

DisCo: Towards Distinct and Coherent Visual Encapsulation in Video MLLMs

Jiahe Zhao^{1,2} Rongkun Zheng¹ Yi Wang^{3,4} Helin Wang⁵ Hengshuang Zhao^{1*}

¹The University of Hong Kong ²University of Chinese Academy of Sciences

³Shanghai Artificial Intelligence Laboratory ⁴Shanghai Innovation Institute ⁵Fudan University
 zhaojiahe22@mails.ucas.ac.cn, hszhao@cs.hku.hk

Abstract

In video Multimodal Large Language Models (video MLLMs), the visual encapsulation process plays a pivotal role in converting video contents into representative tokens for LLM input. While linear projectors are widely employed for encapsulation, they introduce semantic indistinctness and temporal incoherence when applied to videos. Conversely, the structure of resamplers shows promise in tackling these challenges, but an effective solution remains unexplored. Drawing inspiration from resampler structures, we introduce **DisCo**, a novel visual encapsulation method designed to yield semantically **distinct** and temporally **coherent** visual tokens for video MLLMs. DisCo integrates two key components: (1) A Visual Concept Discriminator (VCD) module, assigning unique semantics for visual tokens by associating them in pair with discriminative concepts in the video. (2) A Temporal Focus Calibrator (TFC) module, ensuring consistent temporal focus of visual tokens to video elements across every video frame. Through extensive experiments on multiple video MLLM frameworks, we demonstrate that DisCo remarkably outperforms previous state-of-the-art methods across a variety of video understanding benchmarks, while also achieving higher token efficiency thanks to the reduction of semantic indistinctness. The codes will be available at <https://github.com/ZJHTerry18/DisCo>.

1. Introduction

Multi-modal Large Language Models (MLLMs) [2, 3, 9, 19, 35, 44, 68] have spearheaded the advancement of vision-language learning, gaining impressive visual understanding abilities on a myriad of open-world tasks. While the early exploitations were made on image inputs, recent studies have yielded profound breakthroughs on empowering MLLMs for video understanding [10, 30, 37, 49, 54, 60, 81], contributing to a multitude of real-world appli-

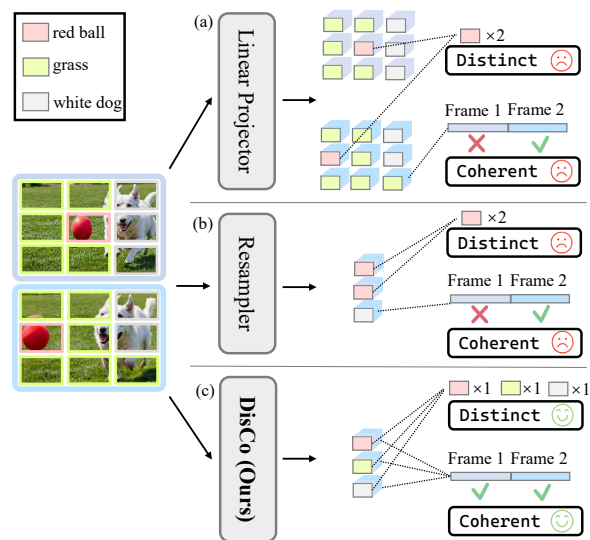


Figure 1. Illustrations of different visual encapsulation methods in video MLLMs. (a) *Linear projector* directly projects tokens of each frame, leading to repetitive semantics for objects appearing in multiple frames, and is incapable of modeling cross-frame temporal coherence. (b) *Resampler* utilizes attention mechanism to derive tokens, which is prone to redundant extraction of same semantics, and cannot guarantee coherent attention across frames. (c) Our proposed *DisCo* can generate high-quality tokens with distinct semantics and coherent temporal correlations.

cations like robotics [51], autonomous driving [66], and AIGC [11]. In contrast to images, video data is characterized by a substantially larger volume of visual information, coupled with inherent temporal complexities, which presents a formidable challenge in effectively encapsulating video inputs to facilitate optimal comprehension by the language model.

For MLLMs, the visual connector [7, 23, 35, 39, 85] emerges as a pivotal component for the encapsulation of visual features into working tokens for LLM. Currently, a major stream of video MLLMs adopt linear projectors [38, 41, 43, 65] for visual encapsulation. While lin-

*Corresponding author.

ear projector proficiently upholds local visual details with simple designs, as shown in Fig. 1(a), applying it to videos usually compromises performance since it introduces semantic indistinctness and temporal incoherence when processing videos. Specifically, the presence of repetitive visual elements across frames leads to redundancy in the projected tokens’ semantics. Moreover, by discretely projecting each visual patch, linear projection fails to encapsulate temporal coherence across frames. More recently, a line of works [7, 16, 39] enhance the information compactness of linear projectors by downsampling or compressing the video patches. Nevertheless, the problems of semantic indistinctness and temporal incoherence are still not relieved due to the locality of their projection mechanisms in both spatial and temporal dimensions.

Different from linear projectors, resamplers [30, 35, 37, 70, 73] exploit cross-attention that transforms video patches into a fixed set of visual tokens. This reduces indistinctness in form and implicitly conducts temporal modeling. However, we note semantic indistinctness and temporal incoherence still exist in the current resampler design. As depicted in Fig. 1(b), in resamplers, there are multiple tokens redundantly focusing on the same semantic instance, while neglecting other crucial instances. Meanwhile, visual tokens display poor temporal coherence by only attending to instances in a part of video frames while neglecting them in other frames. Intuitively, to enable LLMs to accurately comprehend video content, it is crucial to generate high-quality visual tokens representing diverse and distinct semantic concepts while preserving coherent temporal relationships. We argue that the cross-attention mechanism in resamplers is promising for addressing these two limitations, since it allows flexible remodeling of visual cues across spatial and temporal dimensions. The key lies in explicitly guiding this remodeling process towards distinct semantics and coherent temporals, which is absent in current encapsulation techniques.

To this end, we propose **DisCo**, a novel visual encapsulation method that is capable of generating visual tokens with **distinct** semantics and **coherent** temporal cues, as depicted in Fig. 1(c). DisCo features two principal designs: (i) A Visual Concept Discriminator (VCD) module, which aligns each visual token with a distinct semantic concept. Diverging from previous encapsulation methods that uniformly align all visual tokens with the entire video caption, VCD dynamically assigns different visual tokens to discrete text instances extracted from video descriptions. This approach reduces token redundancy and enhances semantic diversity. (ii) A Temporal Focus Calibrator (TFC) module, which aligns the focused instance of each visual token across the temporal dimension. Unlike previous methods that only align visual tokens at video-level, TFC dives into frame-level calibrations between visual tokens and video

instances. We introduce a Frame-level Focus Alignment (FFA) loss to guide each visual token to remain aligned with its designated semantic instance throughout each video frame, ensuring temporal coherence across the video. Extensive experiments demonstrate that DisCo achieves state-of-the-art performances on video understanding. Moreover, by reducing information redundancies in visual encapsulation process, DisCo could improve the efficiency of video MLLMs by utilizing 75% less tokens while maintaining overall performance.

We summarize our contributions as follows:

- We propose DisCo, the first visual encapsulation method that is capable of generating semantically distinct and temporally consistent visual tokens for video LLMs, greatly promoting the quality of visual representations in video-language learning.
- In DisCo, a Visual Concept Discriminator (VCD) module is raised to endow visual tokens with unoverlapped semantic concepts, facilitating semantic distinctiveness in visual representations. Additionally, a Temporal Focus Calibrator (TFC) module is introduced to realize frame-level attention on video instances, ensuring the temporal coherence in visual tokens.
- As a plug-and-play design, DisCo is compatible with various video MLLM frameworks. Extensive experiments on multiple baselines demonstrate the superiority and efficiency of DisCo across a wide spectrum of video understanding benchmarks.

2. Related Works

Multimodal Large Language Models. With the significant advances in Large Language Models (LLMs) [5, 17, 20, 47, 80], there is a surge of investigations on exploring Multi-modal Large Language Models (MLLMs) [2, 33, 74, 86], as they can handle a diverse range of open-ended tasks [26, 42, 53]. Seminal works like Flamingo [2] effectively unified the understanding of vision and text modalities, showing impressive performance on a wide range of multi-modal tasks. Recently, a line of open-source MLLMs like LLaVA [43], Qwen-VL [57] and MM-ICL [78] further incorporate visual instruction tuning data [31, 63, 71] to enhance visual dialogue ability. Based on the success of perceiving static images, several studies leverage extensive video-text data corpus [4, 36, 84] to construct video MLLMs, such as VideoChat [37], Video-ChatGPT [48] and InternVideo [60]. Despite their outstanding capabilities in open-world video understanding [12, 61, 72, 79, 82], recent video MLLMs have not yet deeply explored visual connectors, which hold a critical role in deciding the performance and efficiency of MLLMs. In this study, we investigate developing a visual encapsulation method that contributes to a well-performed and efficient video MLLM.

Visual Encapsulation in MLLM. Visual encapsulation is

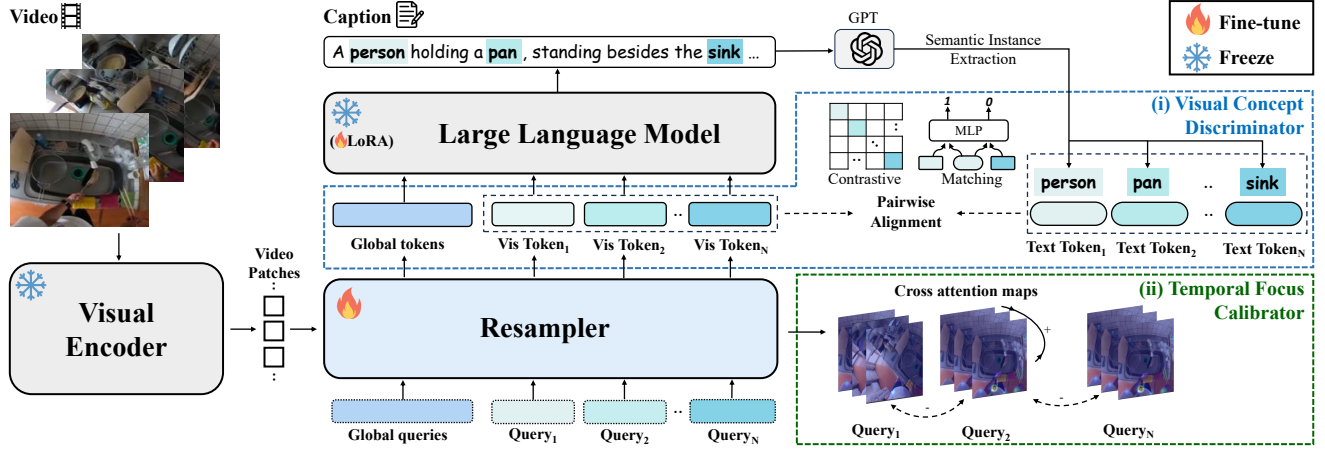


Figure 2. The overall structure of DisCo. DisCo is highlighted by (i) a Visual Concept Discriminator (VCD) module, which conducts a pairwise alignment between visual tokens and semantic concepts, to acquire distinct semantics, and (ii) a Temporal Focus Calibrator (TFC) module, which aligns frame-level focused areas within each visual token, to improve temporal coherence.

a crucial process in multi-modal large language models that bridges visual encoders with LLMs. Among major visual encapsulation methods, linear projection [8, 13, 41, 43, 65] is most widely utilized. This design fully preserves visual information, but leads to high computational load due to the large number of visual patches. Another type of encapsulation uses a resampler [13, 35, 58, 68, 73] to compress the visual patches into a much smaller number of tokens, at the cost of sacrificing the comprehensiveness of visual cues. To achieve token efficiency as well as preserve detailed visual information, works like DeCo [67] and Token-Packer [39] presented token downsampling modules, while Slot-VLM [64] adopted slot attention [46] to capture object-level information. However, the visual representations from these models still lack semantic clarity and temporal coherence. In this work, we address these issues by raising DisCo, a visual encapsulation method that learns semantically distinct and temporally coherent video tokens.

3. Method

As illustrated in Fig. 2, we propose DisCo, a novel visual encapsulation method designed to generate semantically distinct and temporally coherent visual tokens for video MLLMs. DisCo is highlighted by two primary components: (i) A Visual Concept Discriminator (VCD) module, which aligns a set of visual tokens with a group of semantic concepts in a pairwise manner, achieving distinct semantics. (ii) A Temporal Focus Calibrator (TFC) module, which extracts frame-level focused features of visual tokens, and aligns these features across all frames, to ensure coherent temporal attentions. In Sec. 3.1, we will first provide the preliminaries on the structure of DisCo. Then, we will introduce the VCD and TFC modules, in Sec. 3.2

and Sec. 3.3, respectively. Finally, we describe the training scheme in Sec. 3.4.

3.1. Preliminaries

In the video MLLM family, a group of models employ resamplers for visual encapsulation. These models are structurally composed of three main components: a visual encoder, a resampler, and a large language model (LLM).

Visual Encoder. Given a video input sampled into T frames $X = \{x_i\}_{i=1}^T$, a ViT [18] $\mathcal{V}$ is utilized to extract deep video features $V = \{v_i \in \mathbb{R}^{n \times c}\}_{i=1}^T$.

Resampler. Serving as a bridge between the visual encoder and the LLM, in the resampler (e.g., Q-Former [35]), a set of learnable query embeddings $X_q = \{q_i\}_{i=1}^N$ is initialized to interact with video features V through cross-attention layers [56]. This interaction produces a set of visual tokens, denoted as $X_v = \{v_i\}_{i=1}^N$, which contain encapsulated visual representations.

Large Language Model (LLM). Large language model acts as a unified platform to process both vision and language inputs, generating natural language answers accordingly. LLM takes the output tokens of resampler X_v as vision input, and a paired text instruction X_i as language input. The entire video LLM is trained by minimizing the negative log-likelihoods of generating the target answer X_a :

$$\mathcal{L}_{llm} = -\mathbb{E}_{X \sim \mathcal{D}} \left[\sum_{l=1}^L \log p(X_a^l | X_v, X_i^{<l}, X_a^{<l}) \right], \quad (1)$$

where $\mathcal{D}$ denotes the training dataset, and $X_i^{<l}$, $X_a^{<l}$ denotes the instruction and answer tokens before the current generated token X^l .

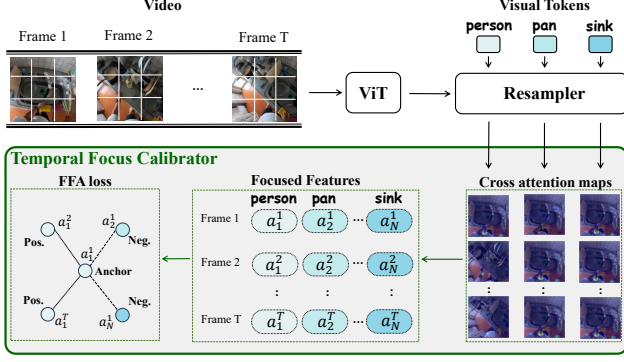


Figure 3. The structure of the TFC module. In TFC, frame-wise focused features are drawn from the cross attention maps in the resampler. Then, the Frame-level Focus Alignment (FFA) loss aligns each frame-wise feature within each visual token, promoting the temporal coherence across video frames.

3.2. Visual Concept Discriminator

In existing resampler-based video MLLMs, visual tokens produced by the resampler often endure semantic indistinctness, with multiple tokens representing the same element. We argue that this issue arises from the lack of explicit guidance on the element-wise contents of each visual token, leading to repetitive semantic information among tokens.

To address this problem, we propose a Visual Concept Discriminator (VCD). It distinguishes itself from previous encapsulation methods by explicitly aligning different visual tokens with distinct semantic concepts in a pairwise manner. To implement this pairwise alignment, both visual tokens and semantic concepts are initially divided into multiple groups. As shown in Fig. 2, for visual tokens, we reorganize the total of N visual tokens into N_g groups, denoted as $\{\hat{v}_i\}_{i=1}^{N_g}$, with each group $\hat{v}_i$ comprising N/N_g tokens. For semantic concepts, we leverage GPT-4 [1] to extract distinct words or phrases that each represent a specific instance in the video caption, forming a set of M semantic concepts. The text embeddings for these semantic concepts are then generated using the resampler’s text processing branch, resulting in embeddings $\{t_j\}_{j=1}^M$.

To achieve a one-to-one alignment between visual token groups and semantic concepts, we perform bipartite matching [6] between visual tokens and text embeddings. For simplicity, we assume $N_g = M$. The bipartite matching algorithm determines a permutation of M elements $\hat{\sigma} \in \mathcal{P}_M$, which pairs i -th visual token with $\hat{\sigma}(i)$ -th semantic concept with lowest cost:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathcal{P}_M} \sum_{i=1}^M c(v_i, t_{\sigma(i)}), \quad (2)$$

where $c(x, y)$ denotes the cosine distance between x and y . This assignment is computed using the Hungarian algo-

Table 1. Comparison of different visual encapsulation methods. DisCo is the first to combine the traits of **Distinct**: semantic distinction, **Coherent**: temporal coherence, **Complete**: information completeness and **Efficient**: token efficiency.

Encapsulation Methods	Distinct	Coherent	Complete	Efficient
Linear Projector	✗	✗	✓	✗
Resampler	✗	✗	✗	✓
DisCo (Ours)	✓	✓	✓	✓

arithm. In circumstances where $N_g \neq M$, our process yields $\min(N_g, M)$ matching pairs, while leaving the excessive visual or textual elements unused in the VCD module.

Upon establishing the one-to-one matching, we use pairwise losses to facilitate learning the alignment between each pair of visual and semantic features. Matched visual-semantic pairs are treated as positive pairs, while others are considered as negative pairs. Following vision-language alignment techniques in [34, 55], we apply a visual-semantic pairwise contrastive (VSC) loss and a visual-semantic matching (VSM) loss, denoted as $\mathcal{L}_{vsc}$ and $\mathcal{L}_{vsm}$, respectively. The VSC loss is defined as:

$$\mathcal{L}_{vsc} = - \sum_i^M \left[\log \frac{S(v_i, t_{\hat{\sigma}(i)})}{\sum_j S(v_i, t_{\hat{\sigma}(j)})} + \log \frac{S(t_{\hat{\sigma}(i)}, v_i)}{\sum_j S(t_{\hat{\sigma}(i)}, v_j)} \right], \quad (3)$$

where $S(v, t) = \exp(\frac{v^T t}{\tau \|v\| \|t\|})$ denotes visual-text similarity score with temperature τ . The VSM loss is expressed as:

$$\mathcal{L}_{vsm} = \sum_i^M \text{CE}(p_\theta(v, t), y_{v,t}), (v, t) \sim (v_i, t_j), \quad (4)$$

where $\text{CE}(p, y)$ denotes the cross-entropy loss between prediction p and ground-truth label y . An MLP is utilized to predict $p_\theta(v, t) = \text{MLP}([v, t])$.

Since the extracted semantic concepts do not completely contain the original video caption (e.g., the term “holding” is not included as shown in Fig. 2), the N_g groups of aligned visual tokens cannot cover complete video information. To ensure comprehensive visual representation, we add a set of global tokens into VCD to capture this uncovered information and preserve the integrity of the visual cues.

3.3. Temporal Focus Calibrator

Despite the improvements brought by VCD, existing resamplers still face the challenge of temporal incoherence. Delving into their mechanisms, it is revealed that each visual token is uniformly attended to video patches from all frames. This approach fails to ensure that each visual token consistently focuses on every individual video frame. As a solution, we introduce a Temporal Focus Calibrator (TFC) module, which pioneers frame-level calibration for resamplers in video MLLMs. The primary aim of TFC is to

Table 2. Performance on video question-answering benchmarks. ‘val’ denotes validation set for PerceptionTest, and ‘subset’ denotes the subset for EgoSchema test set. The best result of each benchmark is boldfaced.

Model	Size	MVBench	STAR	PerceptionTest	EgoSchema	MLVU	VideoMME (w/o & w. sub)			
				val	subset		overall	short	medium	long
Otter-V [32]	7B	26.8	-	-	-	16.7	-	-	-	-
VideoLLaMA [73]	7B	33.6	26.3	36.5	25.6	-	26.5/37.1	25.7/27.8	25.1/35.6	28.6/38.1
VideoChat2 [37]	7B	35.5	59.0	-	64.6	-	39.5/43.8	48.3/52.8	37.0/39.4	33.2/39.2
LLaMA-VID [40]	7B	41.3	-	-	-	18.1	25.9/-	-	-	-
VideoLLaVA [41]	7B	43.0	-	-	-	29.3	39.9/41.6	45.3/46.1	38.0/40.7	36.2/38.1
LLaVA-Mini [76]	7B	44.5	-	-	-	42.8	-	-	-	-
LongLLaVA [59]	9B	49.1	-	-	-	-	43.7/-	-	-	-
ShareGPT4Video [10]	8B	51.2	-	-	-	34.2	39.9/43.6	48.3/53.6	36.3/39.3	35.0/37.9
LLaVA-NeXT-Video [77]	7B	53.1	35.5	48.8	49.1	-	37.3/43.7	39.3/47.8	38.9/46.9	33.9/36.2
VideoLLaMA2 [14]	7B	54.6	57.2	51.4	51.7	48.5	47.9/50.3	54.3/56.1	44.3/47.4	40.1/45.7
VideoChat2-HD [37]	7B	62.3	63.9	54.3	65.6	47.9	45.3/55.7	53.4/59.2	47.3/54.0	37.1/46.7
ST-LLM [45]	7B	54.7	56.7	49.5	55.2	46.7	40.6/-	49.9/-	40.2/-	31.5/-
ST-LLM+DisCo	7B	58.0	60.1	54.4	59.8	48.6	42.1/-	51.8/-	39.6/-	34.8/-
InternVideo2 [60]	7B	60.3	64.5	52.6	64.4	43.9	41.7/51.7	50.3/56.7	37.4/50.1	37.3/48.0
InternVideo2+DisCo	7B	63.3	72.7	61.7	66.2	46.7	42.9/52.8	53.0/59.5	38.7/50.0	37.0/48.7
InternVideo2-HD [60]	7B	66.3	75.7	62.4	67.0	47.1	46.3/56.7	54.5/59.5	42.4/55.3	42.0/55.3
InternVideo2-HD+DisCo	7B	68.2	77.7	67.4	72.2	49.5	47.4/57.9	55.8/61.3	43.8/56.1	42.7/56.2

explicitly supervise each visual token to focus on its corresponding semantic concepts in each frame.

As illustrated in Fig. 3, the initial step of TFC involves extracting the focused features from the cross-attention maps between visual tokens and the video features extracted by ViT. Specifically, for the i -th token, we denote its cross-attention maps with the t -th video frame as $\{C_k^t \in \mathbb{R}^{h \times w}\}_{k=1}^{L_c}$, where L_c is the total number of cross-attention layers. Then, the attention features are as follows:

$$a_i^t = \text{AvgPool}\left(\frac{1}{L_c} \sum_{k=1}^{L_c} C_k^t \cdot V^t\right), i = 1 \sim N_g, t = 1 \sim T, \quad (5)$$

where $\text{AvgPool}(\cdot)$ denotes average pooling along the spatial dimensions, and V^t is the video feature of t -th frame.

To achieve alignment of frame-wise attention features within each visual token, we present a Frame-level Focus Alignment (FFA) loss: given attention feature a_i^t as an anchor, FFA loss pulls a_i^t closer to the attention features of other frames within the i -th token, while pushes a_i^t apart from the attention features of other tokens. Moreover, to improve the stability of frame-wise attention features (particularly in cases where an object may temporarily disappear in some frames), we utilize the feature centroid of each query, defined as $\bar{a}_i = \frac{1}{T} \sum_{t=1}^T a_i^t$. The centroid feature provides a more robust reference for alignment in the FFA loss. Finally, the loss is formulated as follows:

$$\mathcal{L}_{ffa} = - \sum_i \sum_t \left[\log \frac{S(\bar{a}_i, a_i^t)}{\sum_j S(\bar{a}_i, a_j^t)} + \log \frac{S(a_i^t, \bar{a}_i)}{\sum_j S(a_i^t, \bar{a}_j)} \right]. \quad (6)$$

3.4. Training

Following standard training strategies of MLLMs, our training process consists of two stages. Stage 1 focuses on vision-text alignment. In this stage, we leverage a substantial dataset of visual dense captions to align the visual tokens of DisCo with the LLM. Additionally, the VCD and TFC modules are incorporated in this stage. The total training loss is formulated as:

$$\mathcal{L}_{stage1} = \mathcal{L}_{llm} + \lambda_{vsc} \mathcal{L}_{vsc} + \lambda_{vsm} \mathcal{L}_{vsm} + \lambda_{ffa} \mathcal{L}_{ffa}. \quad (7)$$

where λ_{vsc} , λ_{vsm} , and λ_{ffa} are weight parameters. After completing Stage 1, we advance to Stage 2, the instruction tuning stage. In this stage, we utilize a diverse set of image and video caption and question-answer (QA) data to equip the model with strong instruction following ability.

3.5. Discussion

Now we illustrate the difference between DisCo and existing visual encapsulation methods. As shown in Tab. 1, all previous methods endure indistinctness in token semantics, and incoherence in temporal modeling. Instead, DisCo encapsulates the visual token with two defining attributes: (i) **Semantic distinction**: each visual token represents unoverlapped instances, possessing clear semantic difference. (ii) **Temporal coherence**: each visual token attends the dynamics of its corresponding instance at every frame. Moreover, by reducing overlapped semantics, DisCo achieves: (iii) better **Information completeness** by covering more visual elements, and (iv) **Token efficiency** by utilizing less tokens to represent the same amount of visual cues.

Table 3. Comparison with state-of-the-art methods on video conversation benchmarks. ‘CI’, ‘DO’, ‘CU’, ‘TU’, and ‘CO’ denote ‘Correctness of Information’, ‘Detail Orientation’, ‘Context Understanding’, ‘Temporal Understanding’, and ‘Consistency’.

Model	CI	DO	CU	TU	CO	Avg
VideoLLaMA [73]	1.96	2.18	2.16	1.82	1.79	1.98
VideoChatGPT [48]	2.40	2.52	2.62	1.98	2.37	2.38
VideoChat2 [37]	3.02	2.88	3.51	2.66	2.81	2.88
LLaMA-VID [40]	2.96	3.00	3.53	2.46	2.51	2.89
LLaVA-Mini [76]	2.97	2.99	3.61	2.48	2.67	2.94
Chat-UniVi [28]	2.89	2.91	3.46	2.89	2.81	2.99
InternVideo2 [60]	2.88	2.53	3.20	2.51	2.67	2.76
InternVideo2+DisCo	3.13	2.65	3.42	2.56	2.89	2.93
InternVideo2-HD [60]	3.14	2.74	3.53	2.52	2.85	2.96
InternVideo2-HD+DisCo	3.36	3.20	3.76	2.80	3.10	3.24

4. Experiments

Implementation Details. DisCo functions as a plug-and-play module, designed to generally enhance resampler-based video MLLMs. To assess its integration capabilities across different frameworks, we implemented DisCo on two video MLLMs: ST-LLM [45] and InternVideo2 [60]. ST-LLM employs the ViT-G/14 model from EVA-CLIP [21] as its visual encoder and utilizes Vicuna-7B-v1.1 [15] as its LLM. InternVideo2 utilizes InternVideo2-1B as its visual encoder and Mistral-7B [27] for LLM. Both models incorporate Q-Former [34] as the visual connector. Throughout both training stages, we freeze the visual encoder, update the resampler, and fine-tune the LLM using LoRA [25].

In implementing DisCo, for ST-LLM, 8 of the 32 pre-trained query tokens in the resampler are designated as global tokens. The remaining 24 tokens are distributed across $N_g = 12$ visual token groups, each comprising 2 tokens, to ensure comprehensive coverage of each semantic concept. For InternVideo2, 32 of the 96 tokens are assigned as global tokens, and the rest are set into $N_g = 16$ groups. In Eq. (7), we set λ_{vsc} , λ_{vsm} and λ_{ffa} at 1.0.

Datasets. We adopt a wide scope of video captioning and question-answering (QA) data sources for the training of DisCo. In stage 1, we utilize 900K video dense captions from ShareGPTVideo [75], as well as 23K image captions from LLaVA [43]. In stage 2, our approach aligns with the instructional tuning protocols inherent to the foundational Video MLLMs upon which DisCo is based. Specifically, for the ST-LLM-based DisCo, we incorporate WebVid [4], NexT-QA [62], CLEVRER [69], Kinetics-710 [29] and Something-Something-v2 [24]. For the InternVideo2-based DisCo, we adhere to the recipe used in VideoChat2 [37].

Evaluation Metrics. For video question-answering (QA) benchmarks, the accuracy of the model’s responses is assessed using multiple-choice formats. This close-ended approach enhances objectivity and fairness for evaluation. For video conversation benchmarks, we utilize GPT [1] to as-

Table 4. Ablations on the key components of DisCo. ‘SFT’ denotes using the same training corpus as DisCo to directly fine-tune the baseline model. ‘VCD (w)’ and ‘TFC (w)’ denotes adding VCD and TFC to the baseline, respectively. EgoSchema is validated on the *subset*.

Methods	MVBench	STAR	Egoschema
Baseline	66.3	75.7	67.0
SFT	66.5	76.3	68.2
VCD (w)	67.6	77.5	71.8
TFC (w)	67.1	77.4	70.4
DisCo	68.2	77.7	72.2

sign scores for each answer, enabling multi-angled assessments such as detailedness and consistency.

4.1. Comparison with State-of-the-arts

We present quantitative evaluations of our proposed DisCo in comparison to state-of-the-art methods across a broad array of video QA benchmarks, including: (1) Short video benchmarks STAR [61] and PerceptionTest [52] with focus on fine-grained visual details. (2) Long video benchmarks EgoSchema [50] and MLVU [83], stressing complex temporal relationships. (3) Comprehensive benchmarks MVBench [37] and VideoMME [22], covering diverse video QA tasks. To test the capability of DisCo on video conversations, we also validate DisCo on the VideoChatGPT-Bench [48].

As depicted in Tab. 2, DisCo consistently enhances the performance of video MLLMs on the video QA benchmarks with various video lengths, question granularity and task diversity. Moreover, the introduction of DisCo consistently improves the performance of ST-LLM and InternVideo2. As shown in Tab. 3, DisCo consistently outperforms current state-of-the-art methods on video conversation benchmarks. This result validates the comprehensive enhancement DisCo brings to video MLLMs.

In Fig. 4, we present qualitative examples of DisCo. (a) and (b) shows that DisCo possesses stronger abilities on grabbing detailed visual cues like object colors and water steam, leading to better results on detailed understanding. (c) and (d) proves that DisCo captures temporal events more coherently (add wood to fire, arm movements), performing better on temporal reasoning. (e) shows that DisCo yields video captions with more sufficient and fine-grained visual details, demonstrating its superiority on video captioning.

4.2. Ablation Studies

We conduct a thorough analysis on the effectiveness of the primary components and key designs in DisCo. More ablations could be found in the supplementary materials.

Effectiveness of major components. The implementation of DisCo comprises two key components: the Visual

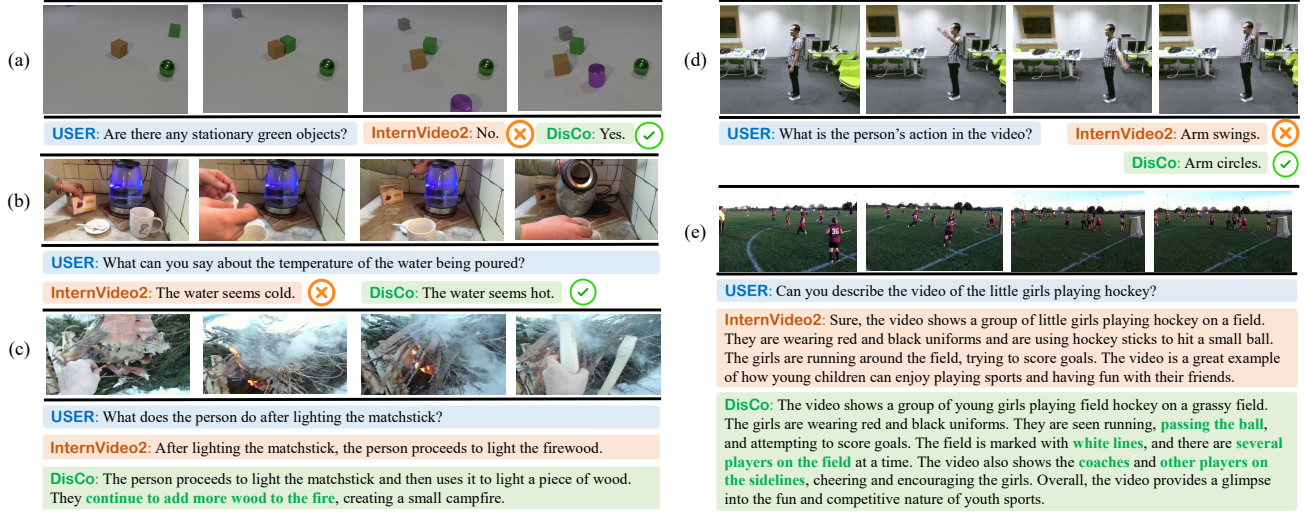


Figure 4. Qualitative examples of video understanding. Utilizing DisCo, video MLLMs achieve (a)(b) better correctness, (c)(d) stronger temporal coherence and (e) richer details in video captioning and QA tasks.

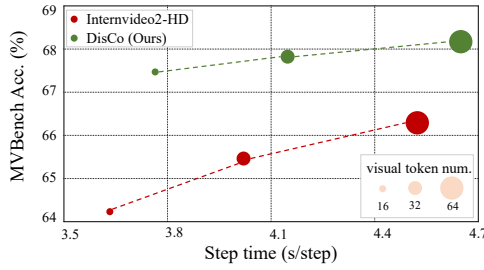


Figure 5. Performance and efficiency under different number of visual tokens. We report performance on MVBench. Efficiency is measured by the average time of each training step.

Table 5. Ablations on the VCD module. ‘ N_g ’ denotes the number of visual token groups, and ‘ N/N_g ’ denotes the number of tokens in each group. ‘Global’ stands for global tokens.

N_g	N/N_g	Global	$\mathcal{L}_{vtc}$	$\mathcal{L}_{vtm}$	MVBench	STAR
4	16	✓	✓	✓	66.5	77.1
64	1	✓	✓	✓	67.3	77.5
16	4		✓	✓	65.8	76.4
16	4	✓	✓		67.4	76.7
16	4	✓		✓	66.8	75.7
16	4	✓	✓	✓	68.2	77.7

Concept Discriminator (VCD) and the Temporal Focus Calibrator (TFC) modules. To assess the effectiveness of each component, we conduct an ablation on these modules. The results, as shown in Tab. 4, indicate that both the VCD and TFC modules contribute significant performance improvements across all three benchmarks.

Furthermore, to ensure that these performance gains originate from the module designs rather than the integra-

Table 6. Ablations on the TFC module. ‘Frame-wise feat’ denotes only using frame-wise attention features when implementing the FFA loss (Eq. (6)). ‘Feat. centroid’ denotes using frame-level average in the FFA loss. EgoSchema is validated on the *subset*.

Methods	MVBench	STAR	EgoSchema
Frame-wise feat.	67.7	76.4	71.0
Feat. centroid (DisCo)	68.2	77.7	72.2

tion of new data, we utilize the same training corpus as DisCo to directly fine-tune the baseline model, resulting in the SFT model. From the results presented in Tab. 4, it is evident that both VCD and TFC achieve higher accuracy compared to SFT by a substantial margin, thereby strongly affirming the efficacy of our component designs.

Improvement on Token Efficiency. As DisCo contributes to alleviating information redundancy in the visual tokens, we explore the potentials of DisCo on improving token efficiency. To this end, we conduct experiments on InternVideo2-HD by varying the number of local visual tokens. From the results in Fig. 5, we can conclude that DisCo could maintain its performance when the token number decreases, and a 16-token DisCo even outperforms a traditional resampler with 64 tokens. This proves that DisCo holds great promise on mitigating training and inference costs. Meanwhile, DisCo only introduces minor training consumptions over resamplers, with training time increasing by less than 5% when token numbers are the same.

Ablations on key designs of VCD. The VCD module is designed to mitigate semantic redundancy in visual tokens by aligning group-wise visual tokens with diverse text instances. Our investigations reveal that the number of groups

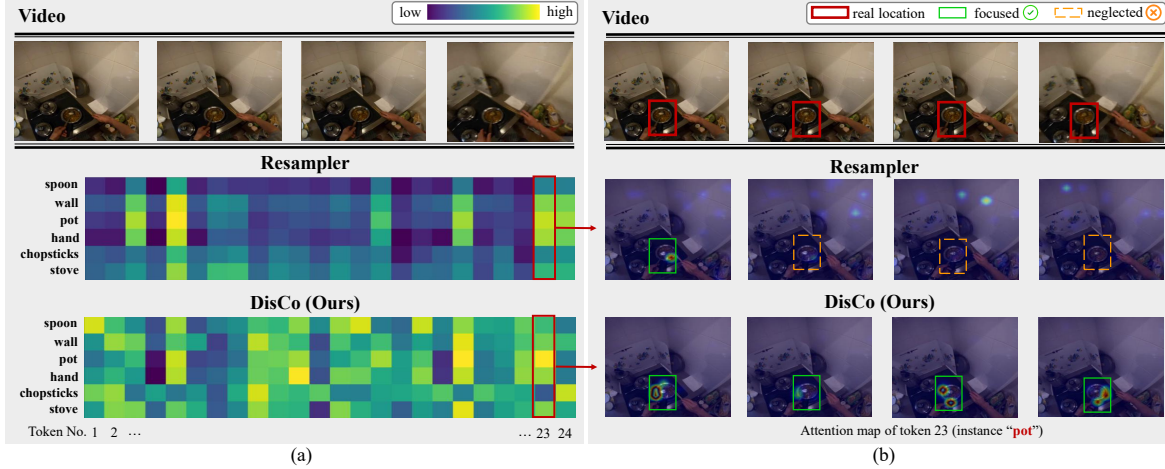


Figure 6. (a) Similarity matrix between visual tokens and text instances. Resamplers exhibit severe semantic redundancy across tokens, while DisCo achieves distinct semantics. (b) Attention maps between the visual token representing ‘pot’ and each video frame. Resamplers fail to consistently attend to the instance ‘pot’, while DisCo attends to it in every frame, demonstrating good temporal coherence.

(N_g) and the number of tokens per group significantly impact VCD’s performance. As demonstrated in Tab. 5, reducing the number of groups to $N_g = 4$ results in a performance decrease of 1.7% on MVBench, indicating that a limited number of discrete visual token groups impairs the model’s ability to capture rich semantic details. Conversely, increasing N_g to 64 also leads to a performance decline, possibly due to insufficient tokens per group, which compromises the informational completeness of visual tokens for each semantic concept. To this end, we choose the optimal N_g as 16. This setting guarantees that the visual token groups could cover most of the text samples (with an average instance number of 9.96, while only 4.95% excess 16 instances), while also making sure there are not too many unused token groups during training. In addition, VCD incorporates a set of global tokens aimed at capturing global information that may be overlooked by the visual token groups. Tab. 5 shows that the existence of global tokens is crucial for DisCo to achieve higher performance.

To further illustrate the effectiveness of VCD, in Fig. 6(a), we present the similarity matrix between visual tokens and semantic concepts. It is clear from the visualization that the tokens from resamplers exhibit severe redundancy, with multiple tokens aligning to ‘wall’, ‘pot’, while instances like ‘spoon’ are almost ignored by all tokens. In contrast, DisCo guides different visual tokens to highlight distinct semantics, and endows visual tokens with more comprehensive representation of the video content.

In Tab. 5, we also validate the necessity of VSC and VSM losses. Performance on MVBench declines by 1.4% and 0.8% when only using VSC or VSM loss, respectively. This proves the effectiveness of utilizing both losses.

Ablations on key designs of TFC. In the TFC module,

to provide a robust foundation for calculating the FFA loss, we employ feature centroids derived from each visual token for contrastive learning. To assess the effectiveness of this approach, we compare the performance of using feature centroids across frames against using frame-wise features. As presented in Tab. 6, the use of feature centroids consistently yields performance improvements across all three evaluation benchmarks. These results underscore the effectiveness of employing feature centroids to enhance temporal consistency in visual token alignment.

To better demonstrate the effectiveness of TFC, we visualize the cross-attention maps between a visual token group and all video frames in Fig. 6(b). We can observe that when using resamplers, the token that highlights the instance ‘pot’ only attends to the pot in the first two frames. Conversely, DisCo consistently tracks the pot across all frames. This illustrates the effectiveness of utilizing TFC on improving the temporal coherence of visual tokens.

5. Conclusion

This paper proposes DisCo, a visual encapsulation method that first builds *semantically distinct* and *temporally coherent* visual tokens for video MLLMs. By incorporating a novel Visual Concept Discriminator (VCD) module and a Temporal Focus Calibrator (TFC) module, DisCo generates visual tokens with distinct semantic information and robust temporal coherence. Extensive experiments verify that DisCo attains state-of-the-art performance and remarkable efficiency on diverse video understanding benchmarks.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. 62422606, 62201484), and National Key R&D Program of China (No. 2022ZD0160101).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 4, 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. In *NeurIPS*, 2022. 1, 2
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv:2308.12966*, 2023. 1
- [4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2, 6
- [5] Tom B Brown. Language models are few-shot learners. In *NeurIPS*, 2020. 2
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4
- [7] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *CVPR*, 2024. 1, 2
- [8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunsang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*, 2023. 3
- [9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023. 1
- [10] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv:2406.04325*, 2024. 1, 5
- [11] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 1
- [12] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *arXiv:2312.06722*, 2023. 2
- [13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 3
- [14] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv:2406.07476*, 2024. 5
- [15] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhonghao 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. 6
- [16] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv:2402.03766*, 2024. 2
- [17] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018. 2
- [18] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 3
- [19] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv:2303.03378*, 2023. 1
- [20] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024. 2
- [21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 6
- [22] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024. 6
- [23] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv:2307.08041*, 2023. 1
- [24] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 6
- [25] 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. In *ICLR*, 2022. 6
- [26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2
- [27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv:2310.06825*, 2023. 6

- [28] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024. 6
- [29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 6
- [30] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *ECCV*, 2024. 1, 2
- [31] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv:2306.05425*, 2023. 2
- [32] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: a multi-modal model with in-context instruction tuning. *arXiv:2305.03726*, 2023. 5
- [33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*, 2024. 2
- [34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4, 6
- [35] 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 *ICML*, 2023. 1, 2, 3
- [36] KunChang Li, Yanan 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
- [37] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 1, 2, 5, 6
- [38] Wei Li, Hehe Fan, Yongkang Wong, Mohan Kankanhalli, and Yi Yang. Topa: Extend large language models for video understanding via text-only pre-alignment. *arXiv:2405.13911*, 2024. 1
- [39] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Token-packer: Efficient visual projector for multimodal llm. *arXiv:2407.02392*, 2024. 1, 2, 3
- [40] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2024. 5, 6
- [41] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122*, 2023. 1, 3, 5
- [42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2
- [43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 6
- [44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1
- [45] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *ECCV*, 2024. 5, 6
- [46] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020. 3
- [47] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*, 2023. 2
- [48] 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
- [49] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv:2406.09418*, 2024. 1
- [50] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023. 6
- [51] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. In *NeurIPS*, 2024. 1
- [52] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2024. 6
- [53] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2
- [54] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *arXiv:2405.16009*, 2024. 1
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3
- [57] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's

- perception of the world at any resolution. *arXiv:2409.12191*, 2024. 2
- [58] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, 2024. 3
- [59] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv:2409.02889*, 2024. 5
- [60] 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. In *ECCV*, 2024. 1, 2, 5, 6
- [61] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021. 2, 6
- [62] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 6
- [63] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2
- [64] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-vlm: Slowfast slots for video-language modeling. *arXiv:2402.13088*, 2024. 3
- [65] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv:2404.16994*, 2024. 1, 3
- [66] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *RAL*, 2024. 1
- [67] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv:2405.20985*, 2024. 3
- [68] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024. 1, 3
- [69] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020. 6
- [70] 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. In *ICLR*, 2024. 2
- [71] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2
- [72] Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *CVPR*, 2023. 2
- [73] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023. 2, 3, 5, 6
- [74] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv:2407.03320*, 2024. 2
- [75] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv:2404.01258*, 2024. 6
- [76] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *ICLR*, 2025. 5, 6
- [77] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. 2024. 5
- [78] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojuan Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv:2309.07915*, 2023. 2
- [79] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? In *ICLR*, 2024. 2
- [80] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanhao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023. 2
- [81] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. Syncvis: Synchronized video instance segmentation. In *NeurIPS*, 2024. 1
- [82] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. Villa: Video reasoning segmentation with large language model. *arXiv:2407.14500*, 2024. 2
- [83] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv:2406.04264*, 2024. 6
- [84] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2
- [85] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 1
- [86] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv:2312.09251*, 2023. 2