LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning

Bolin Lai[1,2,†], Xiaoliang Dai[1], Lawrence Chen[1], Guan Pang[1], James M. Rehg[3], and Miao Liu[1,*]

[1] GenAI, Meta
[2] Georgia Institute of Technology
[3] University of Illinois Urbana-Champaign
bolin.lai@gatech.edu {xiaoliangdai,lawrencechen,gpang,miaoliu}@meta.com
jrehg@illinois.edu

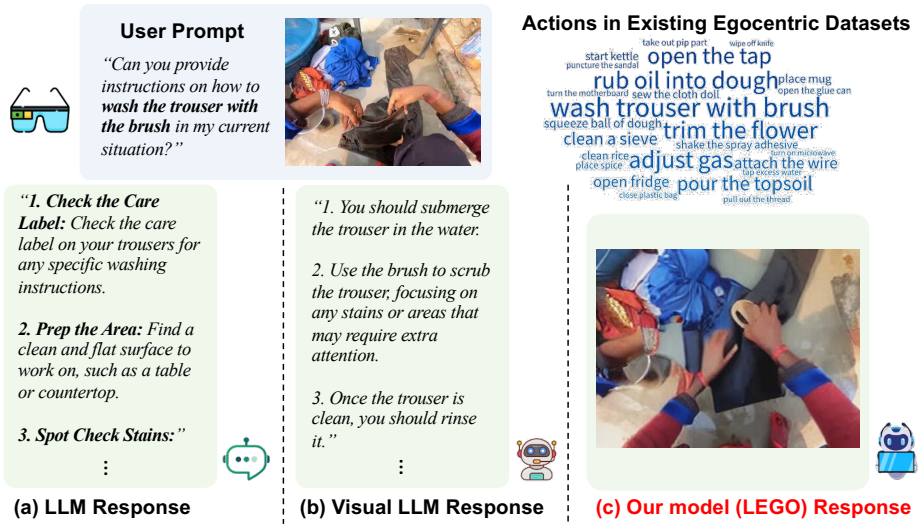

Fig. 1: When a user performing a complex task asks a large language model (LLM) for instructions (a) on how to complete the steps, she receives a generic answer and has to translate the guidance into her specific situation. If she is wearing a camera, then the prompt can be augmented with an egocentric view of the scene and passed to a Visual LLM (b), and the description is now contextualized to her situation. But she still faces the challenge of parsing a written description. When she uses our novel LEGO model (c), however, the combined image and prompt are used to *automatically generate an image* that provides visual guidance tailored to her exact situation from the egocentric viewpoint. Now she can complete her task seamlessly!

Abstract. Generating instructional images of human daily actions from an egocentric viewpoint serves as a key step towards efficient skill transfer. In this paper, we introduce a novel problem – egocentric action

[†]This work was done when the first author was an intern at GenAI, Meta.
[*]Corresponding author.

frame generation. The goal is to synthesize an image depicting an action in the user's context (*i.e.*, action frame) by conditioning on a user prompt and an input egocentric image. Notably, existing egocentric action datasets lack the detailed annotations that describe the execution of actions. Additionally, existing diffusion-based image manipulation models are sub-optimal in controlling the state transition of an action in egocentric image pixel space because of the domain gap. To this end, we propose to Learn EGOcentric (LEGO) action frame generation via visual instruction tuning. First, we introduce a prompt enhancement scheme to generate enriched action descriptions from a visual large language model (VLLM) by visual instruction tuning. Then we propose a novel method to leverage image and text embeddings from the VLLM as additional conditioning to improve the performance of a diffusion model. We validate our model on two egocentric datasets – Ego4D and Epic-Kitchens. Our experiments show substantial improvement over prior image manipulation models in both quantitative and qualitative evaluation. We also conduct detailed ablation studies and analysis to provide insights in our method. More details of the dataset and code are available on the website (`https://bolinlai.github.io/Lego_EgoActGen/`).

Keywords: Egocentric Vision · Instruction Tuning · Diffusion Model

1 Introduction

The emergence of Large Language Models (LLMs) [6, 10, 64, 99], such as ChatGPT, has revolutionized the transfer of knowledge. However, an LLM alone is not a sufficient tool for human skill transfer. Consider the question answering example in Fig. 1(a). The LLM can summarize general world knowledge, but its response may not be directly applicable to the user's current circumstances. To bridge this gap, egocentric visual perception provides a novel means to capture the actions and intentions as well as the surrounding context of the camera wearer. As shown in Fig. 1(b), recent Visual Large Language Models (VLLMs) [1, 12, 46, 52, 102] can generate instructions based on the egocentric visual input. However, such verbose textual instructions are not the optimal medium for enabling efficient human skill transfer (*e.g.*, via AR devices). Neuroscience studies have revealed that the human brain processes text much more slowly than images [4], and that humans can interpret an action from a single static image [22]. Motivated by these discoveries, we seek to design an image generation architecture that can synthesize an image which not only vividly depicts how an action should be conducted, but also seamlessly aligns with the user's visual perspective.

Formally, we introduce the novel problem of egocentric action frame generation as depicted in Fig. 1(c). Given a user query about how to perform a specific action and an egocentric image capturing the moment before the action begins, the goal is to synthesize an egocentric image illustrating the execution of the action in the same egocentric context. We address this problem by harnessing diffusion models [27, 67], which have been shown to be powerful tools for image

manipulation [5,25,44,56,100]. There are two major challenges in using diffusion models to generate action frames from an egocentric perspective. First, the action annotations of the existing egocentric datasets [13,21] are simply composed of a verb and nouns (see word cloud in Fig. 1), and thus lack the necessary details for diffusion models to learn the action state transition and to associate the action with correct objects and body movements. Second, the existing diffusion models are pre-trained primarily on *exocentric* (third-person-view) images, and have limited ability to represent complicated human daily activities from an egocentric perspective. In contrast, our proposed problem requires image generation in the *egocentric* view, conditioned on the user prompt of *actions*. The resulting domain gap impedes existing methods from synthesizing faithful and consistent egocentric action frames.

To address these challenges, we propose to Learn EGOcentric (LEGO) action frame generation with visual instruction tuning. First, we introduce a prompt enhancement scheme to generate enriched action descriptions at scale from an instruction-tuned VLLM. Second, we incorporate the image and text embeddings from finetuned VLLM as additional conditioning into the diffusion model to narrow the domain gap and improve the controllability of action frame generation. Our experimental results suggest that the enriched action descriptions and our innovative utilization of VLLM embeddings both improve the image generation performance. Our model is able to provide a generated key action frame together with a detailed action descriptions to facilitate human skill transfer from the egocentric perspective. Overall, our contributions can be summarized as follows:

- We introduce the novel problem of egocentric action frame generation to facilitate the process of skill transfer and address the challenges of missing action details in prompts and domain gap in existing image diffusion models.
- We propose a prompt enhancement strategy based on visual instruction tuning to enrich egocentric action descriptions, and demonstrate how the enriched descriptions can help the diffusion model understand the action state transition from the egocentric perspective.
- We propose a novel approach to incorporate the text and visual embeddings from the VLLM into the latent diffusion model to bridge the domain gap and improve the performance for egocentric action frame generation.
- We conduct thorough experiments on the Ego4D and Epic-Kitchens datasets to validate the superiority of our model over prior approaches. We also showcase the contribution of each component of our model design through ablation studies. We further provide analysis on how the visual instruction-tuned embeddings benefit model performance.

2 Related Work

Text-Guided Image Manipulation. The recent emergence of diffusion models enables text-guided image manipulation including image restoration [33], style

transfer [75], personalized image synthesis [68,71,87], pose generation [28,36] and generic image editing [15,18,34,38,44,57,60,79,82,82–84,94,98,101]. SDEdit [56] converts the image to the latent space by adding noise through a stochastic differential equation and then denoises the representation for image editing. Rombach *et al*. [67] further expand SDEdit from the original stroke-based editing to text-based editing. Null-text inversion (NTI) [58] inverts a real image by DDIM [35] to yield the diffusion process and then reconstructs the image. The image can then be edited following the same strategies as Prompt-to-Prompt [25]. NTI relies on accurate inversion process which can be improved by using coupled transformations [80] or proximal guidance [24] and accelerated by a blended guidance strategy [61]. To associate the nouns with correct objects, DIFFEDIT [11] generates a mask to localize the manipulated regions. However, most inversion-based methods require accurate image captions, which are largely unavailable in the existing egocentric dataset. Recently, InstructPix2Pix [5] demonstrates the potential to edit a real image without the inversion process and original captions. However, how to leverage the diffusion model to control the state transition of an action within the egocentric image plane remains unexplored.

Large Language Model for Image Generation. LLMs [6, 10, 64, 76, 78, 99] and VLLMs [1, 12, 23, 46, 52, 73, 95, 102] have shown their strong capability of understanding complex human instructions. Recently, LLMs and VLLMs are used to guide image generation [2, 7–9, 51, 93, 102]. Wen *et al*. [88] use a pretrained VLLM to pinpoint the misalignment of the text prompt and the synthetic image and then correct it using the diffusion model. Lian *et al*. [49] propose to generate a layout map using an LLM to improve the understanding of prompts with spatial and numerical reasoning. InstructEdit [83] uses BLIP-2 [46] to infer the objects that will be edited and then generates a mask with SAM [39] for object grounding. A pre-trained LLM can also be used as a controller to connect with various foundation models [70,89,90]. GILL [40] learns text-relevant image embeddings from VLLM in a few additional tokens for image generation. Importantly, all previous efforts apply the off-the-shelf foundational models directly to their problems without finetuning. In contrast, our method uses visual instruction tuning to improve the image and text embeddings from the VLLM, which narrows the domain gap and thereby facilitates the action frame generation from the egocentric point of view.

Egocentric Vision. Recent efforts seek to understand human actions and perceptual attention [16, 29, 30, 37, 41–43, 48, 69, 74, 86], model hand-object interactions [20, 53, 54, 65, 91], and estimate human body poses [45, 55, 77, 81] from the egocentric perspective. Here, we mainly discuss the most relevant works on egocentric visual-language models and egocentric visual content generation. Lin *et al*. [50] propose the first egocentric video-language pre-training model – EgoVLP. Pramanick *et al*. [62] further improve it by incorporating multi-modal fusion directly into the video and language backbone. Ashutosh *et al*. [3] propose to learn a hierarchical video-language embedding for long egocentric videos. Ramakrishnan *et al*. [66] introduce NaQ, which is a data augmentation strategy to train models for long egocentric video search with natural language queries. In terms

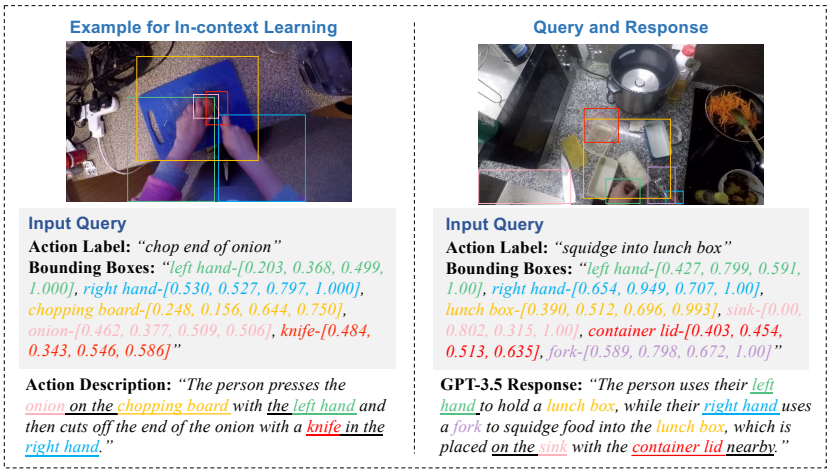

Fig. 2: Examples for data curation using GPT-3.5. We provide detailed action descriptions of several example images as well as their action labels and bounding boxes for in-context learning. In addition, we input the action label and bounding boxes of another action as a query. GPT-3.5 is able to generate descriptions with enriched information (highlighted by underlines) in the response.

of egocentric visual generation, Jia *et al*. [32] leverage GANs [19] to generate future head motion in hand forecasting task. Zhang *et al*. [96] leverage GANs to facilitate future gaze anticipation. Ye *et al*. [92] propose the affordance diffusion model that takes in the image of an object and generates possible hand-object interactions in the egocentric view. In this paper, we present the first work that investigates how to leverage VLLMs and diffusion models to generate action state transition on the egocentric image plane.

3 Method

The problem setting of egocentric action frame generation is illustrated in Fig. 1(c). Given an egocentric image frame $\mathcal{X}$ that captures the user's current visual context as well as a user query prompt $\mathcal{P}$ regarding how to perform an action, our goal is to synthesize the action frame $\mathcal{Y}$ that visually depicts how the action should be conducted in the same situation (*i.e.*, keep a consistent background).

The key insight of our proposed LEGO model is leveraging the strong capability of a VLLM to enhance the diffusion model for egocentric action frame generation. The annotations of existing egocentric datasets do not describe the details of how actions are conducted. As a remedy, we leverage visual instruction tuning to finetune a VLLM that enriches action descriptions based on the egocentric visual prompt. In addition, the existing diffusion-based image manipulation models are limited in understanding egocentric action state transition, due to the domain gap between the exocentric pre-training dataset and the ego-

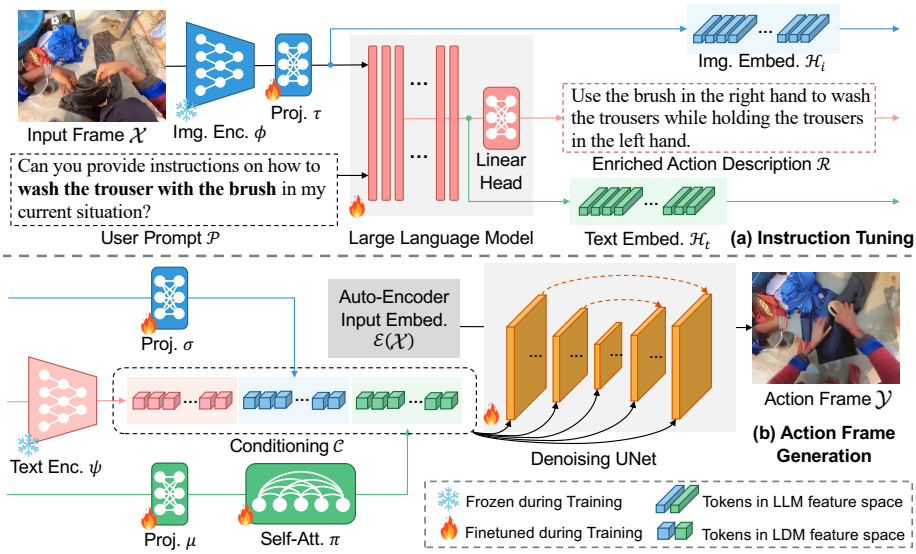

Fig. 3: Overview of our proposed LEGO model. We first finetune a visual large language model (VLLM) to generate the enriched action description with visual instruction tuning. We then project image and text embeddings from the finetuned VLLM to the feature space of the latent diffusion model (LDM). Finally, we train the LDM to synthesize the egocentric action frame conditioning on the input frame, enriched action description, as well as the VLLM image and text embeddings.

centric action dataset for our problem. To bridge this gap, we propose a novel approach that leverages VLLM embeddings to control the state transition of actions and to generate action frames accordingly. We detail the VLLM-based data enrichment pipeline and our model design in the following sections.

3.1 Egocentric Visual Instruction Tuning

Data Curation for Visual Instruction Tuning. As shown in Fig. 2, we use GPT-3.5 to generate detailed action descriptions based on an input query composed of a short action label and object bounding boxes that are provided in the existing datasets. First, we prepare several examples of possible inputs along with their expected output descriptions for GPT-3.5 to perform *in-context learning*. These examples cover a diverse set of scenes in the egocentric action dataset. Each example is composed of an action label, a manually annotated action description, and relative spatial information of hands and objects-of-interests represented by bounding box coordinates. GPT-3.5 can learn from the given examples and generate similar detailed action descriptions based on a new input query. The resulting GPT-3.5 curated data is then further used for visual instruction tuning. More details of the prompt are provided in the supplementary.

Visual Instruction Tuning. We follow the finetuning strategy in prior work [52], as shown in Fig. 3(a). Specifically, we use the pre-trained CLIP visual encoder [63] ϕ to encode the visual representation and then apply a linear projection layer τ to map the CLIP visual features into the semantic space of the LLM, *i.e.*, $\mathcal{H}_i = \tau(\phi(\mathcal{X}))$. To construct the user prompt $\mathcal{P}$, we insert the action label annotation into a prompt template to create a coherent query that is aligned with the instruction-following nature of an LLM. We then tokenize $\mathcal{P}$, and feed both the prompt text tokens and image tokens $\mathcal{H}_i$ as inputs into the LLM. Finally, the LLM is trained to generate enriched action description (denoted as $\mathcal{R}$) based on the user prompt and image input.

User Prompt Enrichment at Scale. Note that the visual instruction tuned VLLM doesn't rely on any object bounding boxes as input. Therefore, we can generate enriched action descriptions for all egocentric action data at scale.

3.2 Egocentric Action Frame Generation

We leverage a latent diffusion model (LDM) [67] to synthesize the action frame conditioning on the input frame and the detailed action description $\mathcal{R}$ generated by our finetuned VLLM (see Fig. 3(b)). Following the regular steps in LDMs [5, 67], the input image is first encoded into a latent space using a pre-trained auto-encoder $\mathcal{E}$. Then the input to the denoising UNet is a concatentation of the latent input representation $\mathcal{E}(\mathcal{X})$ and a Gaussian noise $\mathcal{G}$.

Our key innovation is to design the U-Net conditioning component so that the diffusion model can interpret the egocentric actions correctly. To start, we follow [5] and adopt the conventional pre-trained CLIP text encoder ψ to extract a text representation of $\mathcal{R}$, *i.e.*, $\psi(\mathcal{R}) \in \mathbb{R}^{N \times D}$ where N is the maximum number of text tokens and D is the number of feature channels. We further leverage the image and text embeddings from the visual instruction tuned VLLM as additional LDM conditioning to alleviate the domain gap issue.

Specifically, we feed the VLLM image embedding $\mathcal{H}_i$ into an extra linear layer σ to map it to LDM feature space, *i.e.*, $\sigma(\mathcal{H}_i) \in \mathbb{R}^{M \times D}$, where M is the number of image tokens. Note that $\mathcal{H}_i$ is already projected to the semantic space during visual instruction tuning, and therefore differs from the image embedding $\mathcal{E}(\mathcal{X})$ from the auto-encoder. Moreover, we also extract the text embedding $\mathcal{H}_t$ before the last linear layer of LLM. We adopt a fixed token number N and enforce padding or truncation behavior, as in the CLIP text encoder. The text embedding is then fed to a projection layer μ. In LLM decoder, the response is generated iteratively and each word embedding only conditions on the context ahead of it. To extract the holistic semantic meaning of $\mathcal{H}_t$ in LDM feature space, we further add self-attention layers π after the projection, *i.e.*, $\pi(\mu(\mathcal{H}_t)) \in \mathbb{R}^{N \times D}$. Thus, the U-Net conditioning can be formulated as:

$$\mathcal{C} = \Big[\psi(\mathcal{R}), \sigma(\mathcal{H}_i), \pi(\mu(\mathcal{H}_t))\Big] \in \mathbb{R}^{(2N+M) \times D}. \tag{1}$$

The conditioning $\mathcal{C}$ is fed into the denoising UNet at multiple layers via the cross-attention mechanism [67]. We assume the intermediate feature of a specific

UNet layer is $\mathcal{U}$, which is learned from the UNet input (*i.e.*, the concatenation of input frame representation $\mathcal{E}(\mathcal{X})$ and Gaussian noise $\mathcal{G}$). The cross-attention at this UNet layer can be formulated as:

$$CrossAtt(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right) \cdot V, \tag{2}$$

where $Q = W_Q \cdot \mathcal{U}$, $K = W_K \cdot \mathcal{C}$ and $V = W_V \cdot \mathcal{C}$. Note that W_Q, W_K and W_V are learnable matrices. We also adopt the classifier-free guidance following [5] (see supplementary for details). Finally, the UNet output is converted to the image domain by a pre-trained decoder.

3.3 Implementation Details

All parameters of the VLLM are initialized from the pre-trained LLaVA [52] weights. During training, we freeze the CLIP image encoder and finetune the projection layer and LLM with cross-entropy loss for 3 epochs. To improve the diversity of the text prompts, we randomly select the question template from 10 candidates in each iteration. For LDM training, the text encoder, UNet and auto-encoder are initialized with pre-trained weights [67]. The projection layers σ and μ and the self-attention layers π are initialized using the Xavier algorithm [17]. The text encoder is frozen and the remaining weights are finetuned with L2 regression loss between the predicted noise and real noise for 20,000 iterations. All input and target images are resized to a resolution of 256×256. Please refer to the supplementary for more details of training and inference.

4 Experiments

4.1 Data and Metrics

Datasets. We conduct our experiments on two well-established egocentric action datasets – Ego4D [21] and Epic-Kitchens-100 [13]. Both datasets were densely annotated with action starting time t and ending time $\hat{t}$. In our problem setting, we select an egocentric image frame δ_i seconds before the action begins as the input $\mathcal{X}$, and an image δ_o seconds after the action begins as the target frame $\mathcal{Y}$. On the Ego4D dataset, based on the annotations of Pre-Condition-15 time (PRE-15) t_{pre} , and Point-of-No-Return time (PNR) t_{pnr}, we set $\delta_i = t - t_{pre}$ and $\delta_o = t_{pnr} - t$. For Epic-Kitchens, PNR and PRE-15 annotations are not available. Instead, we empirically select $\delta_i = 0.25$ seconds and $\delta_o = t + \lambda * (\hat{t} - t)$, where $\lambda = 0.6$, for our experiments. More details of data preparation and improvement are elaborated in the supplementary.
Metrics. We adopt image-to-image similarity, image-to-text similarity, and user study as metrics in our experiments.

- **Image-to-Image Metrics.** We implement six metrics to evaluate image-to-image similarity. To begin with, we adopt three contrastive learning based

metrics including image-to-image (1) EgoVLP score [50], (2) EgoVLP$^+$ score [50] and (3) CLIP score [63]. EgoVLP is a contrastive video-language pre-training model trained with egocentric videos. Since EgoVLP takes multiple frames as input, we consider two types of inputs: duplicating the output frame as a static input (*i.e.*, EgoVLP score) and combining the input frame with the output frame (*i.e.*, EgoVLP$^+$ score). As as result, EgoVLP$^+$ can effetively measure whether the generated frame can depict the state transition of an action. Importantly, given that EgoVLP is pre-trained on egocentric data and action labels, we consider EgoVLP and EgoVLP$^+$ score as the *primary* automatic metrics. In addition, we also report (4) Fréchet Inception Distance (FID) [26], (5) Peak Signal-to-Noise Ratio (PSNR) and (6) Learned Perceptual Image Patch Similarity (LPIPS) [97] (with SqueezeNet [31] as the encoder) to make a thorough evaluation. Note that instead of measuring similarity of input and output frames as in prior works [12, 80], in our problem setting, we measure the similarity between the generated action frame and ground truth, which better reflects whether the generation results can illustrate the execution of an action.

- **Image-to-Text Metrics.** We find the widely-used image-to-text CLIP score can not align actions with egocentric images due to the domain gap. Similar misalignment problem is also observed in [14, 59, 72, 85]. In our experiments, we utilize BLIP [47] to generate image captions of the output images and then calculate text-to-text similarity using CLIP text encoder (following [34]). We implement this metric with two BLIP structures: BLIP-B and BLIP-L. Though this solution may still suffer from the same domain gap issue, it is a more appropriate evaluation metric in our problem setting. See more evidence and discussions in the supplementary.
- **User Study.** We also conduct a user study on a subset of test data to further validate the advantage of our model based on human preference. We sample 60 examples from each dataset and present the generated frames from our model as well as the baseline models to raters on Amazon Mechanical Turk (AMT). We also provide the input frames and the corresponding action labels during evaluation. For each instance, we ask the raters to select the image that best aligns with the provided action label while preserving the most contextual information from the input frame. To minimize potential bias, we hire 5 AMT raters to annotate each example thus resulting in 300 samples for each dataset. User study interface is shown in supplementary.

4.2 Comparison with Prior Approaches

We compare our proposed model with previous state-of-the-art methods for text-guided image manipulation, including ProxEdit [24], SDEdit [56] and InstructPix2Pix (IP2P) [5]. For a fair comparison, we finetune these baseline methods with the same data used in our experiments. Specifically, we train IP2P in an end-to-end way on the two datasets with existing egocentric action labels as

Table 1: Comparison with prior image manipulation approaches in image-to-image metrics. ↓ means a lower score in this metric suggests a better performance. The best results are highlighted with **boldface**. The orange row refers to our LEGO model.

	Methods	EgoVLP	EgoVLP$^+$	CLIP	FID ↓	PSNR	LPIPS ↓
Ego4D	ProxEdit [24]	44.51	72.68	68.17	33.01	11.88	40.90
	SDEdit [56]	50.07	72.90	73.35	33.35	11.81	41.60
	IP2P [5]	62.19	78.84	78.75	24.73	12.16	37.16
	LEGO	**65.65**	**80.44**	**80.61**	**23.83**	**12.29**	**36.43**
EK-100	ProxEdit [24]	32.27	52.77	65.80	51.35	11.06	46.35
	SDEdit [56]	33.84	56.80	74.76	27.41	11.30	43.33
	IP2P [5]	42.97	61.06	77.03	**20.64**	11.23	40.82
	LEGO	**45.89**	**62.66**	**78.63**	21.57	**11.33**	**40.36**

Table 2: Image-to-text metrics of our model and baselines. The best results are highlighted with **boldface**. The orange row refers to our LEGO model performance.

Methods	Ego4D		Epic-Kitchens	
	BLIP-B	BLIP-L	BLIP-B	BLIP-L
ProxEdit [24]	17.73	17.35	23.65	23.39
SDEdit [56]	19.80	19.74	21.51	21.30
IP2P [5]	20.00	20.56	25.37	26.36
LEGO	**20.38**	**20.70**	**26.98**	**27.41**

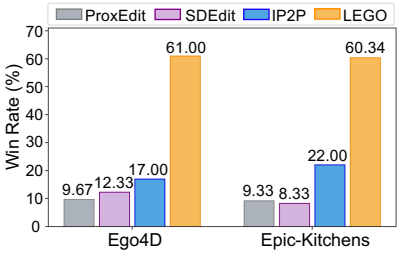

Fig. 4: User study of our model and baselines. Win rate is the percentage of each model being picked as the best.

text conditioning. ProxEdit and SDEdit rely on the off-the-shelf latent diffusion model to synthesize edited images and thus can not be trained end to end. Therefore, we first finetune the latent diffusion model with egocentric images and action labels and then use the finetuned latent diffusion model parameters to implement ProxEdit and SDEdit approaches. Please refer to the supplementary for more baseline implementation details and a thorough comparison with prior methods that leverage LLMs for image generation.

In terms of image-to-image metrics in Tab. 1, both ProxEdit and SDEdit perform poorly on this novel problem, suggesting the challenging nature of generating egocentric action frames. IP2P performs much better in almost all metrics by end-to-end training and serves as the strongest baseline model in our experiments. However, our LEGO model consistently outperforms IP2P by a large margin (3.46%, 1.60%, 1.86%, 0.90, 0.13 and 0.73% respectively) in all six metrics on Ego4D. LEGO also exceeds IP2P notably (2.92%, 1.60%, 1.60%, 0.10 and 0.46% respectively) in five metrics on Epic-Kitchens. Though, LEGO slightly lags behind IP2P in FID score, it still achieves the second best performance.

With regard to image-to-text metrics in Tab. 2, LEGO outperforms the strongest baseline (IP2P) by 0.38% and 0.14% on Ego4D and by 1.61% and

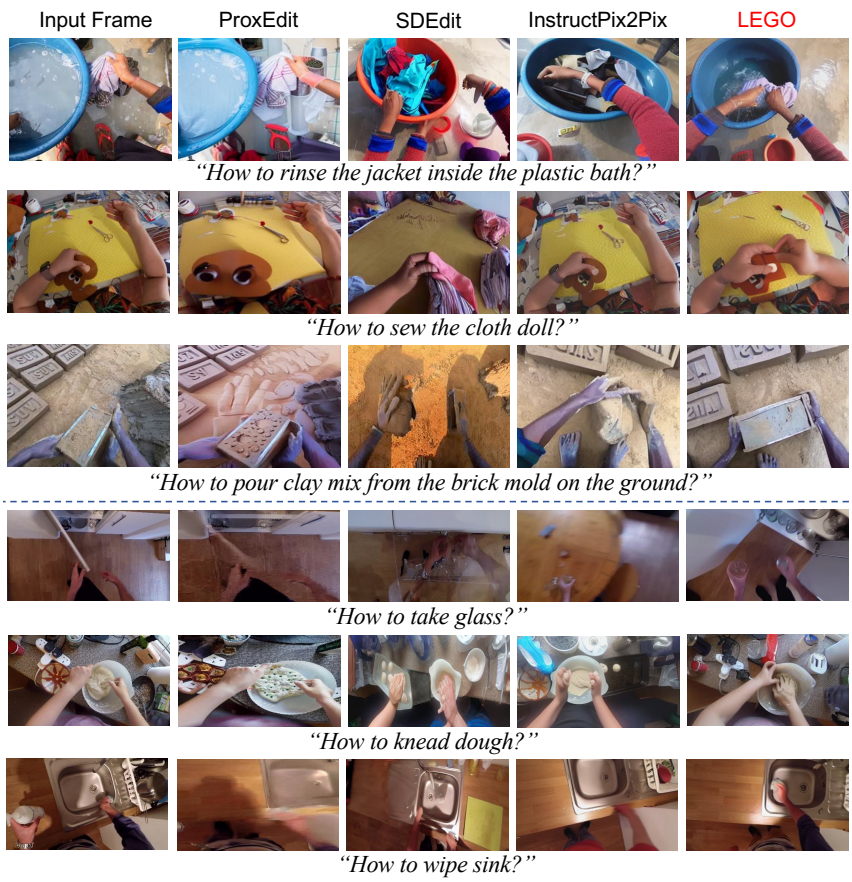

Fig. 5: Visualization of the proposed LEGO model and all baselines on Ego4D (the first three rows) and Epic-Kitchens (the last three rows). The action frames generated by LEGO align with the user prompt better and preserve more contexts in input frames.

1.05% on Epic-Kitchens. The result suggests the synthetic frames from our model can better align with the action descriptions. However, the performance gain is rather limited. We emphasize that this is because the domain gap still exists for the BLIP model. Besides baseline models, we also measure the image-to-text similarity of input frames and action prompts to validate models' capability of learning action state transition in *semantics*. The BLIP-B/BLIP-L scores are 15.44%/15.49% on Ego4D and 20.11%/20.52% on Epic-Kitchens, lagging behind all baseline models. The result evidences diffusion models are able to understand actions and edit the input frame *semantically* towards the action prompt.

Due to the limitation of automatic metrics, we additionally implement user study as a complement. We shuffle the order of the results from our model and the baselines while presenting them to the raters. We define the win rate as the percentage of each model being picked as the best out of the total 300

Fig. 6: Additional visualization of our proposed model. LEGO is able to synthesize faithful action frames as well as preserve the contexts in various scenarios.

samples. Results are illustrated in Fig. 4. Our model surpasses the best baseline by 44.00% and 38.34% respectively on Ego4D and Epic-Kitchens. The prominent gap further validates the superiority of our model.

4.3 Visualization of Generated Frames

We additionally showcase examples of generated image from LEGO and baseline models in Fig. 5. ProxEdit and SDEdit fail to understand the user prompt and thus generate frames of irrelevant actions (*e.g.*, row3). They may also easily synthesize the action in a different environment (*e.g.*, row2). InstructPix2Pix is able to preserve more contexts but still fails to semantically align with the action in user prompt (*e.g.*, row1). In contrast, LEGO can synthesize action frames that better align with the user prompts and retain more contexts in the background. More examples of LEGO output are presented in Fig. 6 and supplementary.

4.4 Ablation Study

We present comprehensive ablation studies to investigate the contribution of enriched action descriptions, the VLLM image embedding and the VLLM text embedding separately. Results are demonstrated in Tab. 3. Given the limitation of automatic metrics, we also provide user study as the *primary* metric. Notably, conditioning on enriched action descriptions can moderately improve the model performance, supporting our claim that expanding the action description can facilitate the learning of the state transition during an egocentric action. Moreover, utilizing the image embedding and text embedding from VLLM as additional conditions both improve the model performance by a notable margin because VLLM embeddings can effectively narrow the domain gap. Interestingly, the image embedding leads to larger performance boost on both datasets. These results suggest the VLLM image embedding $\mathcal{H}_i$ incorporates important high-level semantic meanings that are not captured in the auto-encoder image embedding $\mathcal{E}(\mathcal{X})$ or VLLM text emebdding $\mathcal{H}_t$. Finally, our full LEGO model uses both VLLM image and text embeddings (Desc.+Joint Embed.) and thus achieves the best performance on both datasets in all metrics.

Table 3: Analysis of egocentric action frame generation performance with different conditionings. Joint Embed. refers to incorporating both VLLM image and text embeddings. Similar to Fig. 4, we present win rate as the user study result, *i.e.*, the percentage of each model picked as the best (% is omitted for simplicity). The best results are highlighted with **boldface**. The orange rows refer to our full LEGO model.

Conditioning	Ego4D				Epic-Kitchens			
	User Study	EgoVLP	EgoVLP$^+$	CLIP	User Study	EgoVLP	EgoVLP$^+$	CLIP
Actions Labels	5.33	62.19	78.84	78.75	7.08	42.97	61.06	77.03
Descriptions	13.00	62.91	79.09	79.18	12.50	43.72	61.46	77.47
Desc. + Img Embed.	26.00	65.35	80.13	80.57	24.17	45.82	62.29	78.60
Desc. + Txt Embed.	21.33	63.29	79.40	79.21	22.08	44.68	62.02	77.74
Desc. + Joint Embed.	**34.34**	**65.65**	**80.44**	**80.61**	**34.17**	**45.89**	**62.66**	**78.63**

Table 4: Performance of LEGO without finetuning (w/o FT). The best results are highlighted with **boldface**. The orange rows refer to our full LEGO model.

Conditioning	Ego4D			Epic-Kitchens		
	EgoVLP	EgoVLP$^+$	CLIP	EgoVLP	EgoVLP$^+$	CLIP
Descriptions	62.91	79.09	79.18	43.72	61.46	77.47
Desc.+Joint Embed.(w/o FT)	64.57	79.72	80.23	44.74	61.87	78.35
Desc.+Joint Embed.(w/ FT)	**65.65**	**80.44**	**80.61**	**45.89**	**62.66**	**78.63**

4.5 Analysis of Visual Instruction Tuning

First, we assess the quality of enriched action descriptions from the visual instruction tuned VLLM by user study. For each sample in user study, we ask the raters to select whether the description aligns with the input and target frames (see supplementary for more details of user study setting and interface). The percentage of samples with aligned frames and action descriptions is **87%** on Ego4D and **84%** on Epic-Kitchens. The high alignment percentage suggests the visual instruction tuned VLLM can effectively expand action labels with details captured in the input frame. We additionally implement the same user study for *off-the-shelf* VLLM (*i.e.*, without finetuning). The percentage of alignment is just **27%** on Ego4D and **30%** on Epic-Kitchens with hallucination existing in **92%** of unaligned instances. The notable drop supports our argument that visual instruction tuning is a critical step for high-quality prompt enrichment.

In addition, we further investigate whether the visual instruction tuned embeddings can contribute more to the diffusion model than the off-the-self counterpart, which has not been well studied in prior work. As shown in Tab. 4, without any finetuning, the image and text embeddings from VLLM can still improve the baseline model (Descriptions). However finetuned embeddings yield much larger improvement (*e.g.*, 1.08% and 1.15% in EgoVLP score) on both datasets. The result suggests that visual instruction tuning is a necessary step to learn semantic image and text embeddings that are more aligned with the

Fig. 7: Visualization of generating various actions from the same input frame. The first action (*"open drawer"*) is the existing label for this example in the dataset. We fill another five actions into the user query and synthesize the action frames using our model. All generated frames align well with the actions and preserve most contexts.

egocentric input frame and action prompt, which thus narrows the domain gap and greatly boosts the performance of the latent diffusion model.

4.6 Generation of Various Actions with the Same Contexts

In addition to generating action frames based on the pre-defined action labels in our datasets, we validate the generative nature of our model *i.e.*, whether LEGO is able to synthesize the correct action frames conditioning on the same contexts yet different actions (novel image-action pairs). Results are illustrated in Fig. 7. We feed different actions with the same input frame to our model. The synthesized frames correctly depict the execution of the actions in the same environment. This result further indicates that our model can understand the action state transition and generalize to different user queries.

5 Conclusion

In this paper, we introduce the novel problem of egocentric action frame generation. We also propose a novel model — LEGO that leverages visual instruction tuning and a diffusion model to address this problem. Our key intuition is to use visual instruction tuning to generate informative responses that depict the execution of the egocentric action, and then design the conditioning for the denoising U-Net to exploit the image and text embeddings from a visual instruction tuned VLLM. Our experiments on two large-scale egocentric action datasets validate the advantage of the proposed approach as well as the contribution of each model component. We believe our work provides an important step in understanding the action state transition and controllability of diffusion models, and suggests future research directions in egocentric AI systems, action state transition, image generation and human skill transfer.

Acknowledgements

Portions of this work were supported in part by a gift from Meta. We thank Sangmin Lee for the valuable discussion and suggestions.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Anonymity: Making multimodal generation easier: When diffusion models meet llms. openreview (2023)
3. Ashutosh, K., Girdhar, R., Torresani, L., Grauman, K.: Hiervl: Learning hierarchical video-language embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23066–23078 (2023)
4. Baskin, J.H., Edersheim, J.G., Price, B.H.: Is a picture worth a thousand words? neuroimaging in the courtroom. American Journal of Law & Medicine **33**(2-3), 239–269 (2007)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Chakrabarty, T., Singh, K., Saakyan, A., Muresan, S.: Learning to follow object-centric image editing instructions faithfully. arXiv preprint arXiv:2310.19145 (2023)
8. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
9. Chen, W.G., Spiridonova, I., Yang, J., Gao, J., Li, C.: Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. arXiv preprint arXiv:2311.00571 (2023)
10. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
11. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
12. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems (2023)
13. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision pp. 1–23 (2022)
14. Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J.B., Schuurmans, D., Abbeel, P.: Learning universal policies via text-guided video generation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)

15. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation. In: Advances in Neural Information Processing Systems (2023)
16. Girdhar, R., Grauman, K.: Anticipative video transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13505–13515 (2021)
17. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
18. Goel, V., Peruzzo, E., Jiang, Y., Xu, D., Sebe, N., Darrell, T., Wang, Z., Shi, H.: Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. arXiv preprint arXiv:2303.17546 (2023)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
20. Goyal, M., Modi, S., Goyal, R., Gupta, S.: Human hands as probes for interactive object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3293–3303 (2022)
21. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
22. Hafri, A., Trueswell, J.C., Epstein, R.A.: Neural representations of observed actions generalize across static and dynamic visual input. Journal of Neuroscience **37**(11), 3056–3071 (2017)
23. Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023)
24. Han, L., Wen, S., Chen, Q., Zhang, Z., Song, K., Ren, M., Gao, R., Stathopoulos, A., He, X., Chen, Y., et al.: Proxedit: Improving tuning-free real image editing with proximal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4291–4301 (2024)
25. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
26. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
27. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
28. Huang, J., Liu, Y., Qin, J., Chen, S.: Kv inversion: Kv embeddings learning for text-conditioned real image action editing. arXiv preprint arXiv:2309.16608 (2023)
29. Huang, Y., Cai, M., Li, Z., Lu, F., Sato, Y.: Mutual context network for jointly estimating egocentric gaze and action. IEEE Transactions on Image Processing **29**, 7795–7806 (2020)
30. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Proceedings of the European conference on computer vision (ECCV). pp. 754–769 (2018)

31. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
32. Jia, W., Liu, M., Rehg, J.M.: Generative adversarial network for future hand segmentation from egocentric video. In: European Conference on Computer Vision. pp. 639–656. Springer (2022)
33. Jiang, Y., Zhang, Z., Xue, T., Gu, J.: Autodir: Automatic all-in-one image restoration with latent diffusion. arXiv preprint arXiv:2310.10123 (2023)
34. Joseph, K., Udhayanan, P., Shukla, T., Agarwal, A., Karanam, S., Goswami, K., Srinivasan, B.V.: Iterative multi-granular image editing using diffusion models. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
35. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. In: Advances in Neural Information Processing Systems (2022)
36. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
37. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019)
38. Kim, S., Jang, W., Kim, H., Kim, J., Choi, Y., Kim, S., Lee, G.: User-friendly image editing with minimal text input: Leveraging captioning and injection techniques. arXiv preprint arXiv:2306.02717 (2023)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
40. Koh, J.Y., Fried, D., Salakhutdinov, R.R.: Generating images with multimodal language models. Advances in Neural Information Processing Systems **36** (2024)
41. Lai, B., Liu, M., Ryan, F., Rehg, J.M.: In the eye of transformer: Global-local correlation for egocentric gaze estimation. British Machine Vision Conference (2022)
42. Lai, B., Liu, M., Ryan, F., Rehg, J.M.: In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. International Journal of Computer Vision **132**(3), 854–871 (2024)
43. Lai, B., Ryan, F., Jia, W., Liu, M., Rehg, J.M.: Listen to look into the future: Audio-visual egocentric gaze anticipation. arXiv preprint arXiv:2305.03907 (2023)
44. Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems (2023)
45. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)
46. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. International Conference on Machine Learning (2023)
47. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

48. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European conference on computer vision (ECCV). pp. 619–635 (2018)
49. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)
50. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems **35**, 7575–7586 (2022)
51. Liu, B., Zhang, H., Liu, J., Wang, Q.: Acigs: An automated large-scale crops image generation system based on large visual language multi-modal models. In: 2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). pp. 7–13. IEEE (2023)
52. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems (2023)
53. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 704–721. Springer (2020)
54. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3282–3292 (2022)
55. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. Advances in Neural Information Processing Systems **34**, 25019–25032 (2021)
56. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
57. Mirzaei, A., Aumentado-Armstrong, T., Brubaker, M.A., Kelly, J., Levinshtein, A., Derpanis, K.G., Gilitschenski, I.: Watch your steps: Local image and scene editing by text instructions. arXiv preprint arXiv:2308.08947 (2023)
58. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
59. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329 (2023)
60. Orgad, H., Kawar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. arXiv preprint arXiv:2303.08084 (2023)
61. Pan, Z., Gherardi, R., Xie, X., Huang, S.: Effective real image editing with accelerated iterative diffusion inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15912–15921 (2023)
62. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5285–5297 (2023)
63. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

64. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
65. Ragusa, F., Farinella, G.M., Furnari, A.: Stillfast: An end-to-end approach for short-term object interaction anticipation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3635–3644 (2023)
66. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Naq: Leveraging narrations as queries to supervise episodic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6694–6703 (2023)
67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
68. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
69. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14663–14674 (2023)
70. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580 (2023)
71. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
72. Stein, G., Cresswell, J.C., Hosseinzadeh, R., Sui, Y., Ross, B.L., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, J.E.T., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. arXiv preprint arXiv:2306.04675 (2023)
73. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023)
74. Sudhakaran, S., Escalera, S., Lanz, O.: Lsta: Long short-term attention for egocentric action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9954–9963 (2019)
75. Sun, Z., Zhou, Y., He, H., Mok, P.: Sgdiff: A style guided diffusion model for fashion synthesis. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8433–8442 (2023)
76. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
77. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., De la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
78. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
79. Tsaban, L., Passos, A.: Ledits: Real image editing with ddpm inversion and semantic guidance. arXiv preprint arXiv:2307.00522 (2023)
80. Wallace, B., Gokul, A., Naik, N.: Edict: Exact diffusion inversion via coupled transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22532–22541 (2023)

81. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023)
82. Wang, K., Yang, F., Yang, S., Butt, M.A., van de Weijer, J.: Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. arXiv preprint arXiv:2309.15664 (2023)
83. Wang, Q., Zhang, B., Birsak, M., Wonka, P.: Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. arXiv preprint arXiv:2305.18047 (2023)
84. Wang, Q., Zhang, B., Birsak, M., Wonka, P.: Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path. arXiv preprint arXiv:2303.16765 (2023)
85. Wang, W., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023)
86. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8168–8177 (2021)
87. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
88. Wen, S., Fang, G., Zhang, R., Gao, P., Dong, H., Metaxas, D.: Improving compositional text-to-image generation with large vision-language models. arXiv preprint arXiv:2310.06311 (2023)
89. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
90. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)
91. Xu, Y., Li, Y.L., Huang, Z., Liu, M.X., Lu, C., Tai, Y.W., Tang, C.K.: Egopca: A new framework for egocentric hand-object interaction understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5273–5284 (2023)
92. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22479–22489 (2023)
93. Yu, Q., Li, J., Ye, W., Tang, S., Zhuang, Y.: Interactive data synthesis for systematic vision adaptation via llms-aigcs collaboration. arXiv preprint arXiv:2305.12799 (2023)
94. Yu, Z., Li, H., Fu, F., Miao, X., Cui, B.: Fisedit: Accelerating text-to-image editing via cache-enabled sparse diffusion inference. arXiv preprint arXiv:2305.17423 (2023)
95. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
96. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4372–4381 (2017)

97. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
98. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: Hive: Harnessing human feedback for instructional visual editing. arXiv preprint arXiv:2303.09618 (2023)
99. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
100. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10156 (2023)
101. Zhang, Z., Han, L., Ghosh, A., Metaxas, D.N., Ren, J.: Sine: Single image editing with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6027–6037 (2023)
102. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)