

VideoChat : Chat-Centric Video Understanding

KunChang Li^{*1,4}, Yanan He^{*1}, Yi Wang^{*†1}, Yizhuo Li^{1,3}, Wenhai Wang¹

Ping Luo³, Yali Wang^{4,1}, Limin Wang^{2,1}, Yu Qiao¹

¹OpenGVLab, Shanghai AI Laboratory ²Nanjing University ³The University of Hong Kong

⁴Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

<https://github.com/OpenGVLab/Ask-Anything>

Abstract

In this study, we initiate an exploration into video understanding by introducing VideoChat, an end-to-end chat-centric video understanding system. It integrates video foundation models and large language models via a learnable neural interface, excelling in spatiotemporal reasoning, event localization, and causal relationship inference. To instructively tune this system, we propose a video-centric instruction dataset, composed of thousands of videos matched with detailed descriptions and conversations. This dataset emphasizes spatiotemporal reasoning and causal relationships, providing a valuable asset for training chat-centric video understanding systems. Preliminary qualitative experiments reveal our system’s potential across a broad spectrum of video applications and set the standard for future research. Access our code and data at <https://github.com/OpenGVLab/Ask-Anything>.

1 Introduction

Videos offer a remarkably close representation of how humans consistently perceive the visual world. Intelligent video understanding is crucial for various real-world applications, such as human-robot interaction, autonomous driving, and intelligent surveillance, to name a few. However, current paradigms in video understanding are limited by task-specific tuning of pre-trained video foundation models, restricting a general spatiotemporal comprehension for client-level needs.

Vision-centric multimodal dialogue systems have recently emerged as an essential research area [14]. By utilizing a pre-trained large language model (LLM), an image encoder, and additional learnable modules, these systems can deeply understand images (e.g., recognizing memes or jokes) and perform image-related tasks through multi-round dialogues with user queries [25, 58, 52]. This revolutionizes numerous applications, but existing systems have yet to formally address video-centric tasks from a data-centric perspective using learning machines.

Existing video-centric multimodal dialogue systems ¹ [16] transform video understanding into a natural language processing (NLP) question-answering formulation by textualizing video content using open-source vision classification/detection/caption models. Despite demonstrating decent performance in short-term scenarios with clear objects and actions, converting videos into textual descriptions inevitably results in visual information loss and over-simplification of spatiotemporal complexities. Additionally, almost all utilized vision models struggle with spatiotemporal reasoning, event localization, and causal relationship inference within videos.

To tackle these challenges, we introduce a groundbreaking chat-centric video understanding system that leverages state-of-the-art techniques from both video and language domains. Our approach

* Equal contribution. † Project lead

¹<https://github.com/OpenGVLab/Ask-Anything>

creates a full loop, integrating video and language foundation models in a learnable manner from a model perspective, and provides all techniques required to learn the system from a data perspective.

We begin by presenting our novel video-centric multimodal dialogue system. We propose an innovative system architecture that combines video foundation models and large language models (LLMs) through a learnable neural interface. By a two-stage lightweight training (with only spatiotemporal and video-language alignment modules) on large-scale video-text datasets and self-built video instruction ones, our method excels in spatiotemporal perception & reasoning, and causal inference, marking the first attempt to create a fully learnable and efficient video understanding system that facilitates effective communication.

We introduce a novel video-centric multimodal instruction fine-tuning dataset. We create a unique dataset comprising thousands of videos paired with detailed textual descriptions and conversations generated using dense captions fed to ChatGPT in temporal order. This dataset emphasizes spatiotemporal objects, actions, events, and causal relationships, offering a valuable resource for training video-centric multimodal dialogue systems.

Through these contributions, our work pioneers new frontiers in video and natural language processing integration. By developing a new and effective chat-centric video understanding dialogue system, we pave the way for a wide range of applications across various domains while setting a standard for future research in this field. Our research not only pushes the boundaries of video understanding and reasoning but also offers protocols for both academic and industrial communities.

2 Related Work

Video Foundation Models Large-scale video-text pretraining coupled with downstream task fine-tuning has emerged as the standard paradigm in the video-language domain [27, 23, 55, 49, 21, 20, 55, 49, 13, 11, 36, 51, 38, 60, 46, 5]. Early methods [38, 60] employed pretrained visual and language encoders to derive offline video and text features; however, more recent approaches [23, 27, 13, 11, 41, 44] have proven the viability of end-to-end training. Additionally, prevalent techniques often encompass two or three pretraining tasks, such as masked language modeling [22], video-text matching [43], video-text contrastive learning [49, 46], masked video modeling [41, 44, 46] and video-text masked modeling [12]. Within the realm of video multimodal tasks, VIOLET [12] integrates masked language and masked video modeling, while All-in-one [43] suggests a unified video-language pretraining methodology using a shared backbone, and LAVENDER [22] consolidates the tasks through masked language modeling. Although these approaches yield impressive results in multimodal benchmarks, their training relies on limited video-text data, which leads to difficulties in video-only tasks such as action recognition. On the other hand, MERLOT Reserve [54] compiles 20 million video-text-audio pairs for training joint video representations via contrastive span matching, thereby establishing state-of-the-art outcomes in video recognition and visual commonsense reasoning.

Large Language Models Recent advances in large language models (LLMs) [30, 29, 3, 42, 7, 10, 40] have showcased remarkable capabilities such as language generation, in-context learning, etc. These abilities enable LLMs to tackle complex tasks with user prompts in a zero-shot fashion. GPT-3 [3] shows notable zero-shot performance across numerous benchmarks. InstructGPT models [32] are finetuned using datasets containing prompts with corresponding human-annotated desired behavior. This results in better alignment with users, improved output quality compared to GPT-3, increased truthfulness, and reduced risks. Instruction-tuned models also present remarkable generalization capacity for zero-shot tasks. Therefore, instruction-tuning [28, 8] is crucial in leveraging LLMs’ potential. Besides of GPT family [30, 29, 3], there are multiple LLMs, including OPT [57], LLaMA [42], MOSS [9], and GLM [56], providing high-performance, open-source resources that can be finetuned for various purposes. For instance, Alpaca [40] proposes a self-instruct framework to instruction-tune LLaMA models without heavily relying on human-authored instruction data.

LLMs for Multimodal Understanding The accomplishments of LLMs have accelerated the creation of AI systems that merge vision models with LLMs to enable multimodal reasoning and action [25, 58, 52, 1, 19, 26, 47, 50, 37, 24, 26]. Flamingo [1] pioneered this approach by capitalizing on both vision and language models using web-scale image-text interwoven data, unveiling exceptional zero-shot image-text abilities in a conversational format for the first time. The study in

[14] demonstrates that Kosmos-1 models are naturally equipped to tackle a broad array of perception-intensive tasks, including visual dialogue, visual explanation, visual question answering, image captioning, basic math equations, OCR, and zero-shot image classification using descriptions. Visual instruction tuning introduces an innovative technique for refining large language models on visual instruction tasks, enabling pretrained BLIP and Vicuna to nearly match GPT-4 level conversation performance for image-based tasks [25]. MiniGPT-4 is a multimodal large language model, fine-tuned on multimodal tasks, and exhibits respectable zero-shot image comprehension in dialogues [58].

3 VideoChat

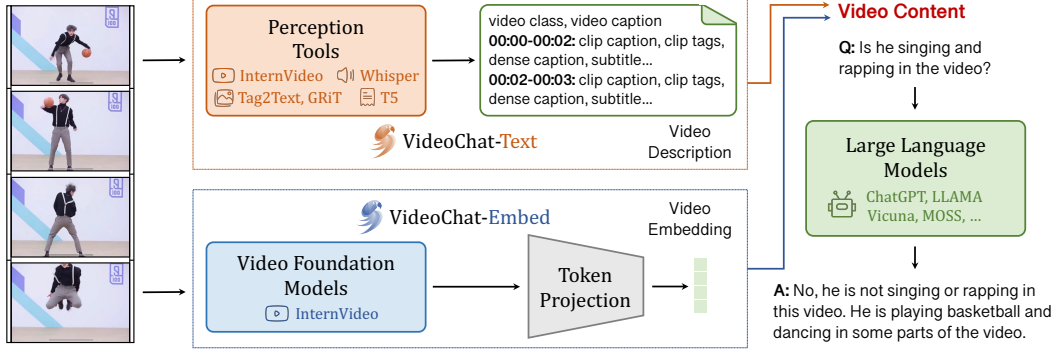


Figure 1: **The framework of VideoChat.** **VideoChat-Text** textualizes videos in stream. **VideoChat-Embed** encodes videos as embeddings. Both video content will be input in LLMs for multimodal understanding.

VideoChat unifies video-related tasks into the formulation of multiple-round video question answering, in which tasks are defined by words in a live inference and no or a few instances are given for learning. In this formulation, we treat an LLM as a universal video task decoder, turning video-related descriptions or embeddings into human-understandable text. This procedure is user-friendly in employing foundation models to address various applications.

Formally, we extract concepts from videos using vision models as:

$$[\mathbf{E}]_i^j = f_{\text{img}}^j(\mathbf{I}_i) \quad \text{or} \quad \mathbf{E}^j = f_{\text{vid}}^j(\mathbf{V}) \quad \text{w.r.t.} \quad \mathbf{V} = [\mathbf{I}_i]_{i=1,2,\dots,T}, \quad (1)$$

where \mathbf{E} denotes a text description or embedding according to context, f_{img}^j denotes the j th image model to predict human-readable annotations or visual feature, while \mathbf{I} and \mathbf{V} denote an image and video, respectively. Then we decode the task prediction from a LLM based on user’s question as:

$$\mathbf{W}_t^a = f_{\text{llm}}(\mathbf{E} | \mathbf{W}_{\leq t}^q, \mathbf{W}_{< t}^a), \quad (2)$$

where \mathbf{W}_t^a and $\mathbf{W}_{\leq t}^q$ stand for the answers from the LLM at the round t and all questions given by users before round t , respectively. f_{llm} denotes an LLM model.

In technical terms, an ideal end-to-end chat-centric video understanding system should utilize a video/vision base model (an encoder) to convert visual sequences into latent features for LLM, guaranteeing the system’s overall differentiability. Prior to this, we verify the efficacy of LLM as a universal video task interpreter through our proposed VideoChat-Text (Section 3.1). This method transforms videos into textual streams for subsequent discrimination/reasoning tasks using LLMs by incorporating various open-source vision models. While VideoChat-Text can tackle typical spatiotemporal tasks such as spatial and temporal perception, it falls short in comprehending intricate temporal reasoning and causal inference. Therefore, we introduce VideoChat-Embed (Section 3.2), a multimodal system that combines both video and language foundation models. Finetuned with video instruction data, it significantly enhances performance in higher-order temporal assignments. We will describe these two approaches in the following sections.

3.1 VideoChat-Text 🦜: VideoChat by Textualizing Videos in Stream

We employ several vision models to convert visual data from videos into textual format. Subsequently, we create purpose-built prompts to temporally structure the predicted text. Ultimately, we rely on a pretrained LLM to address user-specified tasks by responding to questions based on video text descriptions.

In particular, for a given video, we use `ffmpeg` to extract key frames from the video at a low FPS, resulting in T video frames and associated audio. By feeding the extracted frames and audio into various models, we acquire action labels, frame summaries, video tags, comprehensive descriptions, object positional coordinates, video narratives, timestamps, and other segment-related details. We then consolidate related content in the captions considering the timing and generate a timestamped video text description. We will first outline the vision models and prompt schematics employed, and then conclude with an analysis of VideoChat-Text.

3.1.1 Perception Models

Utilizing a combination of video and image models [46, 34, 20, 48, 15, 33], we analyze videos from various aspects such as actions [46, 20, 21], objects [45, 15], object annotations with positions [48], and more. While the majority of these models’ outputs are comparatively independent, we utilize the pretrained T5 language model [34] to refine their descriptions for improved clarity. Moreover, we integrate the Whisper [33] speech recognition model into VideoChat-Text to capitalize on audio data within videos, further enhancing the richness of video descriptions.

3.1.2 Prompt System

We process the video into different visual models to obtain different textualizing videos and then organize them together in a template (shown in Table 1) as inputs to an LLM. Then, we present the LLM with the context that we instruct it to pretend to watch the given video through the formatted texts we feed and then chat with us with the structured video knowledge. This context is realized by the prompt in Table 2.

Analysis Lite perception models enable VideoChat-Text to convert videos into time-stamped text at 1 FPS, processing a 10-second video clip in about 2 seconds using an NVIDIA-A10 GPU. It communicates with users through an LLM. However, using text as the communication medium restricts the representation capabilities of the perception models, as it limits their decoders. To provide richer visual information from videos to the LLM, we must employ more advanced and more perception models, which may conflict with VideoChat-Text’s efficiency. Additionally, VideoChat-Text has limited potential to benefit from popular visual instruction tuning [25].

3.2 VideoChat-Embed 🦜: VideoChat by Encoding Videos as Embeddings

VideoChat-Embed is an end-to-end model designed to handle video-based dialogue. It employs an architecture (Figure 2a) that combines both video and language foundation models with an addition learnable Video-Language Token Interface (VLTF). To achieve better cross-modality optimization, the model incorporates language-friendly video foundation models, inspired by [53, 49, 46, 21]. Considering the video redundancy [41], we introduce the VLTF, using cross-attention to compress the video tokens. It is tuned with video-text data for video-to-language representation alignment. Finally, the video tokens, user queries, and dialogue context are input into the LLM for communication.

3.2.1 Architecture

In this paper, we instantiate the VideoChat-Embed based on BLIP-2 [18] and StableVicuna [10]. Concretely, we incorporate the pretrained ViT-G [39] with Global Multi-Head Relation Aggregator (GMHRA), a temporal modeling module used in InternVideo [46] and UniFormerV2 [20]. For the token interface, we employ the pretrained QFormer with extra linear projection, supplemented by additional query tokens to account for video context modeling. This allows us to obtain compact LLM-compatible video embeddings for future dialogues.

When training, we freeze most of the parameters except the newly incorporated GMHRA, queries and linear projection. Inspired by [21], we introduce image data for joint training. In Stage1, we align

Video Class, Video Caption

00:00-00:02 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...

00:02-00:03 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...

00:03-00:06 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...

...



answering questions, a man and a woman sitting on a couch in a living room with a table in front of them.

00:00-00:11 a man and a girl sitting on a couch in a living room.

a lamp with a white shade a woman sitting at a table: [446, 155, 710, 476]; man wearing a plaid shirt: [361, 44, 581, 337]; man sitting on couch: [10, 63, 324, 350]; the tie is grey: [441, 150, 486, 280]; a glass of beer: [38, 305, 77, 367]; a stack of magazines: [28, 350, 180, 394]; a white tablecloth: [0, 334, 626, 476]; stainless steel oven: [1, 55, 150, 142]; a brown tie on a man: [144, 168, 191, 270]; the couch is white: [0, 119, 730, 472]; a gray binder: [0, 377, 157, 411]; a white couch: [768, 350, 848, 477]; a lamp with a white shade: [582, 26, 713, 195];

00:00-00:02: Hey, Pheebs, you gonna have the rest of that Pop-Tart?

00:02-00:03: Pheebs?

00:03-00:09: Does anyone want the rest of this Pop-Tart?

00:09-00:11: Hey, I might.

Table 1: **Video description with perception models.** We use perception models to obtain the video clip action class, clip/video caption, and dense caption as well as a subtitle. Then we organize them in a template to generate textualizing videos.

You are a chatbot that conducts conversations based on video contexts. You mainly answer based on the given contexts, and you can also modify the content according to the tag information, and you can also answer the relevant knowledge of the person or object contained in the video. **The timing description is a description every 1/FPS second, so that you can convert it into time. When describing, please mainly refer to the timing description. Dense caption is to give content every five seconds, you can disambiguate them in timing.** But you don't create a video plot out of nothing.

Begin!

Video contexts in temporal order: `textualizing_videos`

Question: `question`

Table 2: **System prompt.** This description allows the LLM to understand textualizing videos and respond according to relevant content in the document, avoiding answering questions that do not relate to the video.

the video encoder with LLM via large-scale video-text fine-tuning. In Stage2, we tune the system with two types of video instruction data: in-depth video descriptions and video question-answer pairs. The following section will describe the process of generating instruction data and present the details of two-stage training paradigm.

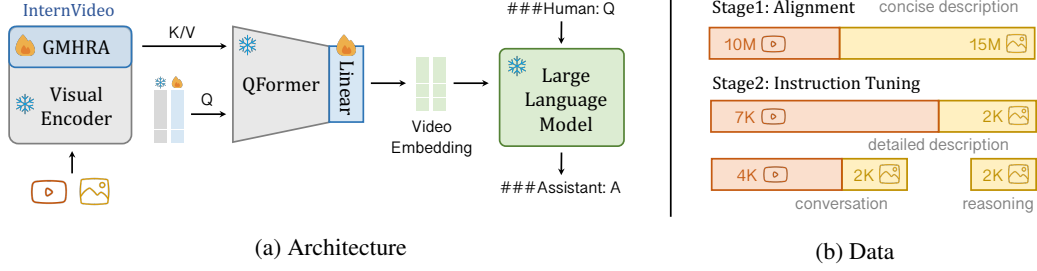


Figure 2: **Architecture and training paradigm of VideoChat-Embed.** The architecture is built on BLIP-2 [18] and StableVicuna [10]. The training contains two-stage alignment and instruction tuning.

3.2.2 Instruction Data

We build a video-centric multimodal instruction data based on WebVid-10M [2]. The corresponding detailed descriptions and question-answer generations are produced by ChatGPT based on video text (aided by VideoChat-Text) with several prompts concerning spatiotemporal features. Compared with detailed video descriptions, video conversations are introduced to further improve the diversity, temporal and casual features in the video instruction data.

Detailed Video Descriptions We condense the provided video description into a video narrative employing GPT-4, as shown in Table 5. This highlights the temporal aspects of the video by illustrating its progression over time. The associated prompts can be found in Table 3 and 4. The first converts the various predicted textual labels into a cohesive, evolving story, while the second one refines the narrative to improve clarity and coherence, as well as minimize hallucination. We generated a total of 7K descriptions from randomly chosen videos.

Give you a video of [origin_caption](#). The content of the video in temporal order is: [textualizing_videos](#). Please use the sequence adverbs "first", "next", "then" and "finally" to describe this video in detail, but don't mention the specific time. Give as many details as possible. Say everything you see. The description should be more than 150 words and less than 200 words.

Table 3: **Prompts for detailed video descriptions.** The [origin_caption](#) is generated from VideoChat-Text.

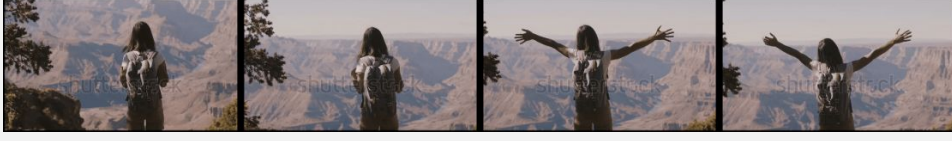
Fix the error in the given paragraph. Remove any repeating sentences, meaningless characters, not English sentences, and so on. Remove unnecessary repetition. Rewrite any incomplete sentences. Return directly the results without explanation. Return directly the input paragraph if it is already correct without explanation.

Table 4: **Prompt for post-processing.** Copied from miniGPT-4 [58].

Video Conversations With the video description, we generate multi-round dialogues with three types of prompts concerning descriptive, temporal, and causal content for videos with ChatGPT. The descriptive part mostly inherits key points from LLaVA [25]. For the temporal and causal parts, we propose prompts (Table 6) focus on temporal perception/reasoning and explanation/uncovering intentions/causes, respectively. We produced multi-round dialogues from 4K randomly chosen videos. One example of the video conversation can be found in Table 7

3.2.3 Two-stage Training

Motivated by MiniGPT-4 [58] and LLaVA [25], we have designed a two-stage joint training paradigm. This approach allows us to benefit from readily-available image instruction data, creating a system capable of handling both images and videos with shared spatial perception and reasoning capacity.



###Human: <Video>video_embed</Video> The video contains T frames sampled at t_0, t_1, \dots, t_T seconds.
 ###Human: Portray the video with a rich, descriptive narrative, capturing its atmosphere and events.
 ###Assistant: **First**, we see a young woman with long brown hair wearing a backpack standing on the edge of a cliff, looking at the beautiful mountain scenery. She is in awe and stands there for a few seconds. **Next**, we see her turning around slowly, with her arms open wide. **Then**, we see her taking a few steps back, still facing the canyon, with her arms still open wide. She seems to be soaking in the grandeur of the view. **Finally**, she stands confidently at the edge of the cliff, her backpack still on her back, as if she has conquered the mountain. Throughout the video, we see the vast expanse of the canyon with the mountains in the background. The sky is overcast in some parts and clear blue in others, creating a beautiful contrast. **Overall**, the young woman seems to be thrilled with the view and enjoys every bit of it with open arms.

Table 5: Example of using the detailed video description.

Stage1: Alignment. To strike a balance between training convergency and efficiency we introduce 25M vision-text pairs for one epoch of fine-tuning. The data consists 10M video-text pairs from WebVid-10M, and 15M image-text pairs from COCO Caption [6], Visual Genome [17], SBU Captions [31], CC3M [35] and CC12M [4]. The input prompts for LLMs are as followed:

- “###Human: <Video>video_embed</Video> video_instruction ###Assistant:”
- “###Human: <Image>image_embed</Image> image_instruction ###Assistant:”

The `video_embed` and `image_embed` are the output from the token interface. Meanwhile the `video_instruction` and `image_instruction` provide concise video and image descriptions randomly sampled from predefined instructions in Table 8. Language models receive corresponding visual descriptions as answers.

Stage2: Instruction tuning. As discussed in Section 3.2.2, our self-built video instruction data consists of 7K detailed video descriptions and 4K video conversations. To improve spatial perception and reasoning capabilities, we also gather 3K detailed image descriptions from MiniGPT-4 [58], 2K image conversations, and 2K image reasoning tasks from LLaVA [25]. With this 18K data collection, we tune the system for 3 epochs. Note we include temporal reasoning sampling information for video data: “The video contains T frames sampled at t_0, t_1, \dots, t_T seconds.”

4 Experiments

4.1 Qualitative Analysis

We give some case studies at this stage. Besides of our VideoChat-Text 🐦 and VideoChat-Embed 🐦, we make qualitative comparisons with LLaVa 🐦 [25], miniGPT-4 🐦 [58], and mPLUG-owl 🐦 [52].

Spatial Perception and Analysis In Figure 7, our approach (VideoChat-Embed) accurately deduces the corresponding music by recognizing Japanese-style clothing and determining the number of individuals present. This confirms the system’s ability to identify objects along with their properties, while also providing pertinent recommendations based on visual elements. Also, we give some image-centric dialogue examples in Figure 10.

Temporal Perception and Reasoning Figure 5, 7, and 9 demonstrate that VideoChat-Embed is capable of performing accurate temporal perception and reasoning. In Figure 5, our system identifies actions over time in a zero-shot fashion, recognizing that the subject played basketball and engaged in

As an AI visual assistant, you are observing a single video. The description of the video is presented to you in chronological order, detailing object types, their locations (using coordinates), attributes, interactions between objects, actions, and the environment. Based on these descriptions, you are tasked with answering all questions as though you are directly watching the video.

Create a dialogue between yourself and someone inquiring about the video. Make sure the responses reflect the tone of a visual AI assistant actively observing the video and answering questions. Include diverse queries and corresponding answers.

Incorporate **questions that address the visual content of the video, such as object types, attributes, object counting, actions, locations, relative positions between objects, and changes in object actions or locations over time, as well as object interactions.** Only include questions with definitive answers:

- Questions whose contents can be confidently observed and answered based on the video.
- Questions whose absence from the video can be confidently determined.

Next, encompass **questions related to temporal perception and reasoning, such as inquiring about what a person did before or after an event, or asking for specific timestamps of certain events or actions**

Also include **complex questions relevant to the video’s content, like those asking about the background knowledge of objects or actions in the video, discussing events occurring in the video, delving into counterfactual topics (e.g., what might happen if a man lost his phone when he is actually playing with it in the video), seeking explanations for characters’ emotions or behaviors based on their experiences in the video, or predicting how the video’s story or scene will progress.**

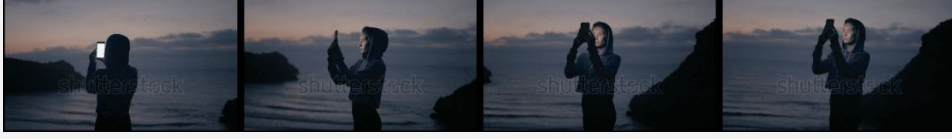
Since you receive video descriptions while viewing the video, **prioritize asking more questions about visual changes over time and the reasons or causes behind these changes rather than questions that can be inferred from a single frame.**

Remember not to inquire about uncertain details. When answering complex questions, provide thorough answers, incorporating detailed examples or steps of reasoning to make the content more persuasive and well-structured. Use multiple paragraphs if necessary. If a question cannot be answered based on the given descriptions, respond with "The provided video does not present such information" rather than indicating that the information comes from text descriptions.

Table 6: **Prompt for video conversations.** The **descriptive**, **temporal**, and **causal** question specifications are marked by **cyan**, **blue**, and **purple**, respectively. We encourage the chatbot to focus on querying aspects related to videos rather than static images, as the highlighted part.

dance movements within a specific timeframe. Additionally, it captures camera motion, showcasing its understanding of filming perspectives. In Figure 9, VideoChat-Embed accurately identifies yoga in the video and provides rational explanations for this activity (practice and enjoyment). Intriguingly, when questioned about the likelihood of the yoga practitioner falling, it asserts that proper safety precautions were taken by the individual.

Casual Inference It is evident that VideoChat-Embed can infer causal relationships using spatiotemporal clues, as demonstrated in Figure 5, 6, 7, and 4. In Figure 5, the model provides an impartial description of the video, primarily highlighting objects, actions, and emotions without commenting on the boy’s dance style. To explain why the video is amusing, VideoChat-Embed cites the erratic and spontaneous nature of the boy’s movements while also conveying an emotional assessment that his dancing appears foolish, accounting for the unexpected humor within the clip. Empirically, we confirm that these visually associated abstract concepts are derived from the video foundation model instead of being hallucinations produced by the utilized LLM. In Figure 6, VideoChat-Embed accurately assesses that a car accident occurred due to the collision of vehicles and the damage sustained to the front car’s license plate visible to the camera. In Figure 7, VideoChat suggests pairing the video with light, enjoyable music, as it can sense the girls’ dancing rhythm and progression over time. This demonstrates that our system can effectively capture and summarize abstract concepts related to movement patterns.



###Human: <Video>video_embed</Video> The video contains T frames sampled at t_0, t_1, \dots, t_T seconds.

###Human: What is happening in the video?

###Assistant: In the video, we see a silhouette of a young woman holding a smartphone and taking pictures of the sunset over the sea on the beach at night. Then, there are several other silhouettes of people taking pictures and looking at their phones.

###Human: Can you describe the **interactions** between the objects in the video?

###Assistant: The silhouettes of people in the video are holding smartphones, and she is taking pictures of the sunset over the sea on the beach at night.

###Human: What is the **location** of the scene?

###Assistant: The scene is located on a beach at night near the sea.

###Human: Can you tell me about any changes that happen in the video **over time**?

###Assistant: The color of the sky changes from orange to grey as the video progresses. Additionally, the actions of the silhouettes of people change as they take pictures and look at their phones on the beach.

Table 7: Example of using the video conversation.

4.2 Comparisons

As depicted in Figure 3, we present a comparison of our approach to recent image-based multimodal dialogue systems in image-related tasks, using a query example from the TVQA dataset. We assess this case through the online demos provided by each respective system. It is evident that our VideoChat-Embed correctly identifies the scene, while other systems inaccurately perceive the conversation setting as indoors. This result highlights the superior spatial perception abilities of VideoChat-Embed in relation to its counterparts. Furthermore, this proficiency remains consistent when dealing with a video from the same dataset, as demonstrated in the right-hand example.

5 Conclusion

We have embarked on a pioneering investigation into general video understanding by developing VideoChat, a multimodal dialogue system specifically designed for videos. Two versions of VideoChat are implemented: a text-based version demonstrates the effectiveness of employing large language models as universal decoders for video tasks, while an end-to-end version makes a preliminary attempt to tackle video understanding using instructed video-to-text formulation. Our end-to-end solution effectively merges video foundation models with large language models through a trainable neural interface. To enhance the system’s performance, we introduced a video-centric instructional dataset, which highlights spatiotemporal reasoning and causality by offering a learning resource for video-based multimodal dialogue systems. Initial qualitative evaluations showcase our system’s promising capabilities across various video applications and drive its ongoing advancement.

Limitations 1) Both VideoChat-Text and VideoChat-Embed struggle with managing long-term videos (≥ 1 min). On one side, effectively and efficiently modeling the context of long videos remains a complex and persistent research issue. Conversely, balancing response time, GPU memory usage, and user expectations for system performance becomes challenging when striving to provide user-friendly interactions while processing longer videos. 2) Our system’s capacities for temporal and causal reasoning remain rudimentary. These limitations stem from the current scale of our instruction data and its construction approaches, as well as the overall system scale and the models employed. 3) Addressing performance disparities in time-sensitive and performance-critical applications, such as egocentric task instruction prediction and intelligent monitoring, is an ongoing challenge.

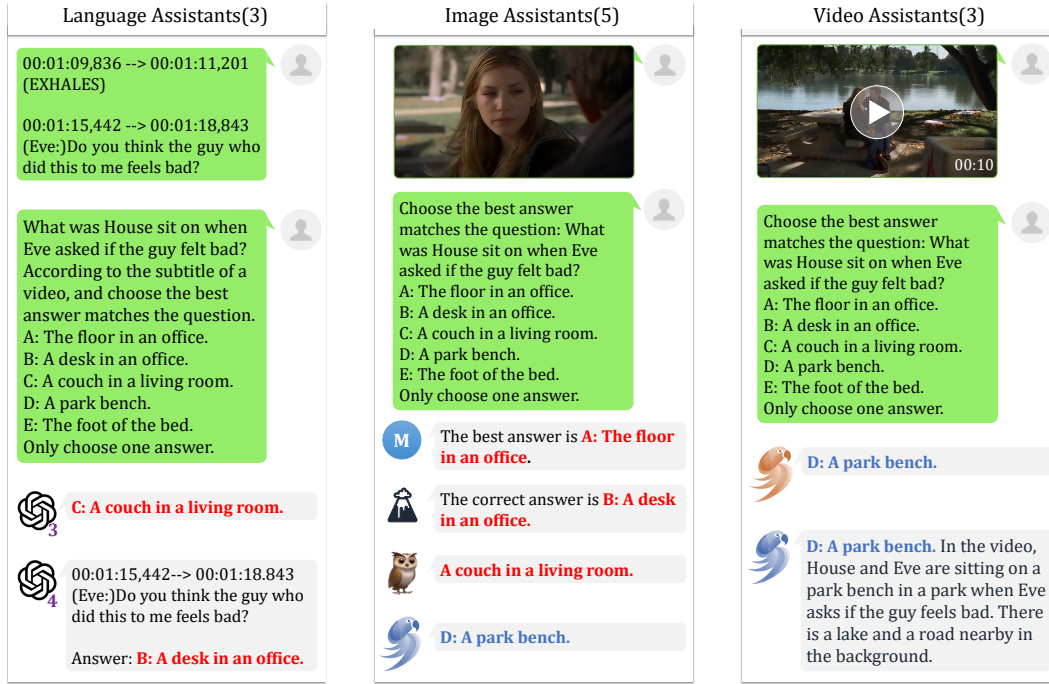


Figure 3: In the TVQA dataset sample, subtitles are fed to language assistants [30, 29], images to the image assistants [59, 25, 52], and videos to the VideoChat. While responding to user questions, only VideoChat utilizes information from videos.

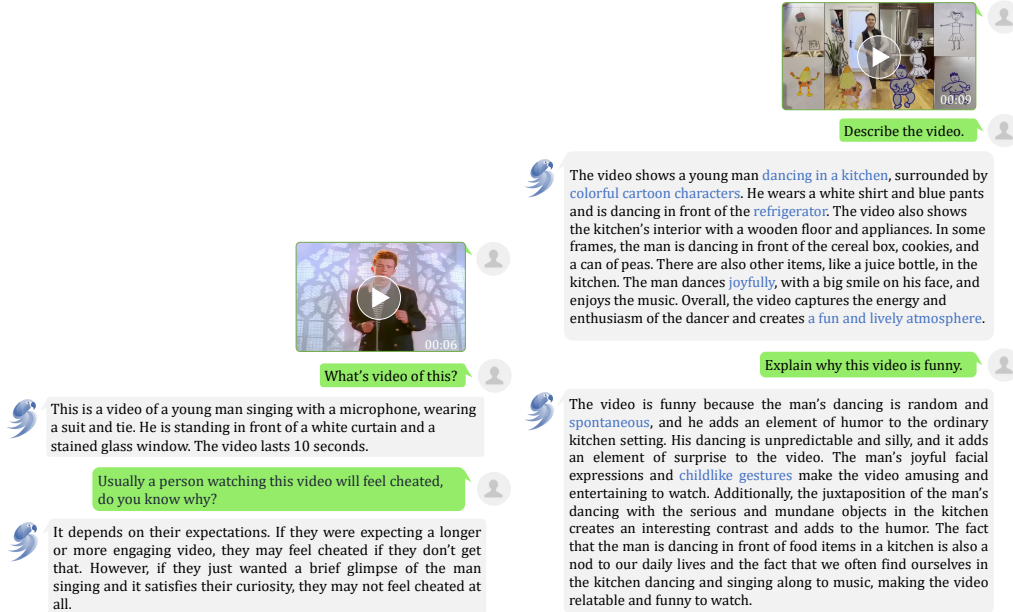


Figure 4: Video meme case. Click here to watch Figure 5: VideoChat can understand why the video is funny. Click here to watch the video.

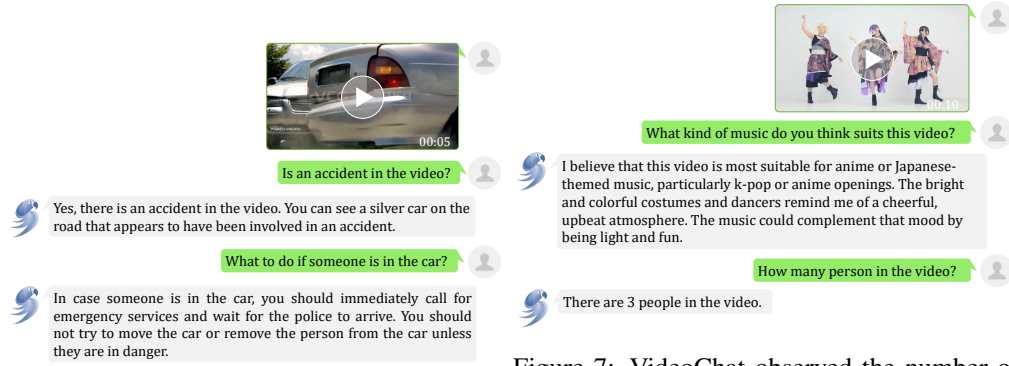


Figure 6: VideoChat finds the accident in the video and gives the advance to the person in the car. Click here to watch the video.

Figure 7: VideoChat observed the number of people and their clothing styles in the space, and perceived temporal actions to recommend light and funny music. Click here to watch the video.

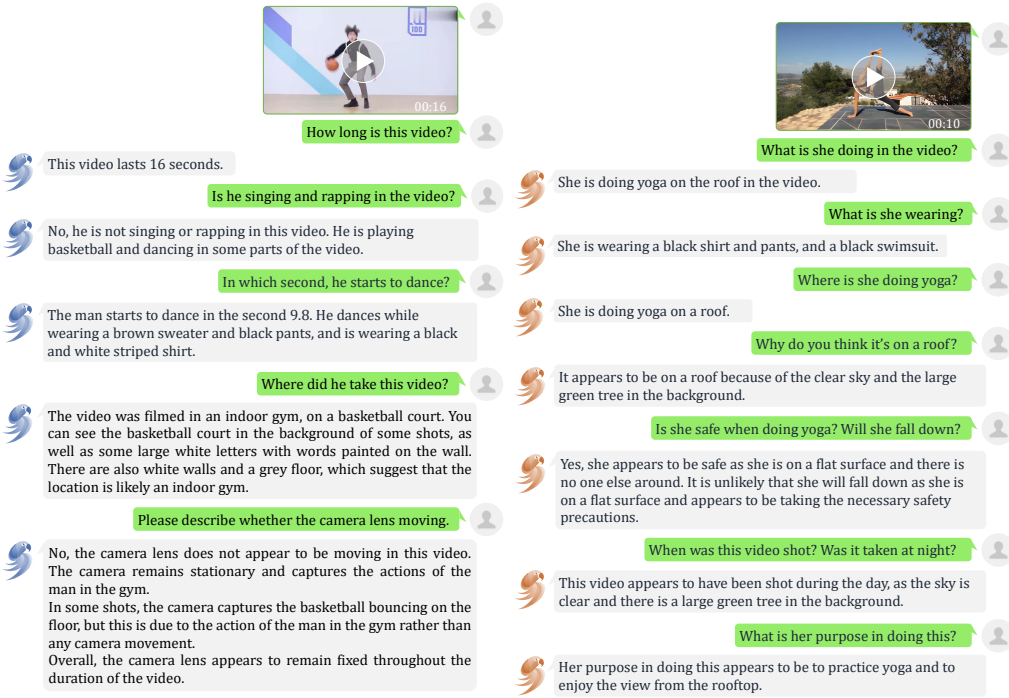


Figure 8: Multi-turn conversation with VideoChat-Embed. Click here to watch the video.

Figure 9: Multi-turn conversation with VideoChat-Text. Click here to watch the video.

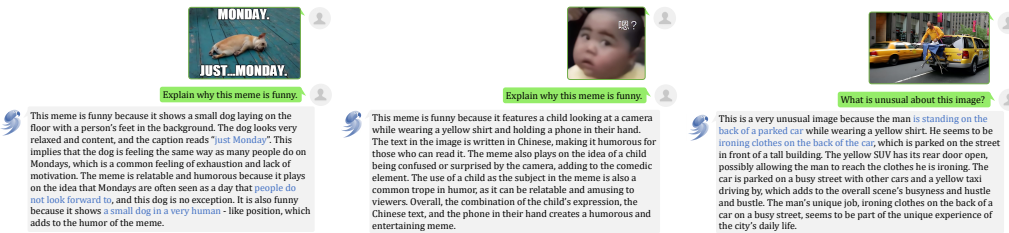


Figure 10: Cases of Image Comprehension. Please zoom in for details.

Future Works Our future works lie in 1) scaling video foundation models both in capacity and data for better spatiotemporal modeling, 2) video-centric multimodal training data and reasoning benchmark for evaluations at scale, and 3) long-term video processing techniques.

References

- [1] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- [5] Guo Chen, Sen Xing, Zhe Chen, Yi Wang, Kunchang Li, Yizhuo Li, Yi Liu, Jiahao Wang, Yin-Dong Zheng, Bingkun Huang, et al. Internvideo-ego4d: A pack of champion solutions to ego4d challenges. *arXiv preprint arXiv:2211.09529*, 2022.
- [6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.
- [8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [9] MOSS contributors. Moss. <https://github.com/OpenLMLab/MOSS>, 2023.
- [10] StableLM contributors. Stablelm: Stability ai language models. <https://github.com/stability-AI/stableLM>, 2023.
- [11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022.
- [12] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021.
- [13] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *CVPR*, 2022.
- [14] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- [15] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.
- [16] Zehuan Huang, Haoran Feng, Chongzhi Zhang, Lu Sheng, Ziwei Liu, and Jing Shao. Dolphin: General video interaction platform based on llms, 2023. <https://github.com/kaleido-lab/dolphin>.
- [17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [20] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022.
- [21] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023.
- [22] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022.
- [23] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *CoRR*, abs/2001.05691, 2020.
- [24] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023.
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.
- [26] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, Kunchang Li, Zhe Chen, Xue Yang, Xizhou Zhu, Yali Wang, Limin Wang, Ping Luo, Jifeng Dai, and Yu Qiao. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. <https://arxiv.org/abs/2305.05662>, 2023.
- [27] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020.
- [28] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.
- [29] OpenAI. Gpt-4 technical report. *arXiv*, 2023.
- [30] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022.
- [31] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.
- [34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hyponymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.
- [36] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.
- [37] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.
- [38] Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *ICCV*, 2019.

- [39] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
- [40] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.
- [42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [43] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022.
- [44] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [45] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [46] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.
- [47] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [48] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [49] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
- [50] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv*, abs/2303.11381, 2023.
- [51] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.
- [52] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.
- [53] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.
- [54] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.
- [55] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 2021.

- [56] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.
- [57] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [58] 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.
- [59] 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.
- [60] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. *CVPR*, 2020.

A Appendix

Instruction for a brief description. Following the brief image instruction in LLaVA, we generate the video instruction with the aid of ChatGPT as shown in Table 8. In Stage1, we randomly sample the instruction to generate brief descriptions of images and videos.

Instruction for a detailed description. Following the detailed image instruction in LLaVA, we generate the video instruction with the help of ChatGPT as shown in Table 9 To build the instruction data used in Stage2, we randomly sample the instruction and combine it with the detailed descriptions.

- "Describe the following image concisely."
 - "Provide a brief description of the given image."
 - "Offer a succinct explanation of the picture presented."
 - "Summarize the visual content of the following image."
 - "Give a short and clear explanation of the subsequent image."
 - "Share a concise interpretation of the image provided."
 - "Present a compact description of the photo's key features."
 - "Relay a brief, clear account of the picture shown."
 - "Render a clear and concise summary of the photo below."
 - "Write a terse but informative summary of the following picture."
 - "Create a compact narrative representing the image presented."
-
- "Describe the following video concisely."
 - "Provide a brief description of the given video clip."
 - "Offer a succinct explanation of the footage presented."
 - "Summarize the visual content of the following video."
 - "Give a short and clear explanation of the subsequent video clip."
 - "Share a concise interpretation of the video provided."
 - "Present a compact description of the clip's key features."
 - "Relay a brief, clear account of the video shown."
 - "Render a clear and concise summary of the video below."
 - "Write a terse but informative summary of the following video clip."
 - "Create a compact narrative representing the video presented."

Table 8: **The list of instructions for brief image and video description.** The image list is inherited from LLaVA [25]. The video list is generated by ChatGPT with examples of the image list.

- "Describe the following image in detail."
 - "Provide a detailed description of the given image."
 - "Give an elaborate explanation of the image you see."
 - "Share a comprehensive rundown of the presented image."
 - "Offer a thorough analysis of the image."
 - "Explain the various aspects of the image before you."
 - "Clarify the contents of the displayed image with great detail."
 - "Characterize the image using a well-detailed description."
 - "Break down the elements of the image in a detailed manner."
 - "Walk through the important details of the image."
 - "Portray the image with a rich, descriptive narrative."
 - "Narrate the contents of the image with precision."
 - "Analyze the image in a comprehensive and detailed manner."
 - "Illustrate the image through a descriptive explanation."
 - "Examine the image closely and share its details."
 - "Write an exhaustive depiction of the given image."
-
- "Describe the following video in detail, including the actions and scenes."
 - "Provide a detailed description of the given video, capturing its key moments."
 - "Give an elaborate explanation of the video you see, including the events and characters."
 - "Share a comprehensive rundown of the presented video, highlighting its main sequences."
 - "Offer a thorough analysis of the video, discussing its various elements and storyline."
 - "Explain the various aspects of the video before you, including the setting and actions."
 - "Clarify the contents of the displayed video with great detail, focusing on its progression."
 - "Characterize the video using a well-detailed description, capturing its essence and events."
 - "Break down the elements of the video in a detailed manner, discussing its key components."
 - "Walk through the important details of the video, describing its scenes and characters."
 - "Portray the video with a rich, descriptive narrative, capturing its atmosphere and events."
 - "Narrate the contents of the video with precision, focusing on its storyline and visuals."
 - "Analyze the video in a comprehensive and detailed manner, discussing its themes and elements."
 - "Illustrate the video through a descriptive explanation, painting a vivid picture of its content."
 - "Examine the video closely and share its details, including the actions, characters, and setting."
 - "Write an exhaustive depiction of the given video, capturing its essence and key moments."

Table 9: **The list of instructions for detailed image and video description.** The image list is inherited from LLaVA [25]. The video list is generated by ChatGPT with examples of the image list.