Task Preference Optimization: Improving Multimodal Large Language Models with Vision Task Alignment

Ziang Yan[*2,1], Zhilin Li[*3,1], Yinan He[*1]
Chenting Wang[4,1], Kunchang Li[5,1], Xinhao Li[6,1], Xiangyu Zeng[6,1]
Zilei Wang[3], Yali Wang[5,1],Yu Qiao[1], Limin Wang[6,1], Yi Wang[†1,7]

[1]Shanghai AI Laboratory [2]Zhejiang University [3]University of Science and Technology of China [4]Shanghai Jiao Tong University
[5]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences [6]Nanjing University [7]Shanghai Innovation Institute

https://github.com/OpenGVLab/TPO

Abstract

Current multimodal large language models (MLLMs) struggle with fine-grained or precise understanding of visuals although they give comprehensive perception and reasoning in a spectrum of vision applications. Recent studies either develop tool-using or unify specific visual tasks into the autoregressive framework, often at the expense of overall multimodal performance. To address this issue and enhance MLLMs with visual tasks in a scalable fashion, we propose Task Preference Optimization (TPO), a novel method that utilizes differentiable task preferences derived from typical fine-grained visual tasks. TPO introduces learnable task tokens that establish connections between multiple task-specific heads and the MLLM. By leveraging rich visual labels during training, TPO significantly enhances the MLLM's multimodal capabilities and task-specific performance. Through multi-task co-training within TPO, we observe synergistic benefits that elevate individual task performance beyond what is achievable through single-task training methodologies. Our instantiation of this approach with VideoChat and LLaVA demonstrates an overall 14.6% improvement in multimodal performance compared to baseline models. Additionally, MLLM-TPO demonstrates robust zero-shot capabilities across various tasks, performing comparably to state-of-the-art supervised models.

1. Introduction

Multimodal large language models (MLLMs) show impressive visual perception and reasoning capabilities, with applications in personal assistants [68], embodied systems [17], and scientific discovery [6]. Considering the

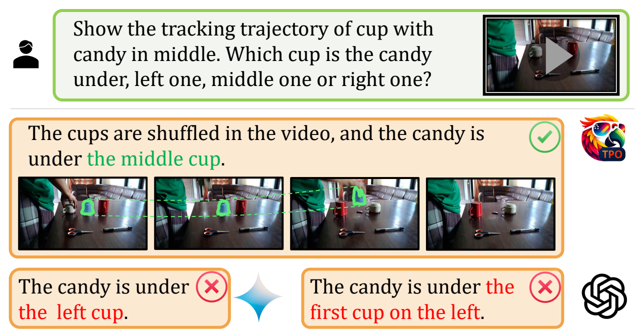

Figure 1. TPO uses differentiable task preferences from dense visual supervisions via task-specific heads to enhance MLLMs in fine-grained understanding.

growing expectations of users in more accurate and detailed perception, taking the shell game as an example, further improving the generality of MLLM requires fine-grained knowledge representation beyond words, *e.g*. MLLM implicitly embeds how to track keys indicated by users.

Existing studies address the enhancement of perceptual granularity in MLLMs by focusing on specific visual tasks (*e.g*. temporal grounding, segmentation, tracking) via MLLMs. They usually fine-tune MLLMs on more task data in text format or enable MLLMs to activate the corresponding task heads. Shikra [7] applies MLLM to localization tasks, transforms object coordinates into dialogue formats, and learns them autoregressively. Meanwhile, TimeChat [63] and VideoChat-T [96] treat event timestamps as text for autoregressive prediction, thereby endowing MLLMs with temporal grounding. They do improve specific task performance significantly, while at the cost of multimodal performance more or less. This is counterintuitive as seminal research proves different visual tasks are correlated and training them together often yields mutual gains [29, 47, 66]. We conjecture that the presentation of different tasks influences this training and that the conflict

*Equal contribution. †Corresponding author.

arises from the learning discrepancies between discrete textual tokens and visually dense predictions. Our experiments in Section 4 validate that a decoupled representation design can effectively address this issue.

To enhance the multimodal capabilities of MLLMs, we explore optimization methods to meet multiple visual task requirements in an end-to-end manner. We propose task preference optimization (TPO), incorporating visual task knowledge into the MLLMs by jointly maximizing the likelihoods of visual task estimations and multimodal dialogue. Inspired by direct preference optimization [60] (DPO) and related methods, we treat visual task annotations as human preferences in particular demands, as shown in Figure 2. DPO aids LLM (or MLLM) in aligning with human preferences through a binary classification that directs the model to generate responses that people prefer. Similarly, TPO enhances MLLMs' visual sensing capabilities through differentiable task optimizations that guide MLLMs to yield dense predictions closely resembling human perception. To achieve this, TPO concretizes MLLM-specific visual perceptions into corresponding task tokens, disentangled from MLLM representation. Then TPO fine-tunes these task tokens and updates the MLLM accordingly.

Specifically, TPO appends visual task heads to the partial output of the MLLM, using several learnable task tokens as inputs for these corresponding heads. During training, TPO enables the MLLM to first distinguish and activate the appropriate task tokens based on user instructions. Subsequently, TPO jointly trains the task tokens and their corresponding heads to enhance the MLLM's understanding of visual tasks. Finally, TPO trains the entire model—including task tokens and heads—on both multimodal and visual-specific task data, promoting the perception and reasoning capabilities of the MLLM. Additionally, we note that multi-task co-training yields greater improvements than single-task training.

TPO demonstrates scalability across various MLLM approaches, encompassing a wide range of visual task categories and data quantities. We validate the effectiveness of TPO within widely used MLLMs, such as LLaVA [48] and VideoChat [43], as detailed in Section 4.3. By fine-tuning these open-source MLLMs with TPO, we enhance their visual understanding capabilities and improve dialogue performance. Additionally, we explore several key spatiotemporal and spatial perception tasks, including spatial grounding, tracking, and temporal grounding. Our findings indicate that these tasks can mutually enhance each other's performance, particularly contributing to the improvement of multimodal dialogue capabilities.

Our contributions can be summarized as:

- We propose a new training method for multimodal large language models, referred to as Task Preference Optimization (TPO). This method leverages supervised information from visual task-specific data to optimize the MLLM through the corresponding heads, resulting in significantly enhanced multimodal perception and reasoning performance. Specifically, TPO achieves an average improvement of 14.6% across multiple image and video multimodal benchmarks [19, 38, 45, 54, 104].
- TPO effectively equips MLLM with the capability to address several key visual perception tasks through the introduced task heads. MLLM-TPO achieves comparable performance in spatial grounding, moment retrieval, highlight detection, referring segmentation, and tracking comparable to expert models across various benchmarks.
- TPO demonstrates scalability across various employed MLLMs, task heads, and scales of task data. We validate the effectiveness of TPO in multiple mainstream MLLMs, such as VideoChat2 [45] and LLaVA [39, 48]. Notably, multi-task joint training based on TPO enhances both multimodal performance and individual visual task, with improvements becoming increasingly significant as additional appropriate heads are introduced. Furthermore, the performance of the MLLM and task heads improves when scaling task data.

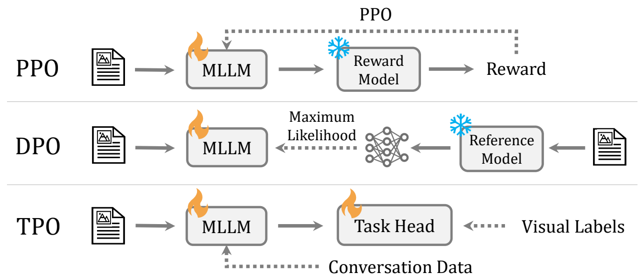

Figure 2. **Comparison of Learning Method.** A solid line indicates data flow, and a dotted line represents feedback. ❄ and 🔥 denote modules that are frozen and unfrozen.

2. Related Work

Vision Foundation Model. Vision foundation models [44, 66, 69, 71, 72, 74, 78, 84, 87, 103] are designed to be adaptable for various downstream tasks through extensive pre-training on large-scale and diverse datasets. VideoPrism [103] achieves the leading results in various video tasks by combining video-text contrastive learning and spatiotemporal token reconstruction using a dataset of public and proprietary videos. InternVideo2 [78] utilizes masked reconstruction, cross-modal contrastive learning, and next-token prediction to enhance the model's perceptiveness, semantic understanding, and reasoning capabilities.

Based on vision foundation models, some studies [41, 51, 77, 106, 107] intend to incorporate downstream task heads into this framework and expect end-to-end task training. Unified-IO [51] introduces a versatile model capable of handling a wide range of tasks across vision and lan-

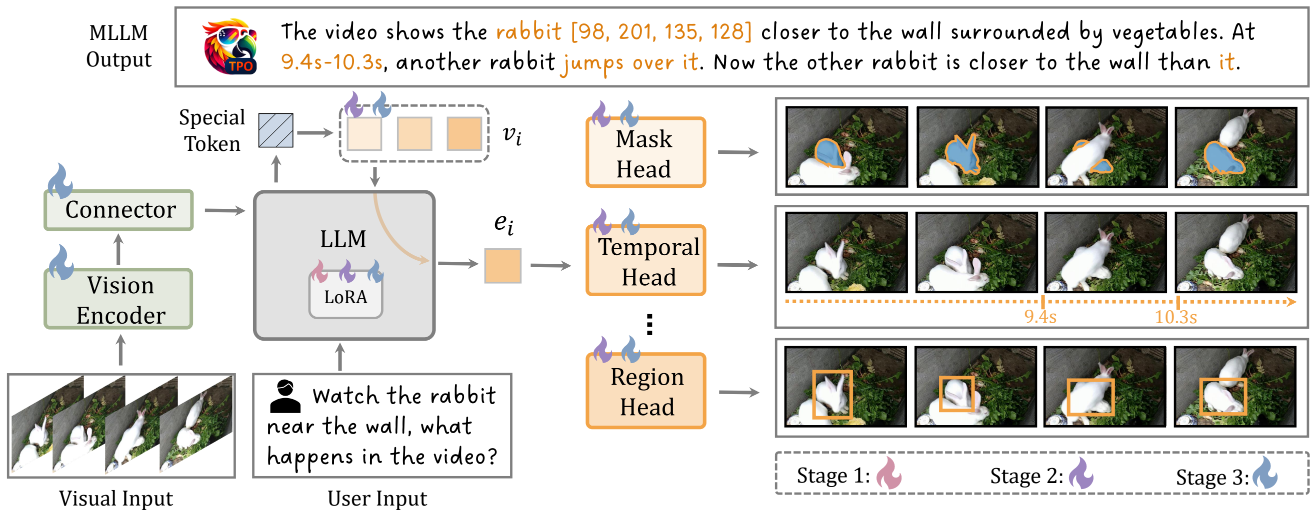

Figure 3. **Overall Pipeline of TPO.** The architecture of Task Preference Optimization (TPO) consists of four main components: (1) a vision encoder, (2) a connector, (3) a large language model, and (4) a series of visual task heads. Differently colored flame symbols indicate which components are unfrozen at various stages of the training process.

guage domains, requiring minimal or no task-specific modifications or additional parameters. Uni-Perceivers [41, 106, 107] formulate different tasks to find the maximum likelihood target for each input through the representation similarity regardless of their modality. Nevertheless, these generalist models are limited to predefined tasks and cannot support flexible open-ended task customization based on language instructions.

Multimodal Large Language Model. The effective understanding and reasoning of LMMs has attracted the attention of numerous researchers. Limited to its input modality, researchers expand the visual capabilities of LLM, leading to MLLMs. Seminal works, such as BLIP-2 [42], LLaVA [48], and mPLUG-Owl [89], introduce image captioning and visual question answering based on LLM using visual instruction-tuning data. Some video-based MLLMs have been proposed, such as VideoChat [43], VideoChat-GPT [52], and Video-LLaMA [99], enabling LLM to gain video understanding capabilities by encoding multiple video frames and using video instruction data.

Vanilla MLLMs have achieved impressive results in visual-related dialogues, but barely address some fine-grained tasks, such as segmentation and temporal grounding. To address this challenge, MLLMs usually take one of the following pipelines: pixel-to-sequence (P2S) and pixel-to-embedding (P2E). For P2S methods [7, 63, 70, 79, 90], MLLM directly outputs the textual predictions. TimeChat [63] introduces a time-aware frame encoder and a sliding video Q-former to enhance temporal perception. For P2E methods [3, 35, 75, 82, 97], MLLM compresses and inputs visuals into the downstream decoder for estimations. LISA [35], NExT-Chat [97], and VideoLISA [3] introduce SAM [33] as a segmentation tool, using a special token as a prompt to connect MLLM and SAM. VisionLLM v2 [82] designs routing tokens and super-link queries to bridge MLLM with multiple decoders.

Alignment in MLLM. Aligning MLLM with human preferences or values is crucial for MLLM's development. Recent works [67, 94, 105] introduce alignment approaches to MLLM, including proximal policy optimization (PPO) [64] and direct preference optimization (DPO) [60], as shown in Figure 2. They usually exploit proprietary models like GPT4-V to build visual preference datasets and then tune MLLMs using PPO or DPO. Llava-RLHF [67] incorporates PPO into a MLLM framework Llava, argumented by image captions or question-answers (QA), while Zhou et al. [105] and Zhang et al. [101] give DPO implementations for MLLMs where visual preference data are created by GPT-4V and other open-sourced MLLMs. RLHF-V [94] collects a dense human preference in segmentation and enhances MLLMs with DPO.

3. Method

Task Preference Optimization (TPO) aims to enable MLLMs to master classical visual perceptions (such as tracking, temporal grounding, etc.) for better task generalization, regarding that many multimodal reasoning tasks require precise visual cues for accurate and reliable responses. As shown in Figure 3, MLLM-TPO has a typical multimodal model M (consisting of a vision encoder E, a vision-language connector C, and a large language model G) and a task preference model P with a series of visual task heads $\{H_i\}_{i=1...n}$ (n denotes the task head number). These heads connect with MLLM using the embeddings $\{\mathbf{e}_i\}_{i=1...n}$ ($\mathbf{e}_i = G(\mathbf{v}_i) \in \mathbb{R}^{1\times\mathrm{C}}$) transformed from the learnable task-specific tokens $\{\mathbf{v}_i\}_{i=1...n}$ ($\mathbf{v}_i \in \mathbb{R}^{1\times\mathrm{C}}$) via MLLM.

TPO employs a local-to-global training scheme, first adapting task heads to the MLLM and then training them jointly. Specifically, the MLLM M starts to recognize $\{\mathbf{v}_i\}_{i=1...n}$ by updating G according to user instructions, then we tune $\{\mathbf{v}_i\}_{i=1...n}$ and $\{H_i\}_{i=1...n}$ for adapting visual heads to M. Finally, we train both M and $\{H_i\}_{i=1...n}$ together. During inference, MLLM-TPO can respond to users' queries in text, and produce structured visual outputs (like masks, timestamps, trajectories, etc.) when users ask (*e.g.* yielding time results for "find when the birthday party starts"). We detail how to form MLLM-TPO structurally and train it in the following.

3.1. Task Preference Models

The task preference model (TPM) P contains a series of task tokens $\{\mathbf{v}_i\}_{i=1...n}$ and heads $\{H_i\}_{i=1...n}$. Before TPM works, the attached MLLM generates special tokens indicating task types from the input queries. Then TPM dynamically calls the task token $\mathbf{v}_j$ based on the special token, transforms it to the task embedding $\mathbf{e}_j$ via the LLM G (the last hidden embedding), and feeds $\mathbf{e}_j$ to the corresponding visual task head $\{H_j\}$ for specific task predictions. Considering existing MLLMs demonstrate remarkable capabilities in common object and scene recognition, yet struggle to accurately locate things or actions, the employed task heads mainly focus on spatiotemporal localization and tracking. Specifically, we give three fundamental task types for compensating mainstream MLLMs' gaps from expert models in visual perceptions: 1) region head, 2) temporal head, and 3) mask head. Their architectures are given below.

- **Region Head.** A two-layer multilayer perceptron (MLP) with ReLU activations is employed for the region head. It takes embeddings from LLM and regresses them to the bounding box coordinates for spatial grounding.
- **Temporal Head.** This head is composed of a video encoder, text encoder, and temporal grounding model for moment retrieval and highlight detection. The video and text queries are input into the video and text encoders to obtain their features respectively. Then we append the temporal task embedding after the text features. With the new text and video features, the MLLM estimates the start/end time and highlight score of the corresponding moment from the query via the temporal grounding model.
- **Mask Head.** Pixel-level tasks pose significant challenges for MLLMs due to the lack of the corresponding learning data and optimization in MLLMs. To this end, we introduce a specialized mask head, utilizing the image encoder, mask decoder, and memory bank components from SAM2 [61], replacing the prompt encoder with a single MLP layer called the mask adapter. Specifically, for the given visuals and user queries, the features extracted from the aforementioned image encoder and the text representation corresponding to the mask task embeddings from LLM are employed, fed into the mask decoder to produce the desired mask.

Most known discriminative vision tasks can be addressed by one or a combination of these three task heads. TPM builds the architectural foundations for leveraging existing vision data with annotations to enhance MLLMs.

3.2. Task Preference Optimization

TPO improves MLLMs with extra supervisions from visual task heads by back-propagating gradients from heads to update MLLMs using visual task data. It enables the language model G in MLLM to discriminate specific task types when users demand (task assignment). Then, TPO trains TPM P via compact task representations $\mathbf{e}_i$ (task optimization) from $\mathbf{v}_i$. Lastly, we train M and P together, tuning M for the refined spatiotemporal perception according to task preferences from P. Its optimization objective is given as:

$$\mathcal{L} = \mathcal{L}_{\text{mllm}} + \underbrace{\mathcal{L}_{\text{assign}}(G(\mathbf{T}_{\text{q}}), \mathbf{s})}_{\text{Task Assignment}} + \sum_{i=1}^{n} \underbrace{\mathcal{L}_{\text{task}}(\mathbf{A}_i, H_i(G(\mathbf{v}_i)))}_{\text{Task Optimization}}, \tag{1}$$

where $\mathbf{T}_{\text{q}}$ and $\mathbf{s}$ denote a user query (some contain specific task needs like tracking) in the form of text token sequence and the ground truth task indicating token, respectively. $\mathbf{A}_i$ is the task annotation of the input visual $\mathbf{X}$, so it could be a number tuple for describing regional localization and area or mask for delineating object shape and position. Here $(\mathbf{X}, \mathbf{T}_{\text{q}}, \mathbf{T}_{\text{a}}, \mathbf{A}_i)$ stands for an input data tuple to MLLM-TPO. Usually, we have $(\mathbf{X}, \mathbf{T}_{\text{q}}, \mathbf{T}_{\text{a}})$ for typical MLLM training while $(\mathbf{X}, \mathbf{T}_{\text{q}}, \mathbf{A}_i)$ for task token and head training. $\mathcal{L}_{\text{assign}}$ is the cross-entropy loss. Meanwhile $\mathcal{L}_{\text{task}_i}$ varies according to the task, and is usually regression- or classification-related loss.

To train MLLM-TPO, we propose a 3-stage local-to-global training scheme, given in Figure 3. Stage 1 learns to identify the task type based on user queries. In stage 2, we train task heads along with their corresponding task tokens, respectively. Lastly, we co-train task heads with MLLM by both task data and multimodal conversation data. Our 3-stage training strategy mitigates the risk of degrading the MLLM's general abilities. We describe them as:

Stage 1: Task Assignment. We design a variety of dialogue templates for different visual tasks to perform task recognition instruction tuning for LLM. The LLM is trained using LoRA [25] in this stage.

Stage 2: Vision Task Training. We integrate task tokens and task-specific heads into the model. By training on task-specific data, the model equips the capacity with fine-grained perception and aligns between the task head and the MLLM. The region head, temporal head, mask adapter, and corresponding task tokens are trained respectively. Similarly, the LLM is updated with LoRA.

Stage 3: Multi-task Training. We tune the entire model on the mixed corpus combining multimodal conversations and specific task data. The synergistic training allows gradients from the heads to MLLM, supervised by human annotations in vision tasks. The vision encoder, connector, task tokens, region head, temporal head, mask adapter, and LLM (with LoRA) are joint-trained in this stage.

4. Experiment

We give the implementation and training & testing specifics of our TPO, and then show its results and ablations.

Implementation Details. We employ VideoChat2 [45], a video-based MLLM, as the primary framework in experiments. For its configuration, we employ UMT-L [44] as the vision encoder, Q-former in $\text{BERT}_{\text{base}}$ [32] as the connector, and Mistral-7B [28] as the language model (LLM).

Regarding the three task heads, the Region Head is initialized randomly. The Temporal Head utilizes CG-DETR [55], with parameters also initialized randomly. The video features input to the temporal head are extracted from the pre-trained InternVideo2 [78], while query features are extracted using the LLM [12]. The Mask Head employs SAM2 [61] and is initialized with its pre-trained weights. Since SAM2 is originally designed solely for tracking, we incorporate a two-layer MLP to encode positioning prompts for spatiotemporal grounding. Additionally, to enable the MLLM to handle the spatial locations provided by the user, we utilize another two-layer MLP to encode the spatial input into the MLLM.

Our model is trained on a variety of visual task datasets and conversation datasets, as shown in Table 1. More training details are provided in the Appendix.

Task	Samples	Datasets
Segmentation	114.6K	MeViS [15], SAMv2 [61]
Temporal Grounding	116.5K	ActivityNet [5], TACoS [62], QVHighlight [36], DiDeMo [24], QuerYD [57], HiREST [95], NLQ [22]
Spatial Grounding	540.0K	Allseeing-V2 [76], Visual Genome [34], RefCOCO [91], RefCOCO+ [91], RefCOCOg [53]
Conversation	3M	YouCook2 [14], ActivityNet [5], VideoChat2-IT [45], ShareGPT-4o [11], LLaVA-Hound-DPO [100], ShareGPT4V [8]

Table 1. **Overview of Datasets Used in TPO for Various Tasks.**

Benchmarks. We evaluate TPO on both general image/video understanding benchmarks (mainly in multiple choice form) and specific visual tasks (e.g. grounding, tracking, and so on). Specifically, we run our model along with other approaches on MVBench, VideoMME, NExT-GQA, MLVU, MMIU, and SEED-Bench2. **MVBench** [45] is designed to evaluate multi-modal fine-grained video understanding tasks for clips (lasting around 8 to 16 seconds), consisting of 20 video tasks relying heavily on temporal perception that can hardly be addressed by single-frame analysis. **Video-MME** [19] is for evaluating MLLMs in both perception and reasoning across varying lengths of videos, annotated by humans. **NExT-GQA** [83] builds on NExT-QA by introducing timestamps that are crucial for understanding questions and determining answers. This requires MLLM to perform multi-step reasoning and emphasize a deeper understanding of both visual and textual content. **MLVU** [104] is constructed from a wide variety of long videos, with lengths ranging from 3 minutes to 2 hours, and includes nine distinct evaluation tasks. **MMIU** [54] is a comprehensive evaluation suite designed to assess MLLMs across a wide range of multi-image tasks. It encompasses 7 types of multi-image relationships, 52 tasks, 77K images, and 11K meticulously curated multiple-choice questions. **SEED2-Bench2** [38] is a MLLM benchmark, featuring 24K multiple-choice questions with human annotations. It spans 27 evaluation dimensions, assessing both text and image generation.

Concerning visual tasks, we test spatial grounding, moment retrieval, highlight detection, tracking, and referring segmentation, including 7 related benchmarks [15, 18, 21, 27, 36, 65, 91] , and the results from several corresponding state-of-the-art expert models.

4.1. General Understanding Evaluation

Multimodal Video Understanding. TPO improves its baseline (VideoChat2) on several video understanding benchmarks with a notable margin. As shown in Table 2, VideoChat-TPO, using a 7B LLM and only 16 input frames, achieves a 66.8 average score on MVBench [45], increasing by 6.4% over VideoChat2 and exceeding the performance of ST-LLM [49] which uses 64 frames. Considering MVBench focuses on characterizing subtle temporal variations involving fine-grained action types, action order, moving direction, and so on, TPO's seamless integration of detailed video understanding (like spatiotemporal grounding and tracking) into its optimized MLLM makes it effectively handle these fine-grained tasks.

On VideoMME [19], VideoChat-TPO outperforms the compared models, achieving a 9.3% improvement over VideoChat2 and demonstrating significant gains across short and medium-length videos, regardless of subtitle availability. On MLVU [104], VideoChat-TPO achieves a M-AVG score of 54.7, surpassing VideoChat2 by 10.2%. This confirms the effectiveness of TPO's enhanced percep-

Model	LLM Params	Frames	MVBench [45]	VideoMME [19]								MLVU [104]
				Overall		Short		Medium		Long		
			AVG	w/o s.	w/ s.	w/o s.	w/ s.	w/o s.	w/ s.	w/o s.	w/ s.	M-AVG
TimeChat [63]	7B	96	38.5	34.3	36.9	39.1	43.1	31.8	33.9	32.1	33.6	30.9
Video-LLAVA [46]	7B	8	43.0	41.1	41.9	46.9	47.3	38.7	40.4	37.8	37.9	47.3
ShareGPT4Video [9]	7B	16	51.2	39.9	43.6	48.3	53.6	36.3	39.3	35.0	37.9	46.4
LLaVA-Next-Video [102]	7B	16	44.0	38.0	40.8	44.6	47.4	37.7	39.4	31.9	35.6	39.3
ST-LLM [49]	7B	64	54.9	37.9	42.3	45.7	48.4	36.8	41.4	31.3	36.9	-
PLLaVA-34B [85]	34B	16	58.1	40.0	35.0	47.2	36.2	38.2	35.9	34.7	32.9	53.6
Chat-UniVi [30]	7B	64	40.8	40.6	45.9	45.7	51.2	41.3	47.3	39.1	43.4	-
VideoChat2(baseline) [45]	7B	16	60.4	39.5	43.8	48.3	52.8	37.0	39.4	33.2	39.2	44.5
VideoChat-TPO	7B	16	**66.8** (+6.4)	**48.8** (+9.3)	**53.8** (+10.0)	**58.8**	**64.9**	**46.7**	**50.0**	**41.0**	**46.4**	**54.7** (+10.2)

Table 2. **Performance on Multimodal Video Understanding.** We compare our model to others using LLMs of the same generation or 16-frame input. w/o s. indicates without subtitle, while w s. indicates with subtitle. M-AVG refers to the mean average of MLVU.

tions of long-form video evaluations. These results across three benchmarks demonstrate the notable advancements in multimodal video understanding achieved through TPO.

Grounded Video QA. Table 3 shows that VideoChat-TPO outperforms other models, achieving superior accuracy (Acc) and intersection over union (IoU) scores in NExT-GQA [83]. Its intersection over prediction (IoP) scores are comparable to those of LLoVi [98], which employs large, closed-source commercial models like GPT-4 [1]. The high Acc@IoP score not only reflects the enhanced capability of TPO-optimized VideoChat in effectively understanding and interpreting video content, but also demonstrates its robustness in handling complex reasoning scenarios. Furthermore, the higher Acc@GQA score indicates that the model successfully integrates fine-grained temporal grounding with complex reasoning tasks, enabling it to accurately provide temporal clues necessary for inferring answers. From this perspective, TPO training framework significantly allows MLLMs to overcome their limitations in identifying and localizing temporal elements within video data. This enhancement positions MLLM-TPO with a competitive edge in temporal-related reasoning.

Multimodal Multi-image Understanding. To explore TPO's effect on multi-image understanding, we test it on the MMIU and SEED-Bench2, as shown in Table 4. On MMIU, VideoChat-TPO achieved an overall score of 40.2, a 5.2% improvement over VideoChat2. Besides temporal changes, MMIU evaluates models' spatial sensing and semantic relations in scenes. On SEED-Bench2, across 27 performance indicators, VideoChat-TPO achieves 41.7% improvement on average performance compared to VideoChat2. VideoChat-TPO's superior performance on multi-image understanding compared to LLaVA-Interleave and InternLM-XComposer demonstrates that TPO's vision enhancements improve the MLLM's spatial perception and image understanding.

4.2. Vision Task Evaluation

Moment Retrieval. Moment Retrieval is to locate the target segments in a video based on queries. Table 5 and 6 compare the zero-shot and fine-tuned moment retrieval performance of VideoChat-TPO against other expert models and MLLMs. In zero-shot settings, VideoChat-TPO achieves a R@1 (IoU=0.5) score of 40.2 on the Charades-STA [21]. It surpasses the previous state-of-the-art (SOTA) MLLM specialized for temporal tasks, ChatVTG [59], and the expert model, UniVTG [47]. This demonstrates VideoChat-TPO's ability to accurately locate video moments corresponding to given text queries, thereby enhancing the practical applicability of MLLMs.

Highlight Detection. Highlight detection generates salient scores for emphasizing frames based on the input language query. We compare the fine-tuning highlight detection performance of VideoChat-TPO with that of other expert models and MLLMs, as shown in Table 6. VideoChat-TPO notably outperforms the previous state-of-the-art MLLM (such as TimeChat [63]) on both Charades-STA and QVHighlight. For expert models like QD-DETR and UniVTG, VideoChat-TPO beats them on Charades-STA non-trivially and achieves a comparable performance on QVHighlight. This also demonstrates the progress in extending MLLMs to broad temporal tasks.

Spatial Grounding. To verify the fine-grained localization ability of the model, we run the spatial grounding task which inputs textual descriptions into the model and gets the corresponding bounding box on RefCOCO [91]. As shown in Table 7, we compare VideoChat-TPO with pixel-to-sequence models, *i.e.*, VisionLLM-H [75], pixel-to-embedding methods, *i.e.*, NExT-Chat [97], and expert models, *i.e.*, G-DINO [50]. VideoChat-TPO performs better than the pixel-to-embedding methods by using only a simple task head and achieves comparable performance to pixel-to-sequence models fine-tuned on a large amount of spatial grounding data, as well as to specialized models.

Model	Acc@IoP	Acc@GQA	mIoP	IoP@0.3	IoP@0.5	mIoU	IoU@0.3	IoU@0.5
VIOLETv2 [20]	54.9	12.8	23.6	25.1	23.3	3.1	4.3	1.3
SeViLA [93]	72.5	16.6	29.5	34.7	22.9	21.7	29.2	13.8
LangRepo [31]	59.6	17.1	31.3	-	28.7	18.5	-	12.2
FrozenBiLM NG+ [88]	73.8	17.5	24.2	28.5	23.7	9.6	13.5	6.1
VideoStreaming [58]	57.4	17.8	32.2	-	31.0	19.3	-	13.3
LLoVi [98]	65.9	24.3	**37.3**	-	**36.9**	20.0	-	15.3
HawkEye [79]	-	-	-	-	-	25.7	37.0	19.5
VideoChat-TPO	**77.7**	**25.5**	35.6	**47.5**	32.8	**27.7**	**41.2**	**23.4**

Table 3. **Performance on Grounded QA.**

Model	MMIU [54]	SEED2$_I$ [38]	SEED2$_M$ [38]
LLaVA-v1.5 [48]	19.2	58.3	39.2
ShareGPT4V [8]	18.5	-	-
OpenFlamingo [2]	22.3	36.6	43.5
LLaVA-Interleave [40]	32.4	-	-
VideoChat2 [45]	35.0	26.5	27.6
VideoChatGPT [52]	-	38.3	49.8
InternLM-XComposer [16]	21.9	65.4	49.8
VideoChat-TPO	**40.2** (+5.2)	**67.3** (+40.8)	**70.0** (+42.4)

Table 4. **Performance on Image Understanding.**

Model	Charades-STA [21]			
	R@0.3	R@0.5	R@0.7	mIoU
UniVTG [47]	44.1	25.2	10.0	27.1
VideoChat2 [45]	38.0	14.3	3.8	24.6
VTimeLLM [26]	51.0	27.5	11.4	31.2
TimeChat [63]	-	32.2	13.4	-
HawkEye [79]	50.6	31.4	14.5	33.7
ChatVTG [59]	52.7	33.0	15.9	34.9
VideoChat-TPO	**58.3**	**40.2**	**18.4**	**38.1**

Table 5. **Zero-Shot Performance on Moment Retrieval.** Gray means no LLM.

Model	Charades-STA [21]				QVHighlight [36]	
	R@0.3	R@0.5	R@0.7	mIoU	mAP	HIT@1
M-DETR [36]	65.8	52.1	30.6	45.5	35.7	55.6
QD-DETR [56]	-	57.3	32.6	-	38.9	62.4
UniVTG [47]	72.6	60.2	38.6	52.2	**40.5**	**66.3**
TimeChat [63]	-	46.7	23.7	-	21.7	37.9
HawkEye [79]	72.5	58.3	28.8	-	-	-
VideoChat-TPO	**77.0**	**65.0**	**40.7**	**55.0**	<u>38.6</u>	<u>65.4</u>

Table 6. **Fine-tuning Performance on Moment Retrieval and Highlight Detection.** Gray means no LLM.

Methods	RefCOCO [91]		
	val	testA	testB
MAttNet ★ [92]	76.4	80.4	69.3
OFA-L [73]	80.0	83.7	76.4
G-DINO-L ★ [50]	**90.6**	**93.2**	**88.2**
VisionLLM-H [75]	-	86.7	-
Shikra-7B [7]	**87.0**	90.6	80.2
NExT-Chat-7B [97]	85.5	90.0	77.9
VideoChat-TPO	85.9	**90.8**	**81.3**

Table 7. **Spatial Grounding Task.** ★ with a refined decoder.

Tracking. In the tracking task, the model receives the object coordinates from the first frame and outputs the coordinates for the remaining frames in the video. We evaluate the VideoChat-TPO on the mainstream tracking benchmarks LAOST [18] and GOT-10k [27], as shown in Table 8. In zero-shot testing, VideoChat-TPO ranks best among all MLLMs, even surpassing some fine-tuned expert models notably with around 10% increase on success rate, such as SiamFC [4] and ATOM [13]. With the TPO training method, the model is optimized from mask sequences that hard to represent by words, allowing the MLLM to achieve strong motion characterization for multimodal generalization in highly dynamic and occluded scenes.

Referring Segmentation. The referring segmentation task requires the model to output the mask sequence corresponding to the specified prompt. This capability is newly activated in VideoChat-TPO, as it is not supported or designed in the original VideoChat2 or SAM2. We compare the referring segmentation performance of VideoChat-TPO with other expert models and MLLMs in Table 9. The zero shot capability of VideoChat-TPO is comparable to the fine-tuning capabilities of other expert models, *i.e*. ReferFormer [81], showing its notable open-world video segmentation performance. Through TPO training, MLLM effectively optimizes its understanding in both tracking and segmenting objects indicated by users. It further enables the model to excel in pixel-level tasks, offering perceptual advantages to it in practical applications like robotic control.

4.3. Ablation Studies

In this section, we analyze the effectiveness of the key components of TPO and evaluate its scalability.

Extending to Other MLLMs. To evaluate the effectiveness of the TPO method on different MLLMs, we apply TPO to LLaVA (LLaVA-OneVision [39]) in addition to VideoChat. In LLaVA, fine-grained video understanding is a weak spot. Despite the conversation data used in LLaVA-OneVision being very similar to, or even more extensive than, what we employ, there remains significant room for improvement in its performance on short-term fine-grained understanding tasks. As shown in Table 10, with TPO, LLaVA not only gains capabilities that it originally lacked but also achieves an 8.1% improvement on MVBench, demonstrating the universality of the TPO method.

Task Preference Model vs. Textualized Task Data. We compare the TPO method with the approach of textualizing task data and training it in an autoregressive manner, which is a straightforward and efficient method for MLLMs to learn specific tasks. We use the same data as the TPO method and convert the data related to the task head into conversation format. Due to the limitations of textual representation, much data must be input in a more discrete manner. For instance, in tracking supervision data, we follow Merlin [90] by converting masks into a sequence of spatial coordinates to serve as both the model's input and output. As shown in Table 11, MLLM-TPO achieves a 2.1% performance improvement on MVBench [45] compared to MLLMs trained with textualized task data. This demonstrates that the task head in the TPO method enables more effective utilization of the original supervision signals compared to the next-token prediction approach, which inevitably incurs information loss when converting fine-grained tasks into textual annotations. The next-token prediction method cannot truly capture the nuanced information present in videos, while TPO allows the MLLM to

Model	LaSOT [18]			GOT-10k [27]		
	Success	P_{norm}	P	Overlap	SR0.5	SR0.75
SiamFC [4]	33.6	42.0	33.9	34.8	35.3	9.8
ATOM [13]	51.5	-	-	55.6	63.4	40.2
SiamRPN++ [37]	49.6	56.9	49.1	51.8	61.8	32.5
SiamFC++ [86]	54.4	62.3	54.7	59.5	69.5	47.9
LLaVA-1.5 [48]	19.4	16.5	12.8	23.5	20.2	9.7
Merlin [90]	39.8	40.2	38.1	51.4	55.9	42.8
VideoChat-TPO	**69.4**	**80.1**	**76.9**	**70.6**	**79.8**	**66.0**

Table 8. **Performance on Tracking Benchmarks.**

Method	Ref-YouTube-VOS [65]			MeViS [15]		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [81]	62.9	61.3	64.6	31.0	29.8	32.2
OnlineRefer [80]	62.9	61.0	64.7	-	-	-
LISA [35]	52.6	52.1	53.0	-	-	-
VideoLISA [3]	63.7	**61.7**	63.7	44.4	41.3	47.6
VideoChat-TPO	**63.9**	52.3	**75.4**	**47.0**	**42.6**	**51.3**

Table 9. **Performance on Referring Segmentation.** TPO is evaluated in a zero-shot on Ref-YouTube.

Model	MVBench
LLaVA-OV [39]	56.7
LLaVA-OV-TPO	**64.8** (+8.1)

Table 10. **Applying TPO on LLaVA-OV.** TPO significantly enhances the ability to understand fine-grained video details.

Model	Charades-STA [21]				MVBench
	R@0.3	R@0.5	R@0.7	mIoU	AVG
VideoChat-TPO	**58.3**	**40.2**	**18.4**	**38.1**	**66.8**
w/o reasoning data	56.4	38.3	17.1	35.6	66.1
replace by simple head	31.5	17.8	6.1	15.4	65.8
textualized task data	33.7	18.6	6.2	16.0	64.7

Table 11. **Ablation of Reasoning Data and Head Performance.**

Exp.	T	R	M	C	R@0.5	Acc@0.5	$\mathcal{J}\&\mathcal{F}$
1	✓				30.2	-	-
2		✓			-	77.3	-
3			✓		-	-	55.1
4		✓	✓		-	80.2	58.3
5	✓	✓	✓		36.7	81.6	61.4
6	✓	✓	✓	✓	**40.2**	**82.0**	**63.9**

Table 12. **Impact of TPO Components and Data.** T, R, M, and C denote temporal head, region head, mask head, and conversation data respectively. R1@0.5 means R1@0.5 in Charades-STA, Acc@0.5 represents the mean of Acc@0.5 in all COCO datasets, $\mathcal{J}\&\mathcal{F}$ means $\mathcal{J}\&\mathcal{F}$ in Ref-YouTube-VOS.

learn perceptual information beyond conversational data.

Impact of Task Head Performance on TPO. A stronger task head is expected to enhance the multimodal capabilities of TPO more effectively than a weaker one. To further explore the impact of the task head on model performance, we replaced the temporal head from CG-DETR [55] with a simple two-layer MLP. As shown in Table 11, the simple temporal head exhibits a significant performance decline compared to CG-DETR on the corresponding tasks. However, it still achieves a 1.1% improvement compared to the model without TPO. In contrast to the simpler head, the well-designed head can provide more accurate expert knowledge through its architecture and pre-trained task preference weights, enabling better utilization of the data.

Impact of Data Scaling. Concerning foundation or large models, we expect to enhance their model performance by incorporating additional data at scale. In our approach, we expand the existing conversation data by integrating two reasoning datasets [5, 14] with temporal information. By incorporating these datasets, we observe a notable enhancement in the capabilities of the temporal head in Table 11. Furthermore, as shown in Table 12, adding conversation data further improves performance across various tasks. Overall, increasing the amount of conversation and fine-grained task data leads to noticeable improvements in model performance, demonstrating the effectiveness of data scaling in training multimodal models with TPO.

Synergistic Gains from Co-training. Table 12 presents the impact of incorporating different visual tasks into TPO on performance. This indicates that the collaborative learning of visual tasks facilitates the transfer of knowledge, ultimately leading to better performance across the board. Overall, these results confirm that co-training not only benefits individual tasks but also creates a synergistic effect that enhances overall capabilities in visual understanding.

5. Conclusions

This paper introduces Task Preference Optimization (TPO). It enhances the overall multimodal performance of MLLMs by enhancing their precise visual understanding. TPO achieves this by integrating differentiable task preferences derived from fine-grained visual tasks. It introduces learnable task tokens that serve as bridges between multiple task-specific heads and the MLLM. Through the joint optimization of these task tokens, heads, and the MLLM, TPO leads to a substantial improvement in multimodal dialogue capabilities and performance across various visual tasks. Our results demonstrate the effectiveness of TPO in scaling MLLMs with task-specific data and seamlessly integrating them with existing expert vision models. We believe this study clarifies the prerequisites for fusing MLLMs with models and data from the pre-large model era, effectively bridging the gap between expert and generalist models, as well as between generation and understanding.

Limitations. Currently, TPO focuses exclusively on discriminative visual tasks, overlooking generative ones. Meanwhile, the framework is supervised by human annotations, which neglects potentially valuable unsupervised or self-supervised learning approaches such as contrastive learning [10, 23]. This limitation inherently restricts the scalability of TPO in terms of both task diversity and requirements. While we demonstrate TPO's potential for enhancing MLLMs through increased task-specific data, a comprehensive investigation into the broader impact of this scaling remains a crucial area for future research.

Acknowledgements

This work is partially supported by the National Key R&D Program of China (No. 2022ZD0160102), Jiangsu Frontier Technology R&D Program (No. BF2024076), and the Science and Technology Commission of Shanghai Municipality under Grant No. 23YF1461900.

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 preprint arXiv:2303.08774*, 2023. 6

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 7

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *arXiv preprint arXiv:2409.19603*, 2024. 3, 8

[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, pages 850–865, 2016. 7, 8

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 5, 8

[6] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023. 1

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 3, 7

[8] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 5, 7

[9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 6

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 8

[11] 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*, pages 24185–24198, 2024. 5

[12] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023. 5

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019. 7, 8

[14] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, pages 2634–2641, 2013. 5, 8

[15] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 5, 8

[16] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 7

[17] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 1

[18] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, pages 5374–5383, 2019. 5, 7, 8

[19] Chaoyou Fu, Yuhan Dai, Yondong 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 preprint arXiv:2405.21075*, 2024. 2, 5, 6

[20] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *CVPR*, 2023. 7

[21] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 5, 6, 7, 8

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 5

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 8

[24] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 5

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 4

[26] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, 2024. 7

[27] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 43(5):1562–1577, 2019. 5, 7, 8

[28] 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 preprint arXiv:2310.06825*, 2023. 5

[29] Fan Jiang and Zilei Wang. Sparse sharing relation network for panoptic driving perception. In *ACMMM*, pages 800–808, 2023. 1

[30] 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*, pages 13700–13710, 2024. 6

[31] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*, 2024. 7

[32] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 5

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3

[34] 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. *IJCV*, 123:32–73, 2017. 5

[35] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024. 3, 8

[36] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *NeurIPS*, 34:11846–11858, 2021. 5, 7

[37] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, pages 4282–4291, 2019. 8

[38] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, pages 13299–13308, 2024. 2, 5, 7

[39] 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 preprint arXiv:2408.03326*, 2024. 2, 7, 8

[40] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 7

[41] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *CVPR*, pages 2691–2700, 2023. 2, 3

[42] 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*, pages 19730–19742. PMLR, 2023. 3

[43] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2, 3

[44] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, pages 19948–19960, 2023. 2, 5

[45] Kunchang Li, Yali Wang, Yinan 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*, pages 22195–22206, 2024. 2, 5, 6, 7

[46] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6

[47] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, pages 2794–2804, 2023. 1, 6, 7

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2, 3, 7, 8

[49] 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

[50] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 6, 7

[51] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2022. 2

[52] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 3, 7

[53] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 5

[54] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024. 2, 5, 7

[55] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023. 5, 8

[56] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, pages 23023–23033, 2023. 7

[57] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP*, pages 2265–2269. IEEE, 2021. 5

[58] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *arXiv preprint arXiv:2405.16009*, 2024. 7

[59] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *CVPR*, pages 1847–1856, 2024. 6, 7

[60] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36, 2024. 2, 3

[61] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4, 5

[62] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 5

[63] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *CVPR*, abs/2312.02051, 2024. 1, 3, 6, 7

[64] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[65] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223. Springer, 2020. 5, 8

[66] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, et al. Intern: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021. 1, 2

[67] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3

[68] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, pages 10078–10093, 2022. 2

[70] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024. 3

[71] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36. Springer, 2016. 2

[72] 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 *CVPR*, 2023. 2

[73] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022. 7

[74] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, pages 6312–6322, 2023. 2

[75] 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, 6, 7

[76] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024. 5

[77] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *ICCV*, pages 1130–1140, 2023. 2

[78] 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. 2, 5

[79] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024. 3, 7

[80] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*, pages 2761–2770, 2023. 8

[81] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022. 7, 8

[82] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Wenhai Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024. 3

[83] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *CVPR*, 2024. 5, 6

[84] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pages 6787–6800, 2021. 2

[85] 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 preprint arXiv:2404.16994*, 2024. 6

[86] Yinda Xu et al. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, pages 140–148, 2020. 8

[87] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv*, abs/2212.04979, 2022. 2

[88] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 7

[89] 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. 3

[90] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *ECCV*, pages 425–443. Springer, 2025. 3, 7, 8

[91] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 5, 6, 7

[92] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 7

[93] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023. 7

[94] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, pages 13807–13816, 2024. 3

[95] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, pages 23056–23065, 2023. 5

[96] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *ICLR*. 1

[97] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023. 3, 6, 7

[98] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 6, 7

[99] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, pages 543–553, 2023. 3

[100] 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 preprint arXiv:2404.01258*, 2024. 5

[101] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 3

[102] 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. 6

[103] Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. In *ICML*, 2024. 2

[104] 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 preprint arXiv:2406.04264*, 2024. 2, 5, 6

[105] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR Workshop on Reliable and Responsible Foundation Models*, 2024. 3

[106] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with

conditional moes. In *NeurIPS*, pages 2664–2678, 2022. 2, 3

[107] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, pages 16804–16815, 2022. 2, 3