EgoM2P: Egocentric Multimodal Multitask Pretraining

Gen Li[1] Yutong Chen[*1] Yiqian Wu[*1,2] Kaifeng Zhao[*1] Marc Pollefeys[1,3] Siyu Tang[1]

[1]ETH Zürich [2]Zhejiang University [3]Microsoft

https://egom2p.github.io/

Abstract

Understanding multimodal signals in egocentric vision, such as RGB video, depth, camera poses, and gaze, is essential for applications in augmented reality, robotics, and human-computer interaction, enabling systems to better interpret the camera wearer's actions, intentions, and surrounding environment. However, building large-scale egocentric multimodal and multitask models presents unique challenges. Egocentric data are inherently heterogeneous, with large variations in modality coverage across devices and settings. Generating pseudo-labels for missing modalities, such as gaze or head-mounted camera trajectories, is often infeasible, making standard supervised learning approaches difficult to scale. Furthermore, dynamic camera motion and the complex temporal and spatial structure of first-person video pose additional challenges for the direct application of existing multimodal foundation models.

To address these challenges, we introduce a set of efficient temporal tokenizers and propose EgoM2P, a masked modeling framework that learns from temporally-aware multimodal tokens to train a large, general-purpose model for egocentric 4D understanding. This unified design supports multitasking across diverse egocentric perception and synthesis tasks, including gaze prediction, egocentric camera tracking, and monocular depth estimation from egocentric video, and also serves as a generative model for conditional egocentric video synthesis. Across these tasks, EgoM2P matches or outperforms specialist models while being an order of magnitude faster. We will fully open-source EgoM2P to support the community and advance egocentric vision research.

1. Introduction

Egocentric video capture has evolved significantly with the integration of multimodal data, including RGB, depth, gaze, and camera trajectories. These modalities interact dynamically, offering the most crucial information for understanding human head motion, intention, and scene geometry. The growing availability of real-world multimodal egocentric datasets [5, 10, 20, 23, 26, 29, 30, 49, 60, 61, 69, 70, 77, 87, 106, 122, 125], such as EgoExo4D, HoloAssist, and HOT3D, provide rich, diverse, and semantically meaningful data. Additionally, large-scale synthetic datasets generated by simulators like EgoGen [52] provide precise ground truth annotations, which are often expensive and time-consuming to obtain in the real world. By combining these complementary data sources, it is increasingly feasible to train large-scale multimodal and multitask models for egocentric vision. These models have the potential to further enhance our understanding of human behavior and real-world interactions, opening new possibilities for applications in augmented reality, virtual reality, and robotics.

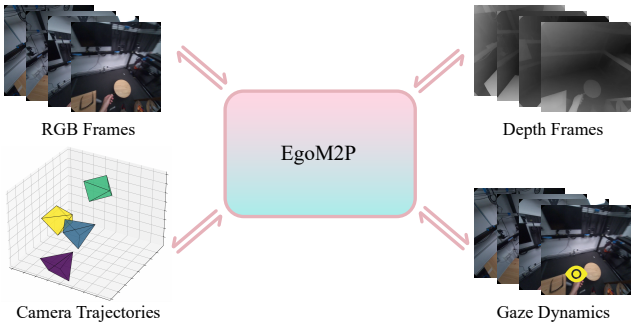


Figure 1. *EgoM2P*: A large-scale egocentric multimodal and multitask model, pretrained on eight extensive egocentric datasets. It incorporates four modalities—RGB and depth video, gaze dynamics, and camera trajectories—to handle challenging tasks like monocular egocentric depth estimation, camera tracking, gaze estimation, and conditional egocentric video synthesis. For simplicity, we only visualize four frames here.

Current large-scale video models predominantly focus on understanding and generating videos from third-person perspectives: Video understanding tasks span video captioning, question answering, retrieval, and segmentation [39, 51, 59, 71, 112, 116, 121, 123, 127]; Video generation models primarily generate third-person view videos from text inputs [11, 12, 17, 22, 31, 35, 37, 45, 67, 68, 78, 85, 88, 100, 107, 111, 117]. Recent works [48, 81] have expanded modalities to include audio, yet these models still operate with a limited range of modalities. While third-person video models have seen significant progress, *ego-*

[*]Equal contribution. Alphabetical order.

2025 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3315-8775-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCV51701.2025.01008

2380-7504/25/$31.00 ©2025 IEEE
DOI 10.1109/ICCV51701.2025.01008

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

centric foundation models have advanced at a much slower pace. Egocentric views present unique challenges, including dynamic camera movements, complex human-object interactions, and the influence of human intentions, which are not adequately addressed by third-person models. Furthermore, the development of large egocentric models was hindered by the limited scale of available egocentric datasets. With the scaling of egocentric datasets, recent efforts in egocentric foundation models [90, 118] have focused on egocentric video question-answering (QA). However, these models often overlook critical human-related modalities, which are naturally captured by head-mounted cameras and are essential for understanding human intentions. Additionally, these models lack 3D or 4D reconstruction capabilities, limiting their ability to fully capture the spatiotemporal aspects of human motions and intentions.

Recent advances in multimodal and multitask vision foundation models primarily center on images, demonstrating remarkable abilities in cross-modal prediction [8, 18, 47, 64, 65, 76, 103, 105]. These models, which build upon Transformers [99], enable versatile any-to-any predictions, facilitating multitasking across various modalities, such as depth, surface normal, segmentation masks, etc. However, these models use pseudo-labeling networks to generate aligned binding data across modalities, but effective pseudo-labelers for egocentric videos remain limited due to the domain gap with third-person views. Moreover, these models focus on single-image prediction, and struggle to maintain temporal consistency when applied to egocentric video sequences with fast-changing camera poses.

In this paper, we present *EgoM2P*, the first multimodal and multitask model for 4D egocentric data. Our approach explores four modalities: RGB video, depth video, gaze dynamics, and camera trajectories. *EgoM2P* supports any-to-any modality predictions and demonstrates its multitasking capabilities across various tasks: gaze estimation, egocentric camera tracking, depth estimation from monocular egocentric video, and conditional egocentric video generation.

Specifically, we build a multimodal token database containing four billion training tokens by curating eight extensive egocentric datasets from both real-world and synthetic data. To address missing modality annotations, we effectively extend multimodal masked pretraining, originally designed for image foundation models that assume the availability of all modalities, to the egocentric video domain. While missing modalities are masked out, *EgoM2P* can still effectively predict them. We design a unified temporal tokenizer architecture to tokenize multimodal data into temporally-aware tokens. By using variable masking rates to mask input and target tokens during the training of *EgoM2P*, we benefit from its parallel inference capability and demonstrate its multitasking performance across various downstream applications, while achieving a significant speed-up. In summary, the contributions of this work are:

1. We introduce *EgoM2P*, the first multimodal and multitask large egocentric model for RGB, depth video, eye gaze dynamics, and camera trajectories.
2. We extend multimodal masked pretraining from the image domain to the egocentric video domain by addressing challenges such as more complex spatiotemporal dynamics and the lack of annotations for certain inherently missing modalities.
3. *EgoM2P* is comparable to or outperforms state-of-the-art algorithms in egocentric camera tracking, gaze dynamics estimation, monocular egocentric depth estimation, and conditional egocentric video synthesis, while being significantly more efficient.

2. Related Work

Image Foundation Models. Image foundation models are pretrained versatile neural networks that serve as a universal foundation of various downstream vision tasks, such as image classification, detection, and segmentation [15, 16, 21, 46, 62, 79]. CLIP [82] aligns image and text embeddings via contrastive learning, unlocking zero-shot and open-vocabulary classification. ImageBind [27] extends multimodal alignment beyond text, connecting images to modalities like audio, depth, and thermal data. Multimodal Language Models [1, 2, 41, 57, 72, 92, 113] enable unified reasoning across text, images, and audio. Recent work 4M [8, 76] trains a multimodal image foundation model using Transformers to enable prediction across any input-output modality pairs. Our work extends 4M by incorporating temporal modeling, training a unified multimodal and multitask model for egocentric vision.

Video Foundation Models. Recent advancements in video foundation models build upon the success of Vision Language Models (VLMs), focusing on video understanding and generation through various approaches, such as video-language contrastive learning [109, 110, 114], masked modeling [25, 95, 102], and autoregressive sequence prediction [4, 53, 55, 71, 91]. Diffusion-based video generative models [3, 11, 78, 81, 93] achieve photorealistic video generation with fine-grained content control through conditioning signals such as text prompts and reference images. These capabilities position video models as promising candidates for world models [3, 14, 32, 74, 97], as their generative process inherently captures the temporal dynamics of real worlds from internet-scale data.

Egocentric Video Understanding and Generation. Understanding the world through egocentric videos is critical for applications in augmented reality, virtual reality, and robotics. Multiple egocentric video datasets [10, 20, 29, 30, 106] have been collected to capture the diversity and complexity of daily life scenarios. These egocentric videos present significant technical challenges, including: activity

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

Datasets \ Modalities	RGB	Depth	Gaze	Camera
EgoExo4D [30]	✓	×	✓	✓
HoloAssist [106]	✓	✓*	✓	✓
HOT3D (Aria) [10]	✓	✓*	✓	✓
HOT3D (Quest) [10]	gray	✓*	×	✓
ARCTIC [23]	✓	✓*	×	✓
TACO [61]	✓	✓*	×	✓
H2O [49]	✓	✓	×	✓
EgoGen [52]	✓	✓	×	✓

Table 1. **Datasets used in our method.** A green checkmark (✓) signifies availability, a red cross (×) denotes unavailability or exclusion due to low quality, and a blue checkmark with a star (✓*) indicates the use of pseudo labels.

recognition [42, 101, 129], hand motion and object interaction estimation [9, 24, 119], egocentric video prediction and generation [56, 58, 108, 115], egocentric camera localization [86, 94], among others. While task-specific methods have been developed for these challenges, there remains a lack of a unified egocentric video foundation model. Our work addresses this gap by enabling cross-modality sequence predictions across RGB, depth, camera poses, and gaze signals, representing a preliminary step toward a foundational multitasking model for a unified egocentric understanding of scenes and human behaviors.

3. Method

This section overviews the data curation pipeline and training paradigm. In Sec. 3.1, heterogeneous egocentric datasets are transformed into unified formats. Sec. 3.2 describes the process of compressing high-dimensional multimodal data into compact discrete tokens, enabling efficient training and inference. Next, Sec. 3.3 covers the embedding of multimodal discrete tokens and the masked pretraining of *EgoM2P*. Finally, Sec. 3.4 explains how final target tokens are predicted by sampling from the pretrained model.

3.1. Data Curation

As shown in Tab. 1, egocentric datasets differ in their data annotation coverage. Due to hardware and processing constraints, frame drops are common in captured depth data, making it difficult for wearable AR glasses [73, 75] to achieve pixel-aligned depth streams. Besides, helmet-mounted Kinect cameras are unable to capture gaze information. This unstructured nature of egocentric data poses challenges for both data curation and model training.

Our data curation pipeline includes 3 steps: 1) splitting, 2) annotation, and 3) standardization:

Splitting. The raw multimodal data are segmented into clips of T frames. These video clips are re-encoded into high-quality mp4 format with the same resolutions.

Annotation. Real-world egocentric depth data is scarce, often contains sensor noise, and suffers from frame drops. We leverage RollingDepth [44] to generate pseudo-labels for the depth annotation to get pixel-aligned depth videos. To further scale up the amount of accurate depth training data, we use EgoGen [52], a novel egocentric synthetic data generator, to generate approximately 30 hours of video data at 30 FPS. This involves letting virtual humans walk in Replica [89] scenes and GIMO [128] scene scans, rendering their egocentric views to obtain accurate depth and camera trajectory annotations. As analyzed in Supp. Mat. Sec. B.3, EgoGen boosts *EgoM2P* performance. For datasets without gaze annotations, we leave them unlabeled due to the lack of effective pseudo-labelers for gaze dynamics.

Standardization. These datasets have various video resolutions and define their world coordinate systems differently, with varying origins, axis conventions, handedness, and scale. We standardize multi-modal data as follows. All data streams are at 30 FPS. Depth videos are encoded using inverse depth representation, with normalization applied per sequence. Noisy Kinect depth labels are preprocessed with hole filling via morphological operations and inpainting, then denoised using median and bilateral filtering. We reproject the eye gaze data, originally represented as a 3D ray with depth, onto the 2D image plane and represent it as a moving point on this plane. Camera trajectories are unified as camera-to-world transformations in OpenCV convention, using the first frame as the reference frame to standardize the world coordinates across different datasets.

3.2. Tokenizers

Given a multi-modal clip with T frames, we represent each modality as follows:

- **RGB Video**: $\mathbf{X}^{\text{rgb}} \in \mathbb{R}^{T\times H\times W\times 3}$, where H and W denote the spatial resolution.
- **Depth Video**: $\mathbf{X}^{\text{depth}} \in \mathbb{R}^{T\times H\times W\times 1}$.
- **Gaze Dynamics**: $\mathbf{X}^{\text{gaze}} \in \mathbb{R}^{T\times 2}$, where each entry corresponds to the 2D gaze coordinates.
- **Camera Trajectory**: $\mathbf{X}^{\text{cam}} \in \mathbb{R}^{T\times 9}$, where each pose is parameterized by the 6D rotation representation [130] and translation, with the reference frame as the first frame.

Our multi-modal dataset is represented as $\mathbf{X} = \{\mathbf{X}^{\text{rgb}}, \mathbf{X}^{\text{depth}}, \mathbf{X}^{\text{gaze}}, \mathbf{X}^{\text{cam}}\}$, where modalities can be missing for each sub-dataset according to Tab. 1.

We leverage the state-of-the-art (SOTA) Cosmos tokenizer [3] to tokenize video modalities, applying a temporal compression rate of 4 and a spatial compression rate of 8 to convert the video into discrete tokens. For other modalities, we train modality-specific tokenizers. To ensure adaptability when incorporating new modalities, we employ a unified tokenizer architecture using a Transformer-based vector quantized variational autoencoder (VQ-VAE) [98]. The network architecture is illustrated in the upper part of

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

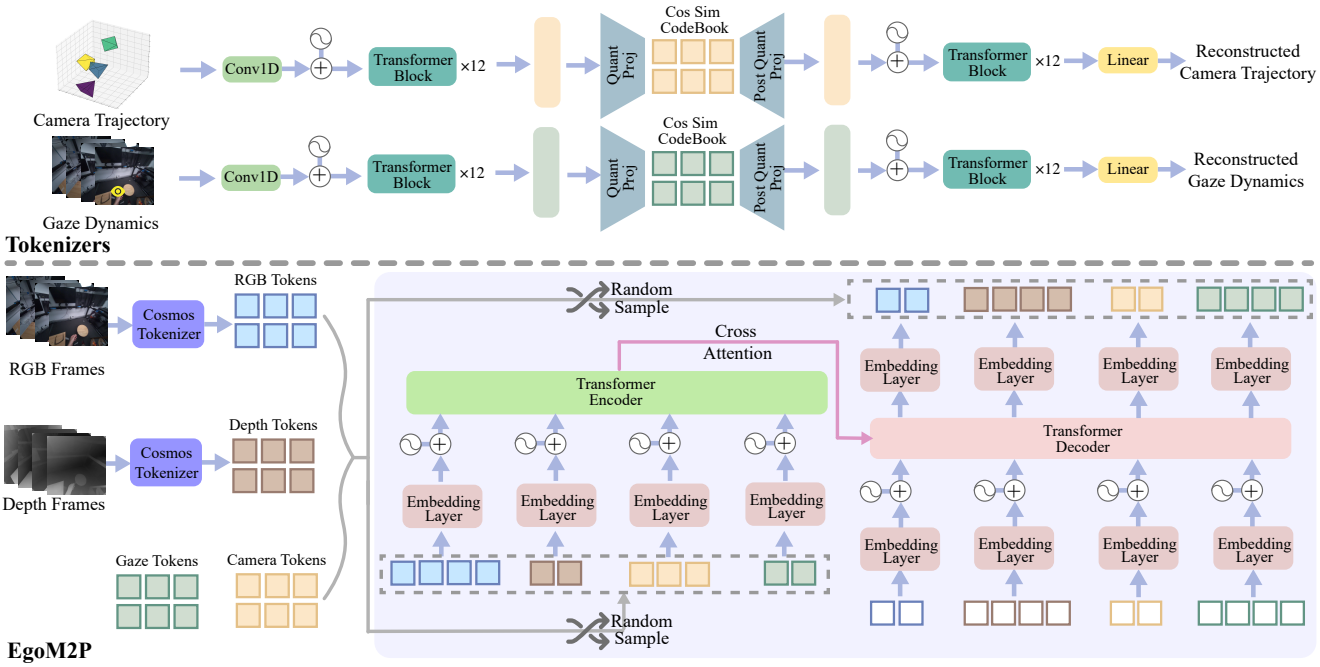


Figure 2. **Network Architecture:** (1) We train VQ-VAE [98] tokenizers for camera trajectories and gaze dynamics (Sec. 3.2), and adopt Cosmos tokenizers [3] to tokenize RGB and depth streams. High-dimensional input modalities, including videos, gaze dynamics, and camera trajectories, are compressed into discrete tokens to serve as our training database. (2) Our *EgoM2P* follows the architecture of T5-Base [83]. We perform multimodal masked pretraining (Sec. 3.3), where we randomly sample a fixed number of input and target tokens from our token database without overlap. For simplicity, we only visualize four frames here.

Fig. 2. For each modality $mod \in \{\text{gaze}, \text{cam}\}$, let $N = \dim(\mathbf{X}^{mod}) - 1$. The encoder $\mathcal{E}$ begins by performing an N-dimensional convolution, represented as $\text{Conv}_N\text{d}(\mathbf{X}^{mod})$. This operation downsamples along the temporal axis by a factor of 2. Next, it adds an N-dimensional positional embedding before passing the embedded data through 12 Transformer blocks. Each Transformer block attends to every pair-wise interaction among all tokens. Then, we quantize the embeddings using the quantizer $\mathcal{Q}$ to learn modality-specific codebooks, employing cosine similarity as the distance metric, following [76, 120]. Finally, discrete codes are decoded through the decoder $\mathcal{D}$, which mirrors the architecture of the encoder:

$$\mathbf{z} = \mathcal{E}(\mathbf{X}^{mod})$$
$$\mathbf{q} = \mathcal{Q}(\mathbf{z})$$
$$\hat{\mathbf{X}}^{mod} = \mathcal{D}(\mathbf{q})$$

The overall training loss is:

$$\mathcal{L}_{tok} = \|\mathbf{X}^{mod} - \hat{\mathbf{X}}^{mod}\|_2^2 + \|\,\text{sg}[\mathbf{z}] - \mathbf{q}\|_2^2 + \beta\,\|\mathbf{z} - \text{sg}[\mathbf{q}]\|_2^2,$$

where $\text{sg}[\cdot]$ is the stop-gradient operator and β balances the commitment loss. Due to hardware constraints, certain modalities, such as gaze, may contain invalid data when tracking is lost. We mask out invalid data and use masked L2 loss instead. See Supp. Mat. Sec. A.1 and C. for details.

3.3. Multimodal Masked Pretraining

4M [8, 76] introduces the Massively Multimodal Masked Modeling training scheme for static image modalities, where a small batch of sampled multimodal target tokens is predicted using another batch of sampled multimodal input tokens. However, when applying masked modeling to egocentric videos, 4M encounters several challenges: 1) There are no mechanisms to ensure temporal consistency. 2) The number of tokens per video sample is significantly larger than in image modalities, hindering efficient training and scalability. 3) The ratio of tokens across different modalities is highly imbalanced; *e.g.*, each sample contains 170 times more video tokens than gaze tokens, which may lead to the neglect of critical information from less represented modalities. 4) Missing annotations in egocentric datasets can pose challenges, whereas 4M pseudo-labeled all modalities.

Multimodal Token and Dataset Balancing. In Sec. 3.2, we leverage our Transformer-based VQ-VAE and Cosmos Tokenizer [3] to tokenize each modality into temporally-aware discrete tokens. The large number of video tokens poses challenges for multimodal token pretraining. While aggressively compressing video tokens during tokenization might be beneficial, it can negatively impact video quality. To address this, we downsample videos to 8 FPS, reducing the token count per video to 1/3 of its original amount.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

After tokenization, the training set comprises roughly 4 billion multimodal tokens, whereas there are just 13 million gaze tokens. In addition, the scale of different datasets is also highly imbalanced, *e.g.*, EgoExo4D [30] has 160 times more samples than H2O [49]. Training directly with these imbalanced datasets can cause the model to ignore modalities that have fewer tokens. Additionally, some datasets might suffer from overfitting, while others remain underfitted. To mitigate this issue, we experiment with different sampling weights for both dataset sampling and token sampling across multi-modalities within the datasets. We discover that initially sampling a dataset with a probability proportional to its size, followed by sampling tokens from its modalities with uniform concentration parameters following 4M, results in the most stable training and optimal performance. See Supp. Mat. Sec. A.2 for details.

Temporal Multimodal Token Embedding. As shown in the lower part of Fig. 2, for each modality, we use modality-specific embedding layers to map input tokens into a high-dimensional unified space, facilitating the alignment and integration of multimodal information. We then add sine-cosine positional embeddings, using 1D for gaze and camera tokens and 3D for video tokens. The same approach is applied to target tokens. Similar to 4M [8, 76], we incorporate a learnable modality category embedding, which is shared for both input and target token embedding modules.

Masking. Masked modeling has demonstrated its efficacy in prior works [7, 8, 33, 76, 95]. 4M requires aligned multimodal annotations and resorts to pseudo-labeling. However, it is not practical to pseudo-label all modalities for egocentric datasets. We represent missing modalities as placeholders and mask them out. See Supp. Mat. Sec. A.4 for details on handling missing modalities. Similar to 4M, we reduce computational costs by applying input and target masking, encoding and decoding only a fixed number of visible tokens. While 4M caps this at 256, given that the number of tokens per video exceeds 5000, we increase this number to 2048 to accommodate more information. For each multimodal data sample $\mathbf{X}_i$, we sample how many visible tokens to use as inputs and targets for each available modality *mod* $\in$ {rgb, depth, gaze, cam}, then sample tokens in each clip $\mathbf{X}_i^{mod}$ within these limits accordingly (See Supp. Sec. A.2). Visible input and target tokens are mutually exclusive.

Model architecture. Apart from the modality-specific embedding layers for the input and target tokens, the main architecture follows T5-Base [83]. The encoder applies self-attention to all sampled visible input tokens, integrating spatiotemporal information from multiple modalities. The decoder input is formed by masking the sampled visible target tokens. The decoder applies cross-attention with the encoder output as context and performs masked self-attention—restricted to tokens within the same modality—to predict those masked target tokens, which ensures that the decoder generates coherent tokens within the same modality. The training loss is cross-entropy. For each modality m:

$$\mathcal{L}_m = -\frac{1}{T_m}\sum_{t=1}^{T_m}\sum_{c=1}^{C_m} y_{t,c}^{(m)} \log\left(\hat{y}_{t,c}^{(m)}\right)$$

$$\mathcal{L}_{\text{total}} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathcal{L}_m$$

where $\mathcal{M}$ is the set of modalities that have available tokens, T_m is the number of tokens in each modality, C_m is the codebook size, and $\hat{y}_{t,c}$ is the predicted probability distribution for each token. See Supp. Mat. Sec. A.3 for details.

3.4. Inference

During training, we use variable masking rates to randomly mask out multimodal tokens that are encoded with temporal information. Prior works [76] have shown that masked image models trained with this scheme function as order-agnostic autoregressive models [38], allowing tokens to be decoded iteratively in random orders for parallel inference. Similarly, we show that *EgoM2P* is able to predict the distribution over masked tokens simultaneously, potentially providing the speed needed for real-time applications. As shown in Tab. 2, our method can predict the camera trajectory for a 60-frame video in 0.18 seconds (300+ FPS).

The parallel inference process can be formulated as a multi-step decoding procedure. We use a linear scheduling approach to predict n target tokens over s decoding steps. At each step, we randomly select n/s target tokens and first perform a forward pass of the pretrained model using all visible input tokens to predict the conditional distribution $\hat{y}_{\text{cond}}$ of selected target tokens in parallel. Next, we mask out all input tokens and perform a second forward pass to predict the unconditional probability $\hat{y}_{\text{uncond}}$ of selected target tokens using cross-attention on masked input tokens. The predicted target token distribution $\hat{y}$ is estimated using classifier-free guidance [34] with weight ω:

$$\hat{y} = (1+\omega)\hat{y}_{\text{cond}} - \omega\hat{y}_{\text{uncond}}$$

The final target token prediction is sampled from the predicted distribution $\hat{y}$ with nucleus sampling [36]. We find that for modalities with a small number of tokens, a single decoding step is sufficient. For video modalities, increasing decoding steps and predicting a subset of tokens at each step is beneficial. See Supp. Mat. Sec. A.5 for pseudocode.

3.5. Implementation Details

For each multimodal clip, we set its length to 2 seconds. Non-video modalities have $T = 60$ frames, while video modalities have $T = 16$ frames due to the reduced FPS. The video resolution is 256×256. After tokenization, the RGB and depth videos have 5120 tokens per sample, and the gaze and camera trajectory have 30 tokens per sample. See more details in Supp. Mat. Sec. A.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

Method	EgoExo4D [30]			ADT [80] (*unseen*)			Time↓
	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓	
DROID-SLAM [94]	0.018	0.005	0.506	0.034	0.010	0.316	2.7s
ACE-Zero [13]	0.028	0.007	0.672	0.049	0.011	0.333	426s
Align3R [66]	0.019	0.006	0.762	**0.028**	0.010	**0.276**	372s
EgoM2P	**0.017**	**0.004**	**0.429**	0.032 **0.026**	**0.006** **0.005**	0.490 0.480	**0.18s**

Table 2. **Evaluation on camera tracking.** Compared to specialist SOTAs that require geometry test-time optimization, *EgoM2P*'s feed-forward tracking results achieve comparable performance yet with significantly higher efficiency. We report the average runtime per sequence. Underlined denotes post-training results (Sec. 4.5).

4. Experiments

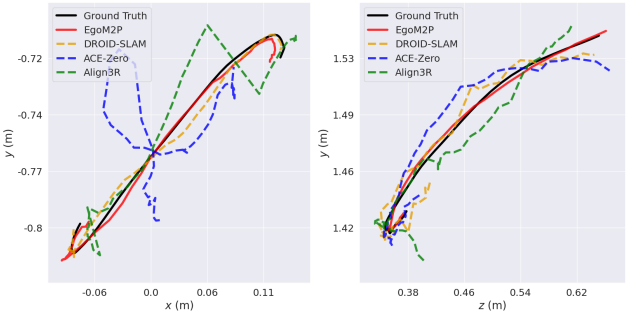
We benchmark *EgoM2P*'s multitasking abilities with SOTA models in downstream tasks, including egocentric perception and synthesis. We also benchmark it on unseen datasets without any fine-tuning to show the strong generalization ability of the pretrained feature. Additionally, we show *EgoM2P* can be easily fine-tuned via post-training.

4.1. Egocentric Camera Tracking

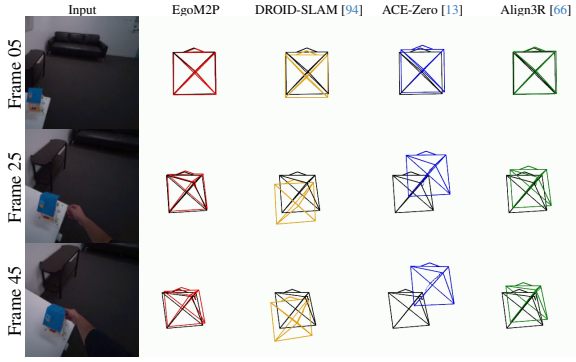
Evaluation Protocols. We sample 200 video clips from the validation split of the EgoExo4D dataset [30] for evaluation. To assess *EgoM2P*'s generalization to unseen dataset, we also evaluate on 200 video clips from the Aria Digital Twin (ADT) dataset [80], which is entirely excluded from our training dataset. We process all input frames to 256×256 resolution and 8 FPS. We use standard error metrics: Absolute Translation Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE).

Baselines. We compare with specialist camera tracking methods, including DROID-SLAM [94], ACE-Zero [13], and Align3R [66]. Note that all baseline methods leverage explicit geometry constraints and perform bundle adjustment during test-time optimization. ACE-Zero and Align3R [66] also reply on off-the-shelf monocular depth and optical flow predictions as additional inputs. In contrast, *EgoM2P* is trained *without* any geometry modeling or 3D inductive bias and can predict camera poses directly from RGB inputs in a single feed-forward pass.

Results. See metrics in Tab. 2. Compared with specialist SOTA models that involve explicit geometry modeling, our versatile model trained without any 3D inductive bias achieves comparable performance. Notably, while all baselines require time-consuming test-time optimization, *EgoM2P* achieves 300+ FPS inference speed, thanks to our parallel decoding approach. Egocentric camera tracking can sometimes be challenging due to rapid motion speed and little camera parallax. As shown in Fig. 3, while baselines may suffer from temporal jitter and error accumulation in some cases, *EgoM2P* can predict smooth and plausible camera trajectories in the sense that it learns to capture the uniqueness of egocentric motion. This capability can even generalize to the out-of-domain ADT test set.



(a) **Visualization of camera trajectories.**



(b) **Comparison of camera tracking on ADT [80].** Ground truth and predictions are represented by black and colored wireframes, respectively.

Figure 3. Egocentric capture often involves rapid head rotations, which challenges baseline tracking methods. However, *EgoM2P* effectively predicts smooth and plausible camera poses in the shown examples. This capability also generalizes to the unseen ADT [80] dataset *without* post-training.

4.2. Egocentric Gaze Dynamics Estimation

Evaluation Protocols. We sample 1,000 videos from the validation split of the EgoExo4D dataset [30]. We normalize both ground truth and predicted labels of all methods to the range $[0, 1]$ and evaluate gaze estimation accuracy using the mean squared error (MSE).

Baselines. We compare *EgoM2P* with two SOTA methods that predict 2D gaze locations from egocentric video: Huang et al. [40] and Lai et al. [50], using their official implementations for evaluation.

Results. We use input videos at 30 FPS for the baselines, and following baselines' respective settings, input video frames are resized to 224×224 for [40] and 256×256 for [50]. Our method achieves the lowest MSE (**0.0162**), outperforming Huang et al. [40] (0.0255) and Lai et al. [50] (0.0175). For qualitative comparisons, refer to Fig. 4 and Sup. Vid. *EgoM2P* produces more consistent gaze predictions, highlighting our capability to understand human intentions as one of its multitasking abilities.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

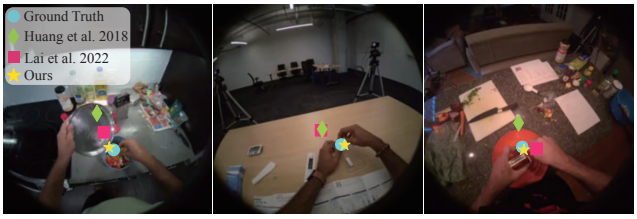


Figure 4. **Gaze dynamics estimation.** *EgoM2P* can predict results that are more aligned with human intentions.

Method	H2O [49] Abs Rel ↓	H2O [49] $\delta_{1.25}$ ↑	HOI4D [60] (*unseen*) Abs Rel ↓	HOI4D [60] (*unseen*) $\delta_{1.25}$ ↑	Time ↓
RollingDepth [44]	0.087	90.5	0.057	97.6	37s
Align3R [66]	0.074	91.8	**0.045**	**98.1**	90s
EgoM2P	**0.055**	**96.0**	0.061 **<u>0.041</u>**	98.0 **<u>99.0</u>**	**0.8s**

Table 3. **Evaluation on egocentric video depth estimation.** Compared to specialist SOTAs requiring geometry-based test-time optimization, the versatile *EgoM2P* achieves comparable performance while being significantly more efficient. With post-training described in Sec. 4.5, *EgoM2P* excels (see <u>underlined results</u>).

4.3. Egocentric Monocular Video Depth Estimation

Evaluation Protocols. We evaluate on the test split of H2O [49], which contains 236 two-second video clips. To validate our method's cross-domain generalization capability, we also evaluate on 100 two-second video clips from HOI4D [60], which is entirely unseen during *EgoM2P*'s training. We temporally downsample each video clip to 8 FPS and resize it to 256×256 as input. We align the estimated relative depth with the GT depth by a sequence-level scale and translation factor and evaluate the depth accuracy with absolute relative error (Abs Rel) and the percentage of predicted depths within a 1.25-factor of true depth ($\delta_{1.25}$).

Baselines. We compare *EgoM2P* with two specialized video depth estimators, RollingDepth [44], and Align3R [66]. Both methods employ pretrained networks [43, 104] to estimate monocular or pair-wise depth maps and then conduct hierarchical sequence-level optimization to temporally align per-frame depths.

Results. We report the results in Tab. 3 and provide visual comparisons in Fig. 5. RollingDepth and Align3R require sequence-level optimization after per-frame predictions, which can take as long as one minute for a two-second sequence. In contrast, *EgoM2P* predicts the entire depth video in a single end-to-end feed-forward pass, achieving at least 30x faster inference speed. Quantitatively, *EgoM2P* achieves comparable performance with baseline methods and even the best $\delta_{1.25}$ on two test sets, particularly the unseen HOI4D dataset, which validates *EgoM2P* 's generalization to out-of-domain egocentric data. Reducing the quantization error of the Cosmos tokenizer [3] would further improve our depth prediction result. Please refer to our supplementary videos for more visualization.

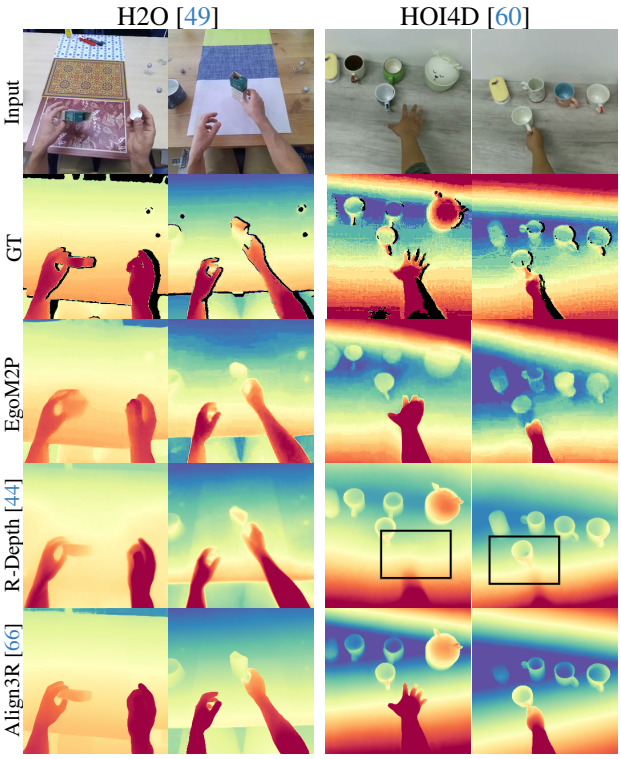


Figure 5. **Egocentric video depth estimation**. *EgoM2P* achieves comparable performance with specialist SOTA methods. Rolling Depth [44] struggles in estimating the depth of hands, an important component in egocentric view, while our method can capture hand movement even in the out-of-domain HOI4D [60] dataset, *without* any post-training or fine-tuning.

4.4. Conditional Egocentric Video Synthesis

EgoM2P can synthesize RGB videos from other modalities. This section focuses on depth-to-RGB video synthesis.

Evaluation Protocols. We randomly sampled 142 depth videos from the HoloAssist [106] test set and 100 depth videos from the ASE [6] dataset. Note that the entire ASE dataset is unseen in our model's training and can be used to assess our method's generalizability. For ASE, we use their labeled fisheye depth videos as inputs. For HoloAssist, since its depth labels are misaligned with the RGB frames, we employ RollingDepth [44] to generate pseudo-depth labels for each RGB frame as input. For quantitative metrics, we employ Fréchet Video Distance (FVD) [96], Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and the perceptual metric LPIPS [124].

Baselines. We compare our method with two SOTA ControlNet-based approaches: Control-A-Video [19] and ControlVideo [126], both support depth as a conditional input.

Results. Since the two baselines require additional text input, we adopt different templates for each dataset. In HoloAssist, we use the template "Two hands with {object}", where {object} represents the item being held

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

Method	HoloAssist [106]				ASE [6] (*unseen*)			
	FVD* ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FVD* ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Control-A-Video [19]	2.309	0.185	9.25	0.677	2.226	0.289	**11.11**	0.817
ControlVideo [126]	1.363	0.235	8.18	0.653	1.392	0.275	10.46	**0.676**
EgoM2P	**0.759**	**0.592**	**15.163**	**0.336**	**1.336**	**0.308**	6.923	0.715
					<u>**0.525**</u>	<u>**0.594**</u>	<u>**16.924**</u>	<u>**0.520**</u>

Table 4. **Evaluation on depth-to-RGB video synthesis.** *EgoM2P* outperforms baselines on the HoloAssist test set, producing higher-quality egocentric videos. On the unseen ASE dataset, it generates videos that more closely resemble real ones with a lower FVD* (FVD$/10^3$). With post-training (Sec. 4.5), *EgoM2P* excels on unseen datasets indicated by <u>underlined results</u>.

in the RGB video. For ASE, since it consists of indoor videos, we use the prompt "Indoor scenes". In Fig. 6, we present a qualitative comparison highlighting the challenges faced by two baselines in generating RGB frames that accurately correspond to the input depth maps. These baseline methods often produce outputs with significant discrepancies in semantic and geometric alignment, especially on the ASE [6] dataset. In contrast, our approach *EgoM2P* demonstrates superior performance by maintaining alignment between the depth maps and the generated RGB frames. The quantitative results are reported in Tab. 4. On the HoloAssist dataset [106], our model outperforms the baseline approaches. On the ASE dataset [6], which is unseen and stylistically different from our training data, our method demonstrates strong generalization capabilities, generating egocentric videos that more closely resemble real ones, as indicated by a lower FVD score. In contrast, baseline models, initialized with the powerful Stable Diffusion [84], tend to produce hallucinations. While they achieve higher PSNR scores, their outputs often deviate from the true egocentric video distribution, as indicated by their higher FVD scores.

4.5. Post-Training

The pretrained *EgoM2P* demonstrates strong generalization abilities on cross-dataset generalization tests. Additionally, in Tab. 2, 3, and 4, we show that *EgoM2P* can be further enhanced via post-training on the training sets of unseen datasets, enabling rapid adaptation to new domains with minimal data. See more details in Supp. Mat. Sec. D.

5. Conclusion

We propose *EgoM2P*, the first multimodal and multitask large egocentric model integrating four common modalities in egocentric vision. To handle complex spatiotemporal dynamics in multimodalities, we propose a unified temporal tokenizer architecture to tokenize gaze and camera trajectory into discrete tokens encoded with temporal information. To address heterogeneity in egocentric datasets, we extend multimodal masked modeling to the video domain and pretrain *EgoM2P* with 400 billion tokens sampled from our 4 billion multimodal token database. *EgoM2P* matches or surpasses state-of-the-art specialist models and demonstrates efficiency in various downstream applications, including egocentric camera tracking, gaze estimation in egocentric videos, egocentric monocular depth estimation, and conditional egocentric video synthesis.

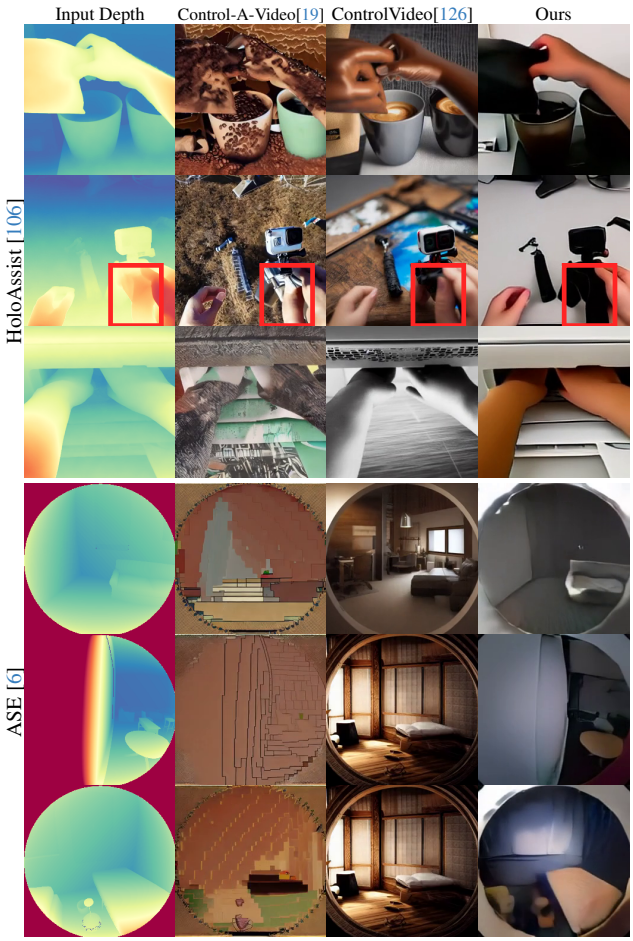


Figure 6. **Comparison of depth-to-RGB video synthesis.** Red boxes highlight incorrectly generated fingers in baselines, while ours generate meaningful hand motion. Our results show improved alignment with the input depth, minimizing hallucinations. *No* post-training was applied for ASE results in this figure.

Limitations. The visual quality of the synthesized videos is inherently limited by current state-of-the-art video tokenizers [3]. Although better tokenizers could reduce video quality loss during quantization, this is not our main focus. Leveraging a diffusion decoder conditioned on discrete video tokens to enhance visual quality could be effective.

Future Work. Wearable devices have constrained computing resources and demand real-time processing for seamless human-computer interactions. We aim to explore performance optimizations for *EgoM2P* on embedded GPUs. In this work, we consider the most common modalities in egocentric vision. Integrating hand motion, audio, text, etc., into the model is a promising direction for future work.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

Acknowledgements. Gen Li was supported by a Microsoft Spatial AI Zurich Lab PhD scholarship, Yutong Chen was supported by the Swiss Innovation Agency Innosuisse, and Kaifeng Zhao was supported by the SDSC PhD fellowship. This work was also supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a03 on Alps. We sincerely thank Siwei Zhang for the fruitful discussions.

References

[1] Gpt-4v(ision) system card. 2023. 2

[2] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft, 2024. 2

[3] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2, 3, 4, 7, 8, 1

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022. 2

[5] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[6] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard A. Newcombe, and Vasileios Balntas. Scenescript: Reconstructing scenes with an autoregressive structured language model. In *ECCV*, 2024. 7, 8, 4

[7] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 2022. 5

[8] Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4M-21: An any-to-any vision model for tens of tasks and modalities. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. 2, 4, 5, 3

[9] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957, 2015. 3

[10] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 1, 2, 3

[11] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 1, 2

[12] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[13] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 6

[14] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 2

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[17] 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. 1

[18] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J. Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. In *Advances in Neural Information Processing Systems*, 2022. 2

[19] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning, 2024. 7, 8

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

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[22] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and Content-Guided Video Synthesis with Diffusion Models . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7312–7322, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1

[23] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[24] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 3

[25] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[26] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[27] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 2

[28] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. 1

[29] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, et al. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[30] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2023. 1, 2, 3, 5, 6

[31] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[32] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 5

[34] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[35] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646. Curran Associates, Inc., 2022. 1

[36] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 5

[37] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[38] Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2022. 5

[39] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23369–23379, 2022. 1

[40] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 789–804. Springer, 2018. 6

[41] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[42] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5492–5501, 2019. 3

[43] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[44] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

Konrad Schindler. Video depth without video models. In *CVPR*, 2025. 3, 7

[45] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15908–15918, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1

[46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[47] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah J. Harmsen, and Neil Houlsby. UVim: A unified modeling approach for vision with learned guiding codes. In *Advances in Neural Information Processing Systems*, 2022. 2

[48] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, et al. VideoPoet: A large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 25105–25124. PMLR, 2024. 1

[49] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 1, 3, 5, 7, 4

[50] Bolin Lai, Miao Liu, Fiona Ryan, and James M. Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 227. BMVA Press, 2022. 6

[51] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2024. 1

[52] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. Egogen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14497–14509, 2024. 1, 3, 4

[53] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

[54] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 4

[55] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[56] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021. 3

[57] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[58] Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Zheng Shou. Exocentric-to-egocentric video generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[59] Ruyang Liu, Chen Li, Yixiao Ge, Thomas H. Li, Ying Shan, and Ge Li. BT-Adapter: Video Conversation is Feasible Without Video Instruction Tuning . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13658–13667, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1

[60] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 1, 7, 4

[61] Y. Liu, H. Yang, X. Si, L. Liu, Z. Li, Y. Zhang, Y. Liu, and L. Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21740–21751, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1, 3

[62] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[63] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1, 2

[64] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26429–26445, 2023. 2

[65] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[66] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *CVPR*, 2025. 6, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

[67] LumaLabs. Dream Machine. https://lumalabs.ai/dream-machine, 2024. 1

[68] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liangsheng Wang, Yujun Shen, Deli Zhao, Jinren Zhou, and Tien-Ping Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, 2023. 1

[69] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 1

[70] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024. 1

[71] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 1, 2

[72] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier Biard, Sam Dodge, Philipp Dufter, Bowen Zhang, Dhruti Shah, Xianzhi Du, Futang Peng, Haotian Zhang, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024. 2

[73] Meta. Project Aria Glasses. https://www.projectaria.com/, 2023. 3

[74] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*, 2022. 2

[75] Microsoft. HoloLens 2. https://www.microsoft.com/en-us/hololens, 2019. 3, 1

[76] David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 4, 5, 3

[77] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 1

[78] OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. 1, 2

[79] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2

[80] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *ICCV*, 2023. 6, 3, 4

[81] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1, 2

[82] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[83] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), 2020. 4, 5

[84] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 8

[85] RunwayML. Gen-3 Alpha. https://runwayml.com/research/introducing-gen-3-alpha, 2024. 1

[86] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *European Conference on Computer Vision*, pages 686–704. Springer, 2022. 3

[87] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *CVPR 2022*. 1

[88] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[89] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard New-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

combe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[90] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L. Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. AlanaVLM: A multimodal embodied AI foundation model for egocentric video understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11101–11122, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[91] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[92] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[93] Kuaishou Technology. Kling ai video generator. https://kling.kuaishou.com/en, 2024. Accessed: 2025-03-03. 2

[94] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-d cameras. In *NeuRIPS*, 2021. 3, 6

[95] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2, 5

[96] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. 7

[97] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 2

[98] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Neural Information Processing Systems*, 2017. 3, 4

[99] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[100] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2023. 1

[101] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5250–5261, 2023. 3

[102] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 2

[103] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 23318–23340. PMLR, 2022. 2

[104] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 7

[105] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[106] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20270–20281, 2023. 1, 2, 3, 7, 8

[107] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[108] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024. 3

[109] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2

[110] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, pages 396–416. Springer, 2024. 2

[111] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.

of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1

[112] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*, pages 53366–53397, 2024. 1

[113] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 2

[114] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for
zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2021. Association for Computational Linguistics. 2

[115] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. X-gen: Ego-centric video prediction by watching exo-centric videos. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[116] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 1

[117] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[118] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, Yinfei Yang, and Bowen Zhang. MMEgo: Towards building egocentric multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[119] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19717–19728, 2023. 3

[120] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. In *International Conference on Learning Representations*, 2022. 4, 1

[121] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. 1

[122] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OakInk2 : A Dataset of Bimanual Hands-Object Manipulation in Complex Task Completion . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 445–456, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1

[123] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 1

[124] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[125] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, pages 180–200. Springer, 2022. 1

[126] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 7, 8

[127] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. 1

[128] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 676–694. Springer, 2022. 3

[129] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 3

[130] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 09:24:37 UTC from IEEE Xplore. Restrictions apply.