

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

Anonymous Author(s)

Affiliation

Address

email

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

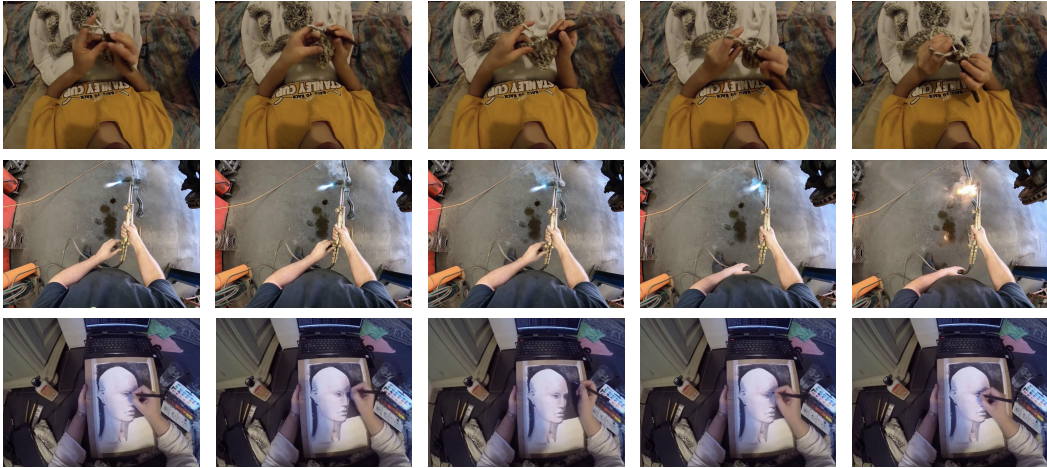


Figure 1: Videos cleaned from the *five-point* optical flow strategy (average optical flow below 3, and the proportion of optical flow (≥ 12 pixels) is greater than 3%). This strategy retains videos with a static background while capturing detailed and extensive motion in hands.

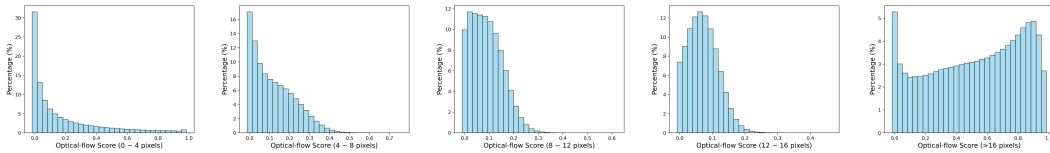


Figure 2: *Five-point* optical flow distribution.

4 A Annotation and Cleaning Details

5 A.1 Kinematic Annotation Details

- 6 To enhance kinematic annotation accuracy, we fuse camera poses from IMU and ParticleSfM [1],
- 7 utilizing the Kalman filter. First, we filter the IMU data to remove gravitational components and
- 8 noise. Next, we employ least squares estimation to determine the initial velocity and scale factor for

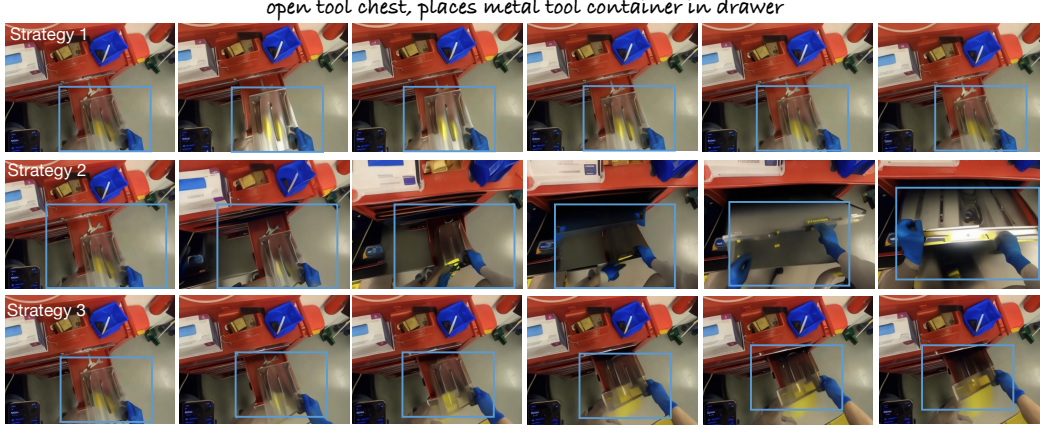


Figure 3: The video visualization comparison across different data cleaning strategies reveals distinct outcomes, where the blue box highlights the difference. Videos generated by strategy-1 fail to capture local motion and tend to be stationary. In contrast, videos produced by strategy-2 exhibit excessive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

9 the ParticleSfM poses. Finally, we align both the IMU poses and ParticleSfM results to the camera
 10 coordinate system (detailed explanations of these processes can be found in the main text). The
 11 Kalman filter implementation involves the following steps:

12 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
 13 position, quaternion, and velocity. The error covariance matrix \mathbf{P} , process noise covariance \mathbf{Q} and
 14 observation noise covariance \mathbf{R} 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
 15 prediction step, the state transition function \mathbf{f} is applied to predict the next state:

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

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

$$\mathbf{P}_{k|k-1} = \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T + \mathbf{Q}, \quad (2)$$

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

$$\mathbf{y}_k = \mathbf{x}'_k - \mathbf{H}\mathbf{x}_{k|k-1}, \quad (3)$$

21 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
 22 the observation model.

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

$$\mathbf{S}_k = \mathbf{H}\mathbf{P}_{k|k-1}\mathbf{H}^T + \mathbf{R}, \quad (4)$$

24 and the Kalman gain is calculated by:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}^T\mathbf{S}_k^{-1}. \quad (5)$$

25 The state estimate is then updated:

$$\mathbf{x}_k = \mathbf{x}_{k|k-1} + \mathbf{K}_k\mathbf{y}_k. \quad (6)$$

26 Finally, the error covariance matrix is updated:

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\mathbf{P}_{k|k-1}. \quad (7)$$

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

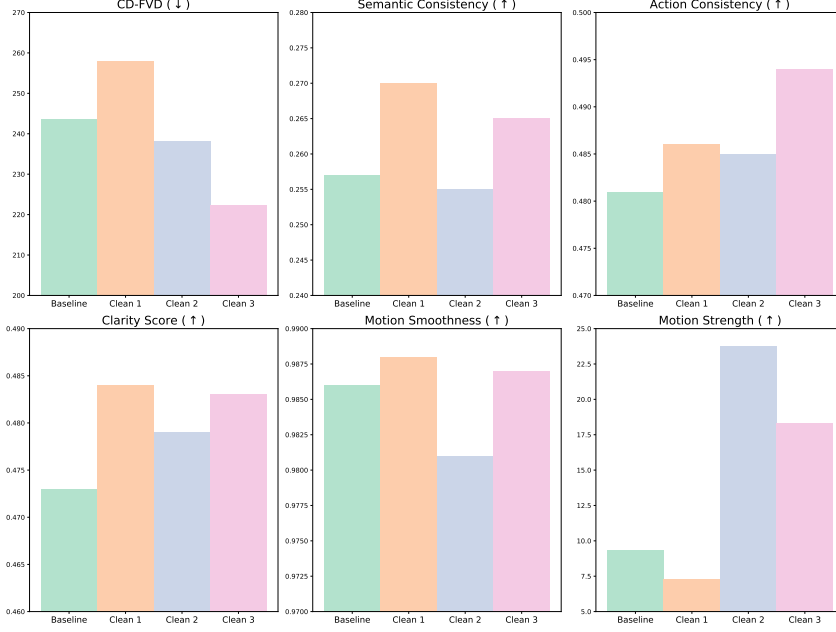


Figure 4: Video generation quantitative comparisons between different data cleaning strategies, where the baseline is DynamiCrafter [2] initialized with its original weights.

28 A.2 Data Cleaning Details

29 **Data Cleaning Strategy Comparison.** Strategy-1 enforces strict text-frame ($\text{CLIP}_{TF} \geq 0.275$)
30 and frame-frame ($\text{CLIP}_{FF} \geq 0.8$) consistency, retaining videos with average optical flow ≥ 3 and
31 DOVER score ≥ 0.3 , forming the subset *EgoVid-IM-1*. DynamiCrafter [2], finetuned for one epoch
32 on this subset, achieves the highest semantic consistency (Fig. 4). However, the strict filtering favors
33 slow-motion videos, leading to weak motion strength and suboptimal generation. Strategy-2 relaxes
34 the thresholds ($\text{CLIP}_{TF} \geq 0.27$, $\text{CLIP}_{FF} \geq 0.75$), and selects videos with optical flow between
35 3 and 40 and DOVER score ≥ 0.3 , yielding *EgoVid-IM-2*. Finetuning improves motion strength
36 but introduces artifacts and fragmentation, degrading text-frame consistency below the baseline.
37 Strategy-3 further lowers the thresholds ($\text{CLIP}_{TF} \geq 0.26$, $\text{CLIP}_{FF} \geq 0.7$) and adds an action
38 consistency constraint ($\text{EgoVideo score} \geq 0.22$). Videos with optical flow between 3 and 35 and
39 DOVER score ≥ 0.3 are retained, along with low-flow videos (flow < 3) if over 3% of pixels exceed
40 12-pixel motion. This yields *EgoVid-IM-3*, which achieves the best CD-FVD score by balancing
41 semantic, motion, and action consistency. The *5-point* optical flow filter emphasizes local motion,
42 capturing fine hand movements (Fig. 4) better than the static results of Strategy-1 and the exaggerated
43 motion in Strategy-2.

44 **Five-Point Optical Flow Filtering.** A typical approach to describe video motion strength is optical
45 flow [3]. Therefore, we first represent video motion by averaging global optical flow. Notably, this
46 approach only encapsulates the average motion magnitude. However, in egocentric scenarios, where
47 a substantial portion of the scene remains static and only foreground elements (e.g., hands) exhibit
48 motion, applying a filtering strategy based solely on average optical flow may result in the inadvertent
49 exclusion of valuable, fine-grained hand movement data. Therefore, as a supplement, we calculate
50 the *five-point* optical flow, which involves the proportion $P_{m \sim n}$ of optical flow score across different
51 pixel intervals:

$$P_{m \sim n} = \frac{\sum_{x,y} \delta(m \leq |F(x,y)| < n)}{N}, \quad (8)$$

52 where N is the total pixel number, F is the optical flow map, δ is the indicator function. Specifically,
53 we calculate $P_{0 \sim 4}$, $P_{4 \sim 8}$, $P_{8 \sim 12}$, $P_{12 \sim 16}$, $P_{16 \sim \infty}$, their distribution is shown in Fig 2. We performed
54 data filtering based on the *five-point* optical flow, as illustrated in Figure. 1, where the average optical
55 flow magnitude is less than 3 pixels, and over 3% of the pixels exhibit motion greater than 12 pixels.
56 The figure shows that although most of the background elements remain static, the hand movements

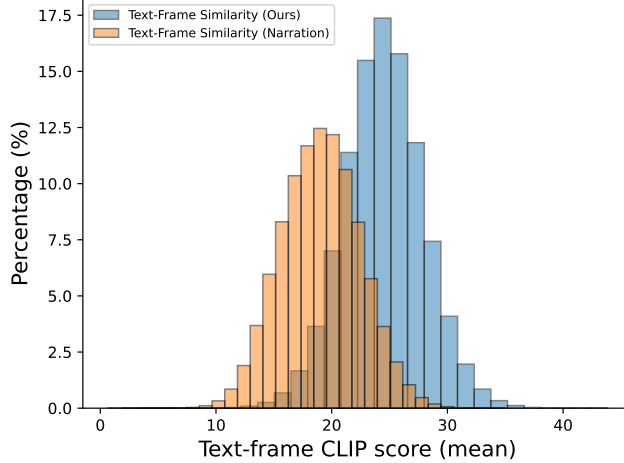


Figure 5: Semantic consistency comparison between our text annotation and the original human narration.

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) [4] to generate detailed captions for the videos. Then, a large language model (LLM) [5] is used to summarize egocentric action descriptions from these detailed captions. We calculate the semantic consistency between captions and the frames using CLIP [6]. As shown in Figure. 5, the semantic similarity of our generated captions is significantly higher than that of the original human narrations.

Visualization Comparison between Cleaning Strategies. As illustrated in Figure. 3, strategy-3 accurately models intricate hand movements, in contrast to the stationary visuals of strategy-1 and the exaggerated motion of strategy-2. Specifically, as indicated by the blue box, videos generated using strategy-1 often fail to capture local motion and appear static. Conversely, videos produced with strategy-2 show too much motion, which undermines semantic coherence. On the other hand, videos generated by strategy-3 successfully model complex hand movements, achieving a balance between motion intensity and semantic accuracy.

B Training Details

We validated the effectiveness of our *EgoVid-5M* using video diffusion baselines with different architectures, including U-Net (SVD [7] and DynamiCrafter [2]), and DiT (OpenSora [8]). The training details are as follows: (1) For SVD, we employ the pre-trained 1.1 version¹ and extend its *img-to-video* architecture to an *Image+Text-to-Video* setup. Specifically, we replace the image CLIP branch with a text CLIP branch², 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 [9] 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 resolution³. Videos are resized to 480p during training, utilizing the DDPM scheduler [10] 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 model⁴, 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 [11, 12] scheduler, a learning rate of $1e-4$, and a batch size of 64, using *EgoVid-1M-3* for one epoch.

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

²huggingface.co/openai/clip-vit-base-patch32

³huggingface.co/Doubiiu/DynamiCrafter_512

⁴huggingface.co/hpcai-tech/OpenSora-STDiT-v3



Figure 6: Visualizations showing that *EgoDreamer* can generate action-driven egocentric videos based on high-level text descriptions.

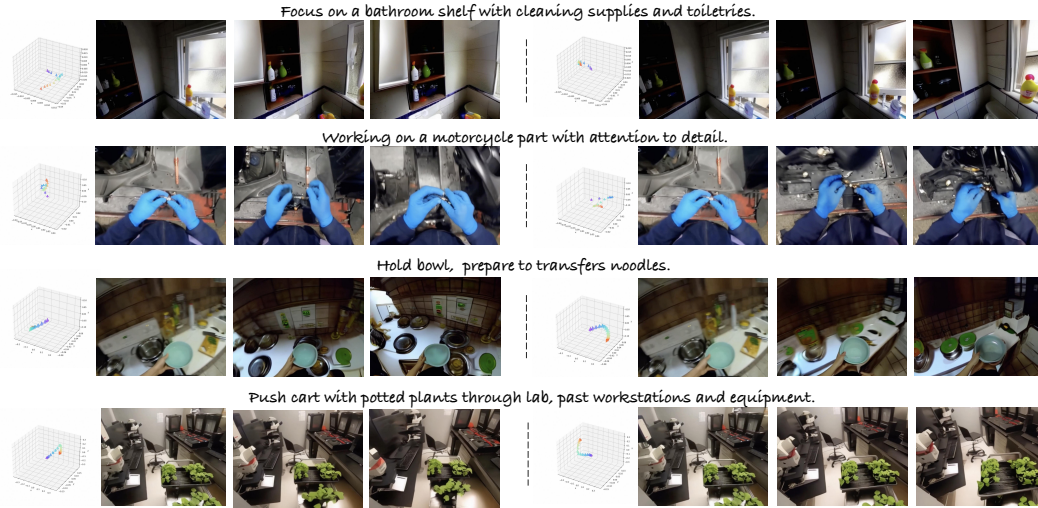


Figure 7: Visualizations showing that *EgoDreamer* can generate action-driven egocentric videos based on low-level kinematic control.

87 For *EgoDreamer*, we first initialize it with the pre-trained model at 512 resolution [2], then *Ego-*
88 *Dreamer* are further trained on *EgoVid-IM-3* to adapt to egocentric scenes, with batch size 64 and
89 learning rate 1e-5. Finally, we finetune the proposed Unified Action Encoder (UAE) and Adaptive
90 Alignment (AA) using *EgoVid-65K*, with batch size 32 and learning rate 1e-5.

91 C Evaluation Details

92 The evaluation metrics are mainly from AIGCBench [13] and VBench [14], along with other metrics
93 such as CD-FVD [15], EgoVideo score [16] and kinematic consistency (Translation Error and
94 Rotation Error) [17, 18]. These metrics are as follows:

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 8: Visualizations verifying that *EgoDreamer* can generate diverse egocentric videos based on action descriptions.

95 **Overall Quality.** CD-FVD⁵ is utilized to measure spatial and temporal quality. Compared with
96 traditional FVD [19], CD-FVD favors both quality and motion of video frames.

97 **Semantic Consistency.** CLIP⁶ [6] is employed to calculate the semantic consistency of text and
98 frames. We uniformly sample four frames from each generated video, calculate the similarity between
99 each frame and the text using CLIP, and then compute the average similarity score.

⁵github.com/songweige/content-debiased-fvd

⁶huggingface.co/openai/clip-vit-large-patch14

100 **Action Consistency.** EgoVideo⁷ [16] is utilized to calculate the action consistency of text and frames.
 101 In this metric, four frames are uniformly sampled from each video to calculate the action similarity
 102 between frames and text.

103 **Motion Strength.** We employed the optical flow score to quantify the motion strength in videos.
 104 Specifically, we utilized the RAFT model⁸ [3] to calculate the optical flow score. For each video, we
 105 sampled frames at 8-frame intervals as input to the model. The motion strength of the video segment
 106 was then determined by averaging the optical flow scores across all sampled frames.

107 **Motion Smoothness.** To assess the continuity of motion in the generated video, we utilize the AMT
 108 model⁹ [16]. Specifically, for a generated video with frames $[f_0, f_1, \dots, f_{2n-1}, f_{2n}]$, we remove the
 109 odd-numbered frames, resulting in $[f_0, f_2, \dots, f_{2n}]$. The AMT model is then employed to interpolate
 110 the omitted frames $[\hat{f}_1, \hat{f}_3, \dots, \hat{f}_{2n-1}]$. Finally, we compute the mean absolute error between the
 111 interpolated frames and the original ones.

112 **Clarity.** We leverage DOVER¹⁰ [20] to calculate the video clarity, and we use the fused score that
 113 focuses on both aesthetic perspective and technical perspective.

114 **Kinematic Consistency.** Following [18, 17], we assess kinematic consistency using translation error
 115 and rotation error, which measures the difference between COLMAP poses and the ground truth
 116 poses in the canonical space:

$$\text{RotErr} = \sum_{i=1}^n \arccos \frac{\text{tr}(\mathbf{R}_{\text{gen}}^i \mathbf{R}_{\text{gt}}^{i\top}) - 1}{2}, \quad (9)$$

117

$$\text{TransErr} = \sum_{i=1}^n \|\mathbf{T}_{\text{gt}}^i - \mathbf{T}_{\text{gen}}^i\|_2, \quad (10)$$

118 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
 119 translation vectors for the generated and ground truth camera translation in the i -th frame.

120 D Visualizations

121 We conducted additional visualizations of the results generated by EgoDreamer. As shown in Figure. 8,
 122 *EgoDreamer* can leverage action descriptions to generate diverse egocentric videos, encompassing
 123 scenes such as householding, cooking, knitting, gardening, and music. These videos include both
 124 subtle hand movements and more extensive movements involving walking. Furthermore, as illustrated
 125 in Figure. 6, given the same initial frame, changing the high-level text descriptions can generate
 126 egocentric videos that comply with semantic control. Lastly, as depicted in Figure. 7, given the same
 127 initial frame, altering the low-level kinematic control can generate egocentric videos that conform to
 128 pose control.

⁷drive.google.com/file/d/1k6f1eRdcL17IvXtdX_J8WxNbju2Ms3AW/view

⁸github.com/princeton-vl/RAFT

⁹huggingface.co/lalala125/AMT/resolve/main/amt-s.pth

¹⁰huggingface.co/teowu/DOVER/resolve/main/DOVER.pth

References

- [1] 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.
- [2] 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.
- [3] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.
- [11] 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.
- [12] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.

- 176 [17] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan
177 Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint*
178 *arXiv:2404.02101*, 2024.
- 179 [18] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vah-
180 dat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint*
181 *arXiv:2406.02509*, 2024.
- 182 [19] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski,
183 and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges.
184 *arXiv preprint arXiv:1812.01717*, 2018.
- 185 [20] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang,
186 Wenxiu Sun Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user
187 generated contents from aesthetic and technical perspectives. In *ICCV*, 2023.