# VideoGLaMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos

Shehan Munasinghe[1] Hanan Gani[1] Wenqi Zhu[2] Jiale Cao [2]
Eric Xing[1,3] Fahad Shahbaz Khan[1,4] Salman Khan[1,5]

[1]Mohamed bin Zayed University of AI, [2]Tianjin University, [3]Carnegie Mellon University,
[4]Linköping University, [5]Australian National University

shehan.munasinghe@mbzuai.ac.ae, hanan.ghani@mbzuai.ac.ae

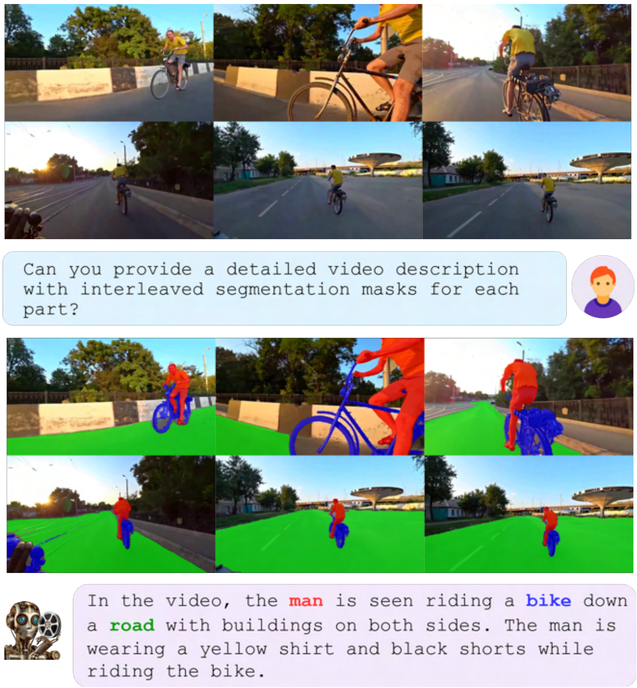
https://mbzuai-oryx.github.io/VideoGLaMM

## Abstract

*Fine-grained alignment between videos and text is challenging due to complex spatial and temporal dynamics in videos. Existing video-based Large Multimodal Models (LMMs) handle basic conversations but struggle with precise pixel-level grounding in videos. To address this, we introduce VideoGLaMM, a LMM designed for fine-grained pixel-level grounding in videos based on user-provided textual inputs. Our design seamlessly connects three key components: a Large Language Model, a dual vision encoder that emphasizes both spatial and temporal details, and a spatio-temporal decoder for accurate mask generation. This connection is facilitated via tunable V→L and L→V adapters that enable close Vision-Language (VL) alignment. The architecture is trained to synchronize both spatial and temporal elements of video content with textual instructions. To enable fine-grained grounding, we curate a multimodal dataset featuring detailed visually-grounded conversations using a semiautomatic annotation pipeline, resulting in a diverse set of 38k video-QA triplets along with 83k objects and 671k masks. We evaluate VideoGLaMM on three challenging tasks: Grounded Conversation Generation, Visual Grounding, and Referring Video Segmentation. Experimental results show that our model consistently outperforms existing approaches across all three tasks.*

## 1. Introduction

The rise of Large Language Models (LLMs) has significantly advanced progress in language-based tasks [7, 10, 13, 35, 47]. Their success in solving language-based complex reasoning tasks has led to their adoption in visual domains, resulting in Large Multimodal Models (LMMs). To align textual and visual modalities, previous works [11, 20, 21, 28, 57] train a projection layer or a cross-attention block that maps visual features into the latent space of an LLM. This straightforward adaptation has enabled advanced spatial understanding, allowing detailed conversations about image content. Recently, these models have been extended to video, aligning textual instructions with the spatio-temporal



Can you provide a detailed video description with interleaved segmentation masks for each part?

In the video, the **man** is seen riding a **bike** down a **road** with buildings on both sides. The man is wearing a yellow shirt and black shorts while riding the bike.

Figure 1. **Grounded Conversation with VideoGLaMM.** Our proposed multimodal video conversational model provides text responses grounded at the pixel level in the input video. The generated masks are spatio-temporally consistent across frames. The fine-grained grounded outputs from VideoGLaMM describe different levels of granularity, e.g., person, objects (bike), stuff (road), and explain object and scene attributes. Existing Video-LMMs do not offer pixel-level grounded conversational capability.

inputs, leading to the development of Video-LMMs.

Existing Video-LMMs [9, 22, 24, 30–32, 52], similar to image-based LMMs, tune single or multiple projection layers to align videos with the language modality using the conventional visual instruction tuning paradigm. Although this simple alignment aids in understanding the global content of videos, it poses challenges in capturing localized object-specific context. Consequently, existing works [22, 24, 32, 52] have demonstrated capabilities in video comprehension and dialogue, they lack the crucial feature of fine-grained visual grounding, which aims to associate the LMM's response to specific objects within the video input. The ability of an LMM to generate visually grounded responses ensures that the model understands fine-grained spatial and temporal details in a video and can relate them with the generated text.

To bridge this gap, we introduce VideoGLaMM, a large video multimodal model capable of pixel-level spatio-temporal grounding. The model responds to natural language queries from the user and intertwines spatio-temporal object masks in its generated textual responses to provide a detailed understanding of video content. VideoGLaMM seamlessly connects three key components: a Large Language Model (LLM); dual vision encoders; and a spatio-temporal pixel decoder. The dual vision encoders extract spatial and temporal features separately, which are jointly passed to the LLM to output responses rich in both spatial and temporal cues. Our spatio-temporal pixel decoder outputs the fine-grained object masks corresponding to the specific objects in the LLM output to visually ground its responses. These components are integrated via tunable Vision-to-Language (V→L) and Language-to-Vision (L→V) adapters that enable close vision-language alignment, unlike existing works that perform alignment with a single adapter.

As there currently exists no instruction-tuning dataset with fine-grained masks associated with video conversations, we present a benchmark instruction tuning dataset curated through a semi-automatic pipeline (Sec. 4). The dataset consists of 38k grounded video-QA triplet pairs with 83k objects and 671k fine-grained masks. The proposed benchmark dataset enables spatio-temporal modeling and significantly augments the capacity of the model to understand videos comprehensively, leading to state-of-the-art performance in grounded conversation generation, temporal grounding, and referring video segmentation tasks under zero-shot settings.

In summary, our contributions are as follows:

- We introduce VideoGLaMM, a video large multimodal model, capable of pixel-level spatio-temporal grounding, featuring an end-to-end alignment mechanism.
- To achieve fine-grained spatio-temporal alignment, we introduce a benchmark instruction tuning dataset consisting of 38k grounded video-QA triplet pairs and 83k objects and roughly 671k fine-grained spatio-temporal masks.
- We assess the performance of VideoGLaMM across diverse tasks spanning grounded conversation generation (GCG), visual grounding, and referring video segmentation, where it achieves state-of-the-art performance.

## 2. Related work

**Large Multi-modal Models (LMMs).** Vision-language models like [37] have made notable advancements, demonstrating impressive zero-shot capabilities using millions of noisy image-text pairs during training. These models have been effective in various applications, from detection and segmentation [6, 23] to more complex tasks such as 3D understanding and video analysis [29, 33, 43, 48]. The rise of LLMs has driven significant progress in Natural Language Processing (NLP) tasks and sparked interest in developing LMMs. Early models [2, 4] incorporate visual information into intermediate embeddings for a frozen LLM using a cross-attention mechanism, trained on billions of image-text pairs to align visual and linguistic modalities. Similarly, BLIP-2 [21] introduces Q-Former to better align visual features with language space. MiniGPT-4 [57] and LLAVA [28] finetune on detailed image descriptions using a single projection layer to align a frozen visual encoder with a frozen LLM. Subsequent LLaVA series models [26] employ a multi-layer perceptron and a two-stage instruction tuning to refine the alignment process. While these works work on static images, our work focuses on efficiently aligning videos with linguistic cues.

**Video LMMs.** Recent advancements in image-based multimodal models have paved the way for video LMMs, which are essential for handling spatiotemporal sequences. Models such as VideoChat [22], Video-LLaMA, Video-ChatGPT [52], Video-LLAVA [24] and Video-GPT+ [31] extend the capabilities of LLMs to video domain by aligning video features with language, followed by instruction tuning on datasets annotated by either GPT models or humans. While these models have shown effectiveness in video comprehension, they still face limitations in fine-grained spatio-temporal modeling and visual grounding. This restricts their ability to accurately understand or localize specific objects and detailed segments within videos, highlighting the need for further advancements in developing better multimodal models capable of visual grounding.

**Visual Grounding.** Recently Grounded LMMs [9, 20, 36, 39, 49, 51, 54] have made significant strides in enhancing visual and language comprehension and excel in complex localization tasks. These models demonstrate proficiency in tasks such as referring expression comprehension and image segmentation, highlighting the advanced image understanding capabilities of LLMs. Approaches such as [9, 36, 49] primarily focus on creating a language-

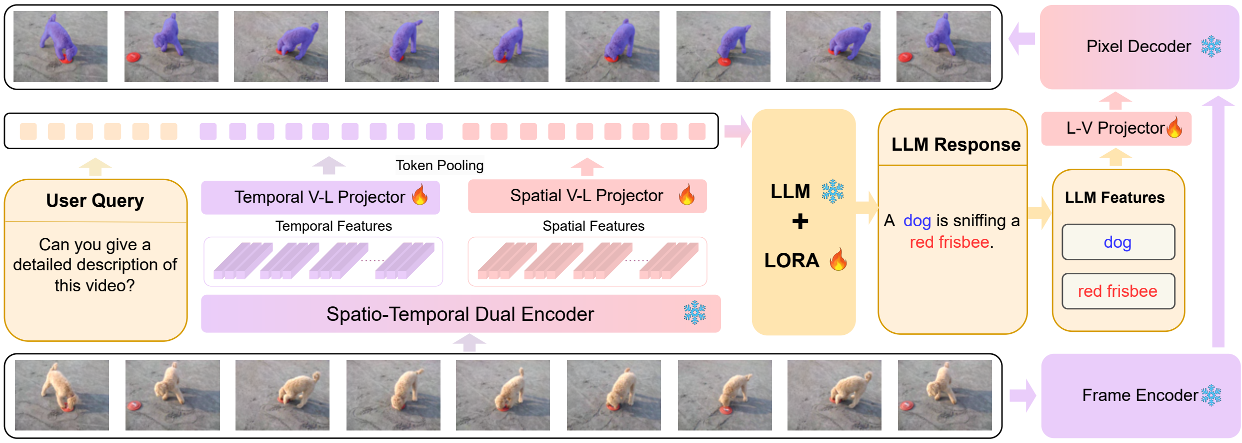


Figure 2. **Working of VideoGLaMM**. VideoGLaMM consists of a dual spatio-temporal encoder for encoding image and video level features. The spatial features represent the local information and the temporal features represent global information. The spatial and temporal tokens are passed through V-L adapters and concatenated with the text tokens, before feeding to LLM. A L-V projector is employed to align LLM's response with the visual space of pixel decoder. Finally, the aligned LLM features along with the frame features from a frame encoder are passed to a grounded pixel decoder, to obtain the fine-grained object masks corresponding to the LLM response.

based context for visual grounding. In contrast, [54] integrates visual elements with language, while [20] leverages vision-language embeddings to produce segmentation masks. Additionally, [39] is adept at generating natural language responses linked with object segmentation masks, facilitating detailed visual-textual interactions. However, these models are limited to image-based applications and do not extend to video understanding. Recently, [32] incorporates audio transcripts alongside visual and textual data for a more detailed video understanding. However, it combines pre-trained modules that cannot be trained end-to-end, which results in lack of fine-grained spatiotemporal modeling. Similarly, [5] introduced a new video grounding model, but their architecture employs only a spatial encoder-decoder setup and does not address the GCG task. To this end, we propose VideoGLaMM, which leverages a novel fine-grained alignment strategy to align language instruction across both spatial and temporal dimensions, facilitating more finegrained video understanding.

# 3. VideoGLaMM

## 3.1. Overview

In this work, we introduce VideoGLaMM, a multi-modal video LMM with spatio-temporal pixel grounding capability. The task of spatio-temporal visual grounding focuses on linking a model's response to a user-specific text query with particular objects and regions within a video, ensuring that both spatial (what's happening in each frame) and temporal (how things change over time) details are accurately reflected in the generated output (Fig. 1). By grounding responses in specific objects and actions across frames, the model demonstrates an understanding of both the evolving and static elements in a video, enabling it to produce responses that align closely with the visual narrative.

Our proposed VideoGLaMM is designed to achieve effective spatio-temporal grounding due to its ability to process spatial and temporal features simultaneously. VideoGLaMM's architecture (Fig. 2) leverages a dual-encoder structure: one encoder focuses on extracting spatial details from images, while the other captures temporal information from video sequences, ensuring complementary representation from both modalities. The visual features from both encoders are then integrated with a LLM using separate spatial and temporal adapters (V→L), guided by specific textual instructions. The LLM outputs are aligned back with the visual space using an L→V adapter and further processed by a pixel decoder, which also takes video frames as input to produce the final grounded outputs.

For end-to-end spatio-temporal alignment, we train VideoGLaMM on our proposed fine-grained benchmark dataset. During training, we finetune the LoRA parameters of the LLM, along with V→L and L→V adapters. This approach seamlessly combines spatial and temporal data through an improved alignment mechanism and a precise grounding framework, enhancing the model's capability for visual grounding and understanding.

## 3.2. Architecture

The overall architecture of our VideoGLaMM consists of following components: (i) Spatio-Temporal Dual Encoder, (ii) Dual Alignment V-L Adapters, (iii) Large Language Model (LLM), (iv) Pixel Decoder. Below we provide a detailed description and working of each of component.

**Spatio-Temporal Dual Encoder.** Our architecture consists of separate image and video encoders for extracting spatial and temporal features, thus leveraging the complementary strengths of both. This enables the model to have both local and global properties. The image encoder $\mathcal{F}_g$, processes the $T$ video frames separately such that the input video $V \in \mathbb{R}^{T\times H\times W\times C}$. The output of the image encoder, represented by $f_g$, produces local spatial features that provide frame-level context.

$$f_g = \mathcal{F}_g(V), \quad V \in \mathbb{R}^{T\times H\times W\times C} \tag{1}$$

Meanwhile, for extracting video features, we use segment-wise Sampling following [31] to obtain fine-grained temporal cues. Given an input video $V \in \mathbb{R}^{T\times H\times W\times C}$, we divide it into $K$ segments, where each segment consists of $s = \frac{T}{K}$ frames. The video encoder $\mathcal{F}_h$, operates on low-resolution video segments $V_k \in \mathbb{R}^{s\times H\times W\times C}$ yielding global features that provide segment-wise temporal context.

$$f_h = \mathcal{F}_h(V_k), \quad V_k \in \mathbb{R}^{s\times H\times W\times C} \tag{2}$$

**Dual Alignment (V→L) Adapters** To align visual features with the LLM space, we use two separate V→L adapters for image and video encoders. $\mathcal{W}_g$ represents the spatial adapter, and $\mathcal{W}_h$ represents the temporal adapter. These adapters project the visual features into the LLM's projection space, thus aligning the two modalities. The spatial and visual features corresponding to image and video samples after projecting from $W_g$ and $W_h$ are represented by $Z_g$ and $Z_h$, respectively.

$$Z_g = \mathcal{W}_g(f_g), \quad and \quad Z_h \quad = \mathcal{W}_h(f_h) \tag{3}$$

**Large Language Model** The tokenized spatio-temporal visual features are then concatenated with the textual tokens $Z_{text} \in \mathbb{R}^{L\times D_t}$ to obtain final feature embedding $\mathcal{Z} = [Z_g, Z_h, Z_{text}]$ which is fed into the **LLM**. Thus, input to the **LLM** contains both the spatial and temporal cues for robust video understanding. We further expand the original **LLM** vocabulary with a new token, i.e., <SEG>, which signifies the request for the segmentation output. Thus the **LLM** response **E** can be described as,

$$\mathbf{E} = \mathbf{LLM}(\mathcal{Z}) = \mathbf{LLM}([Z_g, Z_h, Z_{text}]) \tag{4}$$

The LLM output **E** contains the <SEG> whenever the task requires to generate the segmentation mask.

**Pixel Decoder** Our Pixel decoder consists of a prompt encoder ($\mathcal{H}$) and a mask decoder $\mathcal{D}$, capable of predicting masks with spatio-temporal grounding. The pixel decoder is adapted to videos and can implicitly process temporal information. The last layer embeddings from the LLM denoted as $l_{\text{seg}}$ corresponding to <SEG> token is extracted, which is enriched with both spatial and temporal cues. The LLM embeddings act as prompts for the mask decoder and are processed by the prompt encoder. Simultaneously, we extract visual features of the input frames $V$ using a grounded frame encoder $\mathcal{P}$ which is aligned with pixel decoder and is further equipped with the ability to produce multi-scale features during training. For aligning the output embeddings from LLM with the pixel decoder, we train an (L→V) adapter layer $\mathcal{W}_p$ between the LLM and prompt encoder such that the output from the adapter is denoted as $\mathbf{e}^p_{\text{seg}} = \mathcal{W}_p(l_{\text{seg}})$. The $\mathbf{e}^p_{\text{seg}}$ is fed to prompt encoder $\mathcal{H}$, such that the encoded output $\mathcal{H}(\mathbf{e}^p_{\text{seg}})$ is used to prompt the mask decoder. The encoded prompts $\mathcal{H}(\mathbf{e}^p_{\text{seg}})$ along with the grounded visual features $\mathcal{P}(V)$ are passed to mask decoder $\mathcal{D}$. Subsequently, $\mathcal{D}$ produces the output mask **M**.

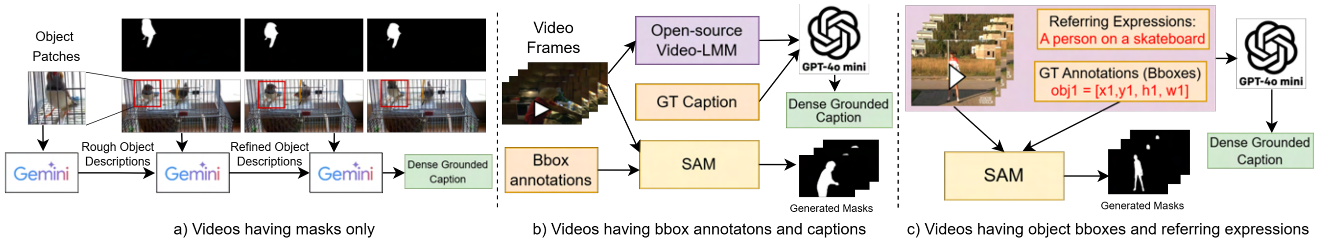
$$\mathbf{M} = \mathcal{D}\Big(\mathcal{P}(V), \mathcal{H}(\mathbf{e}^p_{\text{seg}})\Big) \tag{5}$$

### 3.3. Training Strategy

We train VideoGLaMM end-to-end in a single stage. As stated above, we use a dual encoder consisting of separate image and video encoders for processing spatial and temporal inputs to obtain local and global features, respectively. These encoders are initialized with weights of strong pre-trained encoders. During training, we keep the encoders fixed and only train the V→L adapters $\mathcal{W}_g$ and $\mathcal{W}_h$ associated with these encoders. These adapters are used to project the spatio-temporal visual features in the space of LLM and align the two modules. The spatio-temporal encoder is kept frozen and only the V→L adapters are updated. The textual features from the last layer of LLM, rich in spatial and temporal cues, are projected into the space of the pixel decoder using a multi-layer projection L→V adapter $\mathcal{W}_p$. For the LLM, we keep its weights frozen and only finetune LoRA [14] parameters during training. Both the frame encoder and pixel decoder are instantiated with pre-trained weights We keep the frame encoder and pixel decoder frozen and only train the L→V adapter layer. We optimize the output of the LLM by minimizing the cross entropy **CE** objective between the autoregressively obtained text output and dense grounded ground-truth caption. For the output of mask decoder, we optimize the intersection over union (IOU) between the predictions of mask decoder and ground-truth masks denoted as $\mathcal{L}_{masked}$. The total loss is the sum of **CE** loss and masked loss.

$$\mathcal{L}_{total} = \mathbf{CE} + \mathcal{L}_{masked} \tag{6}$$

The first component of $\mathcal{L}_{total}$ ensures that the LLM generates textual embeddings that not only align with the ground truth but also offer informative spatio-temporal cues to the mask decoder for effective grounding. The second component facilitates efficient grounding by leveraging these textual cues from the LLM.



a) Videos having masks only

b) Videos having bbox annotatons and captions

c) Videos having object bboxes and referring expressions

Figure 3. **Proposed Semi-automatic Annotation Pipeline**. Our dataset for grounded conversation generation (GCG) is built from three video dataset types: i) *Videos having masks only:* Object patches are extracted from video frames using masks and processed by the Gemini-Pro model for initial object descriptions, which are then refined to produce detailed object captions. These refined captions and masks are again fed to Gemini-Pro model to create dense grounded captions. ii) *Videos having bbox annotations and captions:* Frames are first processed with a Video-LMM to generate a comprehensive caption which is combined with the original caption and fed to GPT-4o to obtain dense grounded captions. Masks are generated using frames and ground-truth bounding boxes with the SAM model. iii) *Videos having object bboxes and referring expressions:* Frames, bounding boxes, and referring expressions are input to GPT-4o for dense grounded captions, while masks are generated by feeding frames and bounding boxes to the SAM model.

## 4. Our Benchmark & Annotation Pipeline

Our benchmark video dataset comes from different sources: YTVIS [16], BURST [3] ActivityNet entities [56], Refer-YTVOS [44], MeViS [12], VidSTG [53] and HCSTVG [46]. To create fine-grained grounded captions, we develop a semi-automated pipeline (Fig. 3) that ensures high-quality and scalable annotation. Our annotation pipeline is categorized into three streams based on the availability of the ground truth annotations. We explain each stream below.

**a) Videos with only Mask annotations:** Fig. 3(a) shows the annotation process for the videos having only masks as ground truth labels. To generate the corresponding dense grounded caption, we use following steps: **i)** *Object Description Generation:* For each object in the video, we begin by creating a bounding box based on the ground truth mask provided in the annotation file. This bounding box allows us to crop the object from each frame, producing a sequence of image patches that capture the object throughout the video. We then feed these image patches to the Gemini-Pro model [42] to obtain a rough description of each object in the video. **ii)** *Object Description Refinement:* The bounding boxes from the previous stage are superimposed on the corresponding video frames, and the entire video is then fed into the Gemini-Pro model to obtain a more accurate and detailed description of the objects. **iii)** *Caption Generation:* The bounding boxes of corresponding objects overlayed across the video frames are labeled according to their object IDs. Then, we input these frames into the Gemini-Pro model to obtain dense captions. This results in a comprehensive description of the video. Finally, we manually review the {obj_id} in the generated video captions based on the video content. **iv)** *Detailed Dense Captions.* To enhance the detail and accuracy of the video captions, we leverage two advanced Video LMMs: Video-LLAVA [52] and LLAVA-NeXT [27]. Using the semi-automatically generated captions as a reference, we integrate and refine the outputs from these models, merging their results to produce the final, comprehensive dense captions.

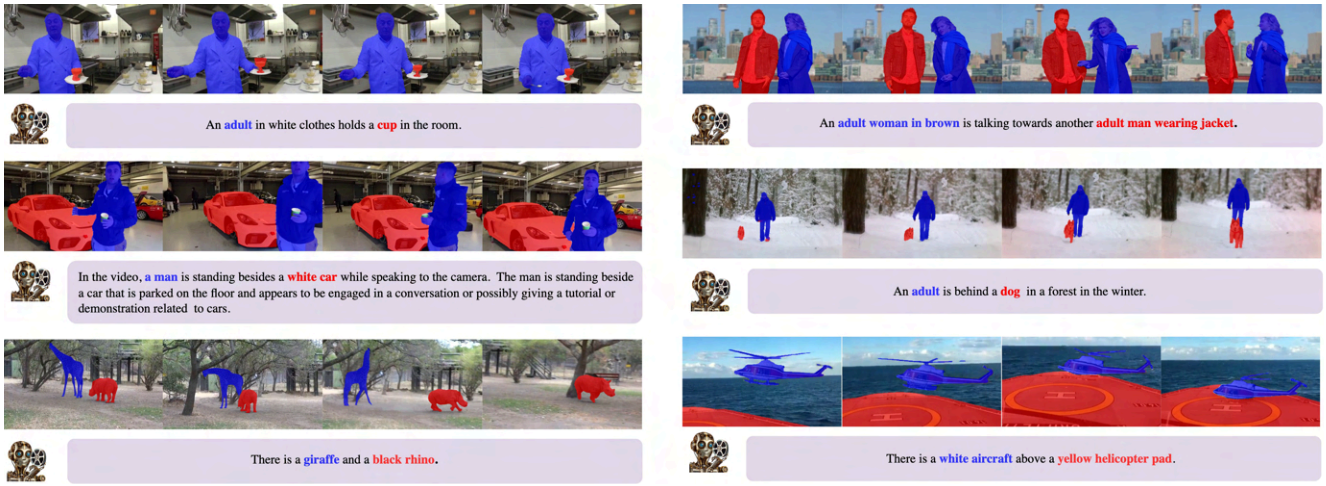
**b) Videos with Bounding Box annotations and Captions:** Fig. 3(b) shows the annotation process for the videos having both captions and object bounding box (Bbox) annotations. To obtain the corresponding dense grounded caption, the video frames are first passed to an open-source Video-LMM [27] to obtain a detailed caption, which is fed along with the reference ground truth caption to GPT-4o mini [34] to obtain the final dense grounded caption. The Bbox annotations are used as a prompts to SAM model [19] which takes the video frames as input and provides the masks corresponding to the objects.

**c) Videos with Bounding Box annotations and Referring Expressions:** Fig. 3(c) shows the annotation process for the videos having object bounding box (Bbox) annotations and referring expressions corresponding to different objects. The video frames along with Referring expressions and Bbox annotations are prompted to GPT-4o mini, which provides the corresponding dense grounded caption. To obtain the masks corresponding to the objects, the video frames are fed to SAM model, which is prompted with Bbox annotations of the objects.

Overall, our proposed GCG dataset has 38,788 grounded video-QA triplets along with 83,877 objects and 6,71,016 fine-grained masks in total. We further curate a separate test set of 308 refined video-QA triplets with 826 objects and 22762 finegrained masks for grounded conversation generation evaluation task.

## 5. Experimental Setup

**Implementation details.** Our spatio-temporal dual encoders follow the design of image and video encoders from [31]. For the image encoder, we use a pretrained CLIP ViT-L/14 (336 × 336)[37] model, and for the temporal encoder, we select the pretrained encoder of InternVideov2 (224 ×


An adult in white clothes holds a cup in the room.

In the video, a man is standing besides a white car while speaking to the camera. The man is standing beside a car that is parked on the floor and appears to be engaged in a conversation or possibly giving a tutorial or demonstration related to cars.

There is a giraffe and a black rhino.

An adult woman in brown is talking towards another adult man wearing jacket.

An adult is behind a dog in a forest in the winter.

There is a white aircraft above a yellow helicopter pad.

Figure 4. **Qualitative results of VideoGLaMM on grounded conversation generation (GCG).** Given user queries, the VideoGLaMM generates textual responses and grounds objects and phrases using pixel-level masks, showing its detailed understanding of the video.

224) [50]. The V→L projectors are initialized with the weights of MLP adapter from [31]. The LLM is instantiated with Phi3-Mini-3.8B [1] weights. Both the frame encoder and pixel decoder are initialized with SAM2 [41] encoder-decoder weights. The training (Sec. 3.3) is carried out end-to-end on 4 Nvidia A100 40GB GPUs with a distributed training based on DeepSpeed [40].

**Datasets.** We train the model on our proposed grounded conversation (GCG) dataset containing 38k grounded video-QA triplets along with 83k objects and 671k fine-grained masks. During training, we also include a variety of other image and video segmentation datasets with our proposed benchmark dataset for more robust alignment. Our choice of image segmentation datasets include: ADE20K [55], COCO-Stuff [8], LVIS-PACO [38], refCOCO, refCOCO+, refCLEF, refCOCOg [15], LLaVA-Instruct-150k [25], ReasonSeg [20] and GranDf [39]. For video egmentation datasets we include train samples from Refer-DAVIS17 [17] and VideoInstruct100K [30].

**Tasks.** We evaluate VideoGLaMM on three challenging tasks: grounded conversation generation (GCG), visual grounding, and referring video segmentation. For grounded conversation generation, we curate a separate dataset of 308 refined video-QA triplets containing 826 objects and 22,762 fine-grained masks, following our proposed annotation pipeline. For Visual Grounding, we evaluate our model on challenging VidSTG [53] dataset, considering only the interrogative sentences as done by [32]. In the case of motion-guided video object segmentation, we leverage the MeViS [12] validation dataset. All the results on MeViS dataset are obtained via official CodaLab evaluation suite. We also report referring video object segmentation results on additional Ref-DAVIS-17 [18] dataset.

**Evaluation metrics.** For GCG task, we use mean Intersection over Union (mIOU) and Recall to determine the correctness of generated masks, and METEOR, CIDEr and CLAIR score for determining the goodness of conversational output. In the case of visual grounding, we report mean Intersection over Union (mIOU) to quantify the performance. Finally, for referring video segmentation, We report Region Jaccard $\mathcal{J}$, Boundary F measure $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$.

**Baselines.** We compare our VideoGLaMM with two challenging baselines employing LLMs capable of visual grounding: PG-Video-LLaVA [32] and GLaMM [39]. Since GlaMM is designed for pixel grounding in images, we enable temporal properties in GLaMM by augmenting its architecture with SAM2. For referring segmentation, we also compare VideoGLaMM with the recently released Video-LISA [5].

**Training Recipe.** VideoGLaMM follows a gradual training schedule. We do not train VideoGLaMM on our GCG dataset directly from the start, rather we take a gradual approach. We first train the model on image and video segmentation datasets until epoch 20 and then introduce our GCG dataset and train the model until epoch 30. This training recipe ensures that model learns both the spatial and temporal cues effectively.

## 5.1. Grounded Conversation Generation

The Grounded Conversation Generation (GCG) task aims to provide video-level detailed captions with specific phrases directly tied to corresponding segmentation masks in the video frames. For example, "`<An adult> in white clothes holds a <cup> in the room`", as shown in the first row of Fig. 4, features how each bracketed phrase is anchored to a unique segmentation mask. This creates a densely annotated caption that aligns textual descriptions with visual regions in the frames, enriching the video's contextual interpretation. To obtain GCG output, we query the model with the following sample prompt: "`Provide a detailed description of the image. Respond with interleaved segmentation masks for the corresponding parts of the answer.`" The model generates a detailed caption along with interleaved segmentation masks, employing the format "`<p>An adult woman in brown</p><SEG> is talking to another <p>adult man wearing jacket</p><SEG>`" as shown in the third row of Fig. 4. We use special tokens, namely `<p>`, `</p>` and `<SEG>`, to delineate the start and end of each phrase and its corresponding region mask, respectively.

As shown in Table 1, our proposed Video-GLaMM performs better in generating detailed captions containing references to objects in the video frames, as is evident from high METEOR, CIDEr and CLAIR scores. Regarding the quality of masks, VideoGLaMM consistently outperforms baselines in terms of mIOU and Recall scores, signifying a higher overlap with ground-truth masks. Fig. 4 further shows the qualitative visualizations of VideoGLaMM on GCG samples.

| Model | mIoU | Recall | METEOR | CIDEr | CLAIR |
|---|---|---|---|---|---|
| PG-Video-LLaVA [32] | 24.03 | 0.093 | 0.10 | 0.01 | 15.0 |
| GLaMM [39] + SAM2 [41] | 28.60 | 0.117 | 0.097 | 0.15 | 22.9 |
| VideoGLaMM | **62.34** | **0.375** | **0.103** | **0.59** | **28.2** |

Table 1. **Evaluation on grounded conversation generation (GCG):** VideoGLaMM shows superior performance in generating accurate video-level captions which are tied to corresponding segmentation masks in the video frames.

## 5.2. Referring Video Segmentation

For referring video segmentation, the output should be grounded as per the given phrase, pointing towards specific instances in the video. Given a sentence or referring expression containing a specific object instance, the goal is to localize the object instances present across the video frames. This task operates in an open vocabulary setting, assessing the model's ability to localize objects both spatially and temporally. Given a referring phrase expression `Phrase`, we prompt the model using the following instruction prompt to obtain the instance masks: "`What is {Phrase} in this video? Respond with segmentation masks`". Table 2 shows results on challenging MeViS dataset for motion-guided referring video segmentation. Both the region Jaccard $\mathcal{J}$ and boundary F-measure $\mathcal{F}$ are high in the case of VideoGLaMM, significantly outperforming the baselines. Similarly, the mean $\mathcal{J}\&\mathcal{F}$ follows the same trend. Additionally, the scores corresponding to VideoLISA are reported with post-processing step. Notably, VideoLISA involves an additional post-processing step to boost performance. Therefore, we further fine-tune the VideoGLaMM on the task of referring segmentation post epoch 30 until epoch 40. Clearly, VideoGLaMM outperforms the VideoLISA (post-processed) on both $\mathcal{J}$ and $\mathcal{F}$, including the mean $\mathcal{J}\&\mathcal{F}$. Additionally, VideoGLaMM outperforms baselines on Ref-DAVIS-17 dataset. The improved performance of VideoGLaMM can be credited to its training pipeline, which seamlessly integrates spatio-temporal dynamics into the model.

| Model | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|
| LMPM (MeViS baseline) [12] | 37.2 | 34.2 | 40.2 |
| PG-Video-LLaVA [32] | 18.35 | 19.39 | 18.87 |
| GLaMM [39]+ SAM2 [41] | 35.80 | 41.50 | 38.66 |
| VideoLISA [5] | 41.30 | 47.60 | 44.40 |
| VideoGLaMM | **42.07** | **48.23** | **45.15** |

Table 2. **Performance comparison of VideoGLaMM on MeViS:** VideoGLaMM shows superior performance on motion grounding and segmenting referring objects in the videos.

| Model | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|---|
| LISA-7B [20] | 61.9 | 54.9 | 58.4 |
| LISA-13B [20] | 64.6 | 56.8 | 60.7 |
| TrackGPT-7B [58] | 67.0 | 59.4 | 63.2 |
| TrackGPT-13B [58] | 70.4 | 62.7 | 66.5 |
| VideoLISA [5] | 72.7 | 64.9 | 68.8 |
| VideoGLaMM | **73.3** | **65.6** | **69.5** |

Table 3. **Performance comparison of VideoGLaMM on Ref-DAVIS-17:** VideoGLaMM shows superior performance on segmenting referring objects in the videos.

## 5.3. Visual Grounding

| Model | VidSTG (interrogative mIoU) |
|---|---|
| PG-Video-LLaVA-7B [32] | 34.20 |
| PG-Video-LLaVA-13B [32] | 35.10 |
| GLaMM [39] + SAM2 [41] | 38.63 |
| VideoGLaMM | **39.66** |

Table 4. **Performance comparison of VideoGLaMM with other models on spatial grounding:** Results on VidSTG (interrogative) benchmark highlights VideoGLaMM's superior ability in correlating textual instructions with the visual frames.

To quantitatively assess VideoGLaMM's visual grounding capability, we conduct quantitative evaluations on the benchmark test set of VidSTG dataset. The visual grounding task measures the adeptness of the model at correlating textual descriptions with visual elements in the video, a critical aspect of contextual comprehension. This ability is crucial for applications that integrate continuous visual data with language. The output of this task is refined masks that correlate with the given caption `{caption}`. To obtain the visual grounding output, we query the model with interrogative captions. For these captions, the prompt format follows "`{caption} Please respond with a segmentation masks.`". Table 4 shows VideoGLaMM's improved visual grounding precision as it outperforms the baselines, demonstrating its fine-grained understanding.

In addition to the above downstream tasks, in Sec. B of supplementary, we also integrate VideoGLaMM into a conditional video generation model [45]. VideoGLaMM provides temporally coherent masks that guides generative model in editing videos effectively. Please refer to Sections A and B of supplementary for additional quantitative and qualitative results respectively.

## 5.4. Ablation studies

**Effect of Spatio-Temporal Dual Encoder.** We employ separate image and video encoders to process spatial and temporal information. While spatial processing induces local information, temporal processing helps learn global features. Both are necessary from the perspective of grounding. To verify the effectiveness of dual spatio-temporal encoder, we conduct an ablation study to measure the effectiveness of each encoder for grounded conversation generation (GCG) task (see Table 5). We notice that using only an image encoder gives suboptimal results, as we notice a drop in both the localization and captioning metrics. Using only video branch leads to the highest mIOU; however, relatively lower METEOR, CIDEr, and CLAIR scores. To obtain an optimal mIOU and good conversational abilities, VideoGLaMM uses both image and video encoders.

| Encoder Configuration | mIoU | Recall | METEOR | CIDEr | CLAIR |
|---|---|---|---|---|---|
| Image encoder | 60.06 | 0.395 | 0.081 | 0.371 | 18.9 |
| Video encoder | 64.62 | 0.375 | 0.097 | 0.568 | 26.5 |
| Dual encoder | 62.34 | 0.375 | 0.103 | 0.590 | 28.2 |

Table 5. **Effect of Spatio-Temporal Dual Encoder:** We obtain low performance using only spatial (image) encoder. Using only a video encoder gives the highest mIOU but lower scores on CLAIR, METEOR and CIDEr. For a better trade-off, we employ dual (image and video) encoders to have accurate, grounded conversations.

**Spatial vs Spatio-temporal Pixel decoder.** Pixel decoder in VideoGLaMM can operate in two configurations. The first configuration processes video frames individually, ignoring temporal consistency. The second configuration employs both spatial and temporal branches for spatio-temporal context. Table 6 demonstrates the impact of spatiotemporal decoder on the GCG task. Results indicate that using only the spatial configuration reduces performance, with a nearly 3% drop in mIOU scores compared to the spatio-temporal configuration. Similarly, metrics like METEOR, CIDEr, and CLAIR also show a decline, underscoring the importance of using spatio-temporal configuration for pixel decoder.

**Effect of number of frames for Pixel Decoder.** The pixel decoder receives the raw input frames encoded via frame encoder as input for predicting fine-grained grounded masks. During training, the pixel decoder also receives ground-truth masks which act as supervision signals. To provide more temporal supervision, we feed the pixel decoder with multiple input frames to enhance its temporal understanding. This allows it to learn semantic information that generalizes across frames. Table 7 shows the performance when 4 and 8 frames are input to the decoder. We observe that while the mIOU with 8 frames is slightly lower compared to 4 frames, the conversational quality measured by METEOR and CLAIR is higher. Hence, to achieve a decent mIOU with higher conversational output, we stick to 8 frames in the paper.

| Decoder Configuration | mIoU | Recall | METEOR | CIDEr | CLAIR |
|---|---|---|---|---|---|
| Spatial decoder | 59.68 | 0.369 | 0.097 | 0.553 | 26.7 |
| Spatio-temporal decoder | 62.34 | 0.375 | 0.103 | 0.59 | 28.2 |

Table 6. **Spatial vs Spatio-temporal Pixel decoder:** We observe that using Pixel decoder without the temporal branch gives limited performance as the model faces difficulties in temporal grounding. When using temporal branch, the performance on both the temporal grounding and grounded LLM response improves indicating the importance of temporal processing in VideoGLaMM.

| Decoder Input frames | mIoU | Recall | METEOR | CIDEr | CLAIR |
|---|---|---|---|---|---|
| 4 frames | 63.82 | 0.37 | 0.094 | 0.659 | 27.2 |
| 8 frames | 62.34 | 0.37 | 0.103 | 0.590 | 28.2 |

Table 7. **Effect of number of frames for Pixel Decoder:** We observe that using 4 supervision frames for pixel decoder gives better mIOU but relatively modest conversation quality measured by METEOR and CLAIR. With 8 supervision frames, mIOU slightly decreases while the conversational quality increases.

**Limitations and Future Work:** Our GCG dataset plays a key role in enhancing the model's grounding capabilities. While we validated annotations manually, some noise may still be present. Also, each scene contains several objects and the video descriptions do not exhaustively cover all objects in the scenes. A higher-quality densely annotated set could further boost model performance but would require substantial annotation resources. Additionally, VideoGLaMM struggles with objects of varying granularities, likely due to limited representation in the training data. Another improvement is to extend VideoGLaMM for longer videos, as the current GCG dataset mainly focuses on short-medium duration clips.

## 6. Conclusion

We introduce VideoGLaMM, a LMM specifically designed to address the challenge of fine-grained pixel-level grounding in videos. By integrating a dual vision encoder with a spatio-temporal decoder and employing tunable Vision-Language adapters, our model achieves precise alignment between video content and textual instructions. To facilitate this alignment, we introduce a refined instruction-tuning dataset curated via a semi-automatic annotation pipeline. Our experimental evaluations across Grounded Conversation Generation, Visual Grounding, and Referring Video Segmentation tasks demonstrate that VideoGLaMM consistently outperforms existing models.

## References

[1] Marah 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. *arXiv preprint arXiv:2404.14219*, 2024. 6

[2] 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

[3] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023. 5

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2

[5] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *arXiv preprint arXiv:2409.19603*, 2024. 3, 6, 7, 1

[6] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 2

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proc. of NeurIPS*, 2020. 1

[8] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6

[9] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proc. of NeurIPS*, 2023. 1

[12] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 5, 6, 7

[13] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: general language model pretraining with autoregressive blank infilling. In *Proc. of ACL*, pages 320–335, 2022. 1

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. 6

[16] Lei Ke, Henghui Ding, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Video mask transfiner for high-quality video instance segmentation. In *European Conference on Computer Vision (ECCV)*, 2022. 5

[17] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. *arxiv: 1803.08006*, 2018. 6

[18] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 6

[19] 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. 5

[20] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1, 2, 3, 6, 7

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[22] 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:2305.06355*, 2023. 2

[23] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2

[24] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 6

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[27] 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. 5

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[29] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 2

[30] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2023. 2, 6

[31] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 2, 4, 5, 6

[32] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 2, 3, 6, 7

[33] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2

[34] OpenAI. Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023. 5

[35] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proc. of NeurIPS*, 2022. 1

[36] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[37] 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, 5

[38] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 6

[39] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 2, 3, 6, 7

[40] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, pages 3505–3506, 2020. 6

[41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6, 7

[42] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 5

[43] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 2

[44] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 5

[45] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7393–7402, 2024. 8, 1

[46] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 5

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, 2023. 1

[48] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[49] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 2

[50] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 6

[51] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2

[52] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023. 2, 5

[53] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020. 5, 6

[54] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 2, 3

[55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[56] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019. 5

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2

[58] Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with human-intent reasoning. *arXiv preprint arXiv:2312.17448*, 2023. 7, 1