 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 the supplement, 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 shown in the supplement, 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-1M-3* subset. Compared to the previous two strategies, the model finetuned on *EgoVid-1M-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 visualizations in supplement, 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 [16], DynamiCrafter [17], and OpenSora [33] are selected as baselines, which are initialized with their original weights, and then we employ *EgoVid-1M-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 Table 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 Figure 4. 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 signals. Additionally, the efficacy of the proposed UAE and AA modules will be validated. In our experiments, we initialize *EgoDreamer* using weights from [17]. The results are presented in Table 3. In Row-1, the low-level kinematic control signals are integrated via ControlNet [77], which resembles [56, 57]. 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 [73] 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 Figure 5. Under initial frame conditions, varying the input text descriptions enables *EgoDreamer* to realize distinct action controls. Furthermore, as illustrated in Figure 6, 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 Conclusion and Limitations

Conclusion We present *EgoVid-5M*, the first large-scale, high-quality dataset tailored for egocentric video generation, containing 5 million clips with fine-grained action annotations. To ensure data quality, we introduce a robust cleaning pipeline that enforces temporal consistency and motion coherence. We also propose *EgoDreamer*, a generative framework that synthesizes egocentric videos conditioned on both action labels and kinematic signals. We hope *EgoVid-5M* will foster further research in egocentric video generation and benefit applications in VR, AR, and immersive simulation.

Limitations Despite its scale and quality, *EgoVid-5M* focuses primarily on short clips and pre-defined action categories, which may limit generalization to long-horizon tasks or open-ended activity understanding. Additionally, while *EgoDreamer* integrates kinematic control, its current form assumes pre-specified trajectories rather than learning them jointly, leaving room for future improvements in closed-loop control and interactive generation.

7 Acknowledgment

This work was supported by Alibaba Research Intern Program.

References

[1] 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *CoRL*, 2023.

[10] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

[11] Danijar Hafner and Timothy P. Lillicrap and Mohammad Norouzi and Jimmy Ba. Mastering atari with discrete world models. In *ICLR*, 2021.

[12] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[13] 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.

[14] 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.

[15] Anastasis Germanidis. Introducing general world models. 2023.

[16] 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.

[17] 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.

[18] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, April 2024.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan 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.

[24] 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.

[25] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015.

[26] 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.

[27] 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.

[28] 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, Abrham 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, Oluwatumininu 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen 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.

[41] Anastasis Germanidis. Introducing gen-3 alpha. 2024.

[42] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *ICCV*, 2023.

[43] 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.

[44] 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.

[45] 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.

[46] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *CVPR*, 2022.

[47] 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.

[48] 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. *NeurIPS*, 2022.

[49] 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.

[50] 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.

[51] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021.

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

[53] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *ACMMM*, 2021.

[54] 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.

[55] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

[56] Dejia 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.

[57] 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.

[58] 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.

[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*, 2024.

[60] 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.

[61] 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.

[62] Yi Wang, Yinan 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.

[63] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019.

[64] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. EgoGen: An Egocentric Synthetic Data Generator. In *CVPR*, 2024.

[65] 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.

[66] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024.

[67] 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.

[68] 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.

[69] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[70] Schuhmann Christoph. Improved-aesthetic-predictor, 2022.

[71] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023.

[72] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *NeurIPS*, 2021.

[73] 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.

[74] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[75] 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.

[76] 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.

[77] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[78] 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.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract accurately summarizes the paper's key contributions, including the introduction of a large-scale egocentric video dataset (*EgoVid-5M*) and a novel generation model (*EgoDreamer*). It also clearly outlines the challenges addressed and the scope of the work, aligning well with the content of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper acknowledges several limitations in Section 6, including the focus on short clips and pre-defined action categories in *EgoVid-5M*, which may constrain generalization. It also notes that *EgoDreamer* currently relies on pre-specified kinematic trajectories, highlighting the need for future work on joint trajectory learning and interactive generation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper primarily focuses on dataset construction and model design for egocentric video generation and does not present formal theoretical results or proofs. Therefore, this question is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of the dataset annotation protocols (Section 3.1), cleaning process (Section 3.2), model architecture (Section 4), training procedures, and evaluation metrics (Section 5.1). These details are sufficient to reproduce the main experimental results and validate the core claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset is already uploaded to `https://huggingface.co/datasets/Jeff-Wang/EgoVid-5M`, and the authors have provided detailed instructions in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 includes key training and testing details such as dataset splits, model architecture, hyperparameters, optimizer choice, and training schedules.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper focuses on qualitative results and standard quantitative metrics for video generation but does not emphasize statistical significance testing or report error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.2 specifies the computational resources used for data cleaning and annotation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper adheres to the NeurIPS Code of Ethics by responsibly collecting and processing data, clearly disclosing limitations, and aiming to advance research in a transparent and reproducible manner. No ethical concerns related to privacy, consent, misuse, or harm are identified in the presented work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the potential positive societal impacts of egocentric video generation in applications such as VR/AR and human-computer interaction. It also acknowledges possible negative implications, such as privacy concerns and misuse of realistic video synthesis.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper does not explicitly describe specific safeguards or measures to mitigate risks related to misuse or ethical concerns in the release of the dataset or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly credits all external datasets, codebases, and prior works utilized.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper thoroughly documents the newly introduced *EgoVid-5M* dataset, including its structure, annotation details, and data cleaning pipeline.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing experiments or human subject research, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects or direct participant interaction, so there are no risks to disclose or IRB approvals required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 3.1 describes the use of Vision Language Model (VLM) as part of data annotation.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.