This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Enhancing Visual Planning with Auxiliary Tasks and Multi-token Prediction

Ce Zhang[1*] Yale Song[2] Ruta Desai[2] Michael Louis Iuzzolino[2]
Joseph Tighe[2] Gedas Bertasius[1] Satwik Kottur[2]
[1]UNC Chapel Hill [2]Meta

Abstract

*Visual Planning for Assistance (VPA) aims to predict a sequence of user actions required to achieve a specified goal based on a video showing the user's progress. Although recent advances in multimodal large language models (MLLMs) have shown promising results in video understanding, long-horizon visual planning remains a challenging problem. We identify two challenges in training large MLLMs for video-based planning tasks: (1) scarcity of procedural annotations, limiting the model's ability to learn procedural task dynamics effectively, and (2) inefficiency of next-token prediction objective to explicitly capture the structured action space for visual planning when compared to free-form, natural language. To tackle data scarcity, we introduce Auxiliary Task Augmentation. We design and train our model on auxiliary tasks relevant to long-horizon video-based planning (*e.g.*, goal prediction) to augment the model's planning ability. To more explicitly model the structured action space unique to visual planning tasks, we leverage Multi-token Prediction, extending traditional next-token prediction by using multiple heads to predict multiple future tokens during training. Our approach, VideoPlan, achieves state-of-the-art VPA performance on the COIN and CrossTask datasets, surpassing prior methods by **7.3%** and **3.4%**, respectively, when predicting 3 future actions. We further extend our method to the challenging Ego4D Long-term Action Anticipation task, and show that it is on par with the state-of-the-art approaches despite not using specialized egocentric features. Code is available at* https://github.com/CeeZh/VideoPlan.

1. Introduction

With the rapidly increasing interest in assistive technologies, ranging from personal virtual assistant to a physical robot, the ability to anticipate future actions in a goal-oriented setting is crucial for such embodiments to better aid humans. Aimed at this desired ability, Visual Planning for Assistant (VPA) task [35] focuses on predicting a sequence of future actions (*e.g.*, *'install sofa legs'*, *'put on sofa cover'*) necessary to achieve a specified goal (*e.g.*, *assemble sofa*), based on a video that captures the user's progress. VPA has broad applications, such as helping people learn new skills (*e.g.*, drawing, cooking) or guiding them in unfamiliar household tasks (*e.g.*, assembly). To be successful in this task, a model would need to handle long context videos from untrimmed video, understand complex real-world goals, and generate consistent action plans towards them as illustrated in Fig. 1.

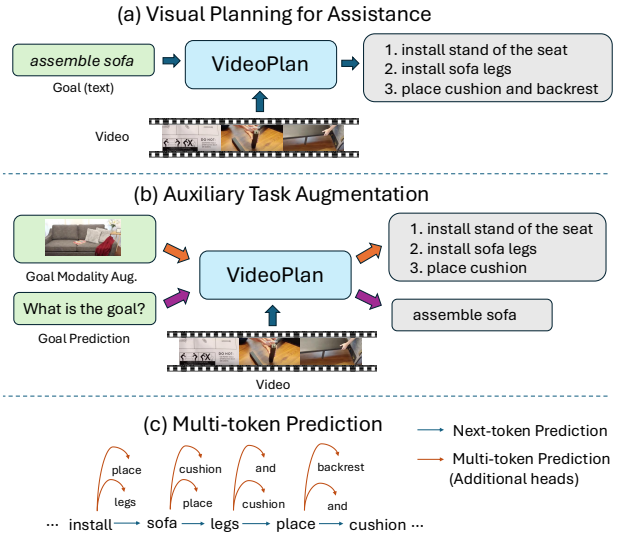


Figure 1. **(a) Visual Planning for Assistance (VPA):** predict a sequence of future actions (grey) given a video observation of user's progress and a succinct goal in text (green). (b) **Auxiliary Task Augmentation:** Construct additional tasks related to long-term visual planning. Inputs (green) and outputs (grey) for a given auxiliary tasks are connected via same colored arrows. (c) **Multi-token Prediction:** Extend next-token prediction by also modeling future tokens (red arrows) via additional heads. VideoPlan leverages (b) and (c) to overcome data scarcity and inefficiency of next-token prediction to explicitly reason about future tokens at current step, to achieve state-of-the-art on VPA.

Recent developments in Multimodal Large Language Models (MLLMs) have made significant progress in video

*Work done during an internship at Meta.

understanding, such as visual question answering [17, 23, 39, 49], temporal grounding [13, 37], and visual captioning [19]. Prior works [11, 24, 38] have shown that large language models (LLMs), which usually form the heart of MLLMs, possess procedural knowledge and can plan well in the text domain. Inspired by this success, we investigate MLLMs as a natural modeling choice for visual planning. However, applying MLLMs to long-horizon planning tasks like VPA uncovers two main challenges. First, training an MLLM-based agent capable of assisting humans with complex, long-horizon daily tasks necessitates a substantial volume of training data, each containing long sequences of procedural annotations. Unlike short video-text pairs, procedural annotations require detailed and step-by-step labeling making data collection expensive, time consuming, and cumbersome; thereby are prohibitively resource intensive. The scarcity of such annotated procedural data limits the model's ability to learn the task-specific dynamics necessary for accurate action prediction. Second, the action space for visual planning exhibits more structure compared to natural, free-form language–both in terms of individual and sequence of labels. For instance, in a cooking scenario, the actions are limited to a set of cooking-related steps, where actions like *'install sofa legs'* will never appear. At the sequence level, there are strong long-term temporal dependencies amongst the constituent actions, *e.g.*, *'open microwave door'* is likely followed by *'take an item out'* or *'put an item in.'* Traditional MLLMs are trained using the standard next-token prediction loss that shows strong generalizability when trained on a large scale of data. However, next-token prediction might not fully capture these strong temporal structured dependencies in visual planning tasks, due to the lack of explicit reasoning about future tokens when predicting the next token. As a consequence of this design, the resultant models miss out on a crucial learning signal for reliable long-horizon planning, the effects of which are more pronounced in data-scarce tasks like visual planning.

In this work, we propose two strategies to address the above related challenges. First, we introduce Auxiliary Task Augmentation, which enhances the model's planning capabilities by training it on auxiliary tasks relevant to long-horizon visual planning. Specifically, we employ two types of auxiliary tasks: (1) Goal Modality Augmentation: We modify the goal modalities, for example, changing it from text to image. (2) Goal Prediction: The model predicts the human's goal based on video or text inputs. Our augmentation strategy generates additional training data, helping to address the scarcity of procedural annotations. This ultimately enables the model to better capture human intentions and task dynamics, which are essential for effective assistance and long-horizon planning.

Second, we employ Multi-Token Prediction (MTP) [9]. Unlike next-token prediction, which focuses on predicting only the next token, MTP introduces multiple additional heads on top of a shared model trunk during training to predict multiple future tokens simultaneously. As a result, the model, when predicting each token, not only reasons about the next token but also explicitly reasons about the future tokens, thus essentially emulating a mild form of 'planning' even at the token level. During inference, the model removes the additional heads and generates the next token autoregressively. In essence, MTP serves as an additional regularizer well-suited for visual planning, without sacrificing the expressibility of a language model to generate open-vocabulary actions, unlike competing approaches that use a closed-vocabulary action classifier on top [5, 35]. Our experiments demonstrate that Multi-token Prediction captures the structured action space of planning tasks more effectively than the standard next token prediction approach, improving the model's ability to handle the structured long-term temporal dependencies.

We conduct extensive evaluations on the COIN [40] and CrossTask [55] datasets for the VPA task. Both Auxiliary Task Augmentation and Multi-token Prediction individually enhance our baseline MLLM model. Combining them, our model achieves state-of-the-art results on both datasets, outperforming previous methods [14, 15, 35] by 7.3% (absolute) and 3.4% (absolute), respectively, on success rate for predicting 3 future steps. It is worth noting that VideoPlan uses a smaller LLM compared with the prior SOTA method [15], which is important for practical applications of MLLMs for VPA in the real world [42]. We further extend our method to the challenging Ego4D Long-term Action Anticipation task [10], which requires predicting 20 future actions without a specified goal. Despite not being pre-trained on egocentric data, our approach achieves comparable results to the state-of-the-art methods.

2. Related Works

Multimodal Large Language Models (MLLMs). Recent advancements in LLMs [2, 41, 53] have sparked interest in MLLMs that leverage the knowledge within LLMs to enhance multimodal perception. Flamingo [3] integrates cross-attention modules into the LLM to handle interleaved multimodal sequences. BLIP-2 [18] utilizes Q-former to encode visual information and align visual features with the LLM's input space, while LLaVA [25] and MiniGPT-4 [54] employ linear layers to connect the visual encoder and the LLM, streamlining the integration process. To extend MLLMs to videos, most existing approaches uniformly sample frames from videos and align them with the LLM [17, 19, 23]. Recent works focus on advanced frame selection or token compression methods [21, 39, 49] to boost performance. Recent works also explore MLLMs for downstream tasks, such as temporal grounding [13, 37], spatial reasoning [6], and planning [30, 46, 52]. For plan-

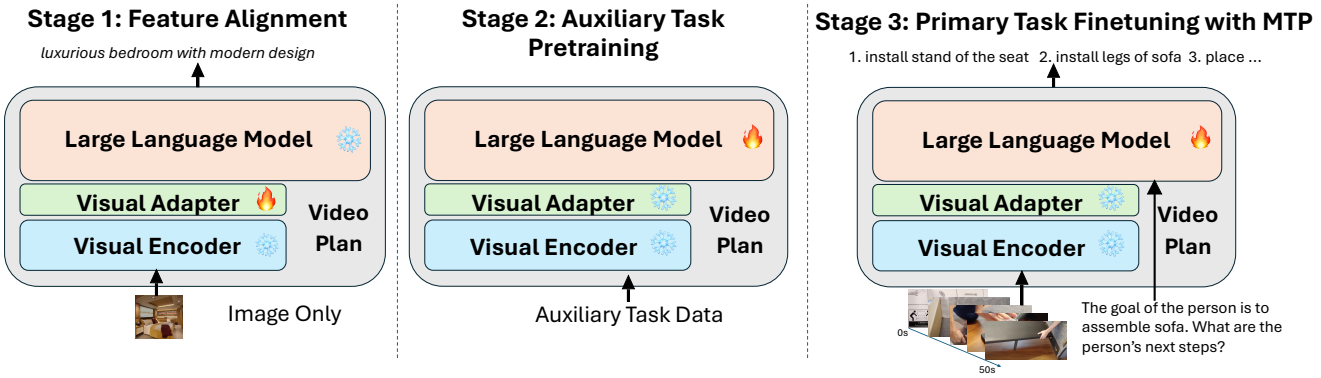


Figure 2. **Our three-stage training pipeline.** Stage 1 aligns the features of the visual encoder with the LLM embedding space by only training a visual adapter. Stage 2 helps the model better learn visual planning dynamics by training on other related auxiliary tasks. Finally, Stage 3 finetunes the model on VPA, the desired task at hand.

ning tasks, these works typically assume the agent can interact with the environment. Instead, EgoPlanBench [7] and VPA [35] introduce benchmarks and models for video-based planning without interaction, focusing on future action prediction. Our work also aims to enhance MLLMs' capabilities for long-horizon visual planning where interaction with the environment is infeasible.

Planning in Instructional Videos. Procedure Planning [5] in instructional videos aims to predict multiple steps to finish the given goal based on the current observation, where both the observation and the goal are images. Prior works employ many techniques to solve procedure planning, including using intermediate states as additional supervision [34, 43, 50], leveraging LLMs or diffusion models for planning [15, 26, 32, 44, 47], or using temporal prior or task prior [22, 32, 43, 50]. However, in practice, the visual state of terminal state (goal) is not available, reducing the usefulness of procedure planning as a real-world application. To address this issue, recent work [35] introduces Visual Planning for Assistance (VPA), which requires the model to output future steps based on a textual goal and a video showing the user's progress. Our work focuses on VPA as it is a more realistic real-world assistance setting.

Long-term Action Anticipation (LTA). This line of work aims to predict multiple future actions directly from the input video [1, 10, 31]. Most prior methods explore learning useful representations for future prediction [4, 27, 33, 48]. Recent works focus more on using LLMs and VLMs for LTA, leveraging the knowledge from large-scale pretraining data [16, 28, 45, 51]. LTA can be viewed as a special case of long-horizon visual planning where the goal is not specified. Therefore, we evaluate our method on LTA to show the generalizability of our proposed approach.

3. Proposed Approach

In this section, we introduce VideoPlan, an MLLM designed for video-based long-horizon planning. VideoPlan is composed of a visual encoder, a visual adapter, and an LLM (Sec. 3.2). To enhance planning capabilities, we adopt two key strategies: Auxiliary Task Augmentation (Sec. 3.3) and Multi-Token Prediction (Sec. 3.4) that effectively tackle the shortcomings in naively training an MLLM for visual planning. Finally, we describe our three stage process used to train VideoPlan in Sec. 3.5.

3.1. Task Formulation

The task of Visual Planning for Assistance (VPA) aims to predict a sequence of actions $\mathcal{A} = \{a_1, a_2, \ldots, a_H\}$ given the user's goal $\mathcal{G}$ and current observation $\mathcal{O}$, where H is the planning horizon [35]. Specifically, the observation $\mathcal{O}$ is presented as an untrimmed video history capturing the user's progress. The goal $\mathcal{G}$ is specified in a short natural language description (*e.g. assemble sofa*). The actions $\mathcal{A}$ are annotated in free-form language (*e.g. 'install sofa legs'*, *'put on cover'*) but form a closed set of action vocabulary.

3.2. Model Architecture

Visual Encoder. Given a video $\mathcal{O}$, we uniformly sample N_V frames represented as $V \in \mathbb{R}^{N_V \times H \times W \times C}$, where H, W, C are height, width, and number of channels, respectively. For each frame, we utilize a pretrained visual feature encoder $f_v(.)$ to extract the representation $v_i = f_v(V)$.

Visual Adapter. Following prior works [25, 29, 35], we use a visual adapter $f_a(.)$ to map frame feature v_i into the input embedding space of the LLM, denoted as $z_i = f_a(v_i)$.

LLM. Given the goal $\mathcal{G} = \{G_i\}_{i=1}^{N_T}$, with N_T number of tokens, we encode it with the model prompt using the embedding layer of the LLM $f_e(.)$, resulting in $\{g_i\}_{i=1}^{N_T}$ with $g_i = f_e(G_i)$. We then concatenate the aligned visual features $\{z_i\}_{i=1}^{T}$ with the goal text embeddings $\{g_i\}_{i=1}^{N_T}$ as the input to the LLM transformer trunk. The LLM thus generates the future action sequence conditioned on the video content and given text prompt that includes the goal text, as shown in Fig. 2. The full text prompt is given in appendix.

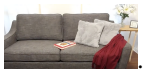
Task Type	Observation	Instruction	Output
Goal Modality Aug.		Goal: Install Sofa. What are the next 3 steps?	1. install legs of sofa 2. install armrest on sofa 3. put on sofa cover.
	The legs of the sofa are separate from the base.	Goal:  What are the future 4 steps?	1. install legs of sofa 2. install armrest on sofa 3. put on sofa cover 4.put every parts mentioned together.
		Goal: N/A. What are the future 3 steps?	1. install legs of sofa 2. install armrest on sofa 3. put on sofa cover.
Goal Prediction		What is the person's goal?	Install Sofa

Table 1. **List of auxiliary tasks.** For each of these, the observation can either be a video, image, or in text. **Goal Modality Augmentation (GMA)** varies how the goal is specified as an input. **Goal Prediction (GP)** requires the agent to understand the observation and predict the goal of the user.

3.3. Auxiliary Task Augmentation

To enhance the model's planning abilities, we design a set of auxiliary tasks relevant to visual planning and generate task-specific data. Our model is then trained jointly on the VPA and these auxiliary tasks. The core idea is to re-use existing annotations in novel ways that go beyond the input-output combination for VPA (see Fig. 1), without the need for any additional human labeling. Different from prior instruction data construction approaches [20, 25], our method is specifically designed for the visual planning task. We focus on generating instruction tuning data for long horizon video-based planning and use long instructional videos for this purpose. Additionally, Auxiliary Task Augmentation includes variants that change input modality (video to image or text) while most prior methods [20, 25] retain the input video. The proposed two types of auxiliary tasks and corresponding instructions are shown in Tab. 1. Below, we introduce the auxiliary tasks in detail.

Goal Modality Augmentation (GMA). In the VPA task, the observation $\mathcal{O}$ is a video and the goal $\mathcal{G}$ is a natural language text, as described in the Sec. 3.1. We generate modality-augmented auxiliary tasks by either changing the goal $\mathcal{G}$ from text to image or discarding it. To convert the $\mathcal{G}$ from text to image, we utilize the existing action segment annotations in our datasets. Specifically, we first identify the end time of the last action to be predicted, and use the corresponding last frame as the image for the goal $\mathcal{G}$.

Goal Prediction (GP). Given the current observation $\mathcal{O}$, the model needs to predict the goal $\mathcal{G}$ of the person in the form of natural language. This task, along with the auxiliary tasks described above, has variations where the observation $\mathcal{O}$ is represented as either a video, text, or an image. To change $\mathcal{O}$ from video to text, we generate textual object states to replace the video. Specifically, we feed the action label and the high-level task goal to the LLMs and prompt for descriptions about possible object states before that action. More details can be found in the appendix. To change the observation from video to image modality, we simply take the last frame of the input video as the observation.

3.4. Multi-Token Prediction

Recall that visual planning entails a structured action space both in terms of individual and sequence of actions. The standard next-token prediction is too unconstrained to take advantage of such a setting, especially in a low-data regime. Instead, we propose to use multi-token prediction [9] as a way to force the model to explicitly reason about future tokens while predicting the next-token, a beneficial trait for the visual planning task at hand.

Background. Suppose $x_{1:T} = \{x_1, x_2, \ldots, x_T\}$ are the input embeddings to the language model, where T is the number of input tokens. Traditional language models use next-token prediction loss as the training objective:

$$\mathcal{L}_{next} = -\sum_{t=1}^{T} \log P_\theta(x_{t+1} \mid x_{1:t}) \tag{1}$$

where θ are trainable parameters of the language model.

In multi-token prediction [9], the model needs to predict N future tokens at once during training. The training objective minimizes the cross-entropy loss for each future token:

$$\mathcal{L}_{multi} = -\sum_{i=1}^{N}\sum_{t=1}^{T} \log P_\theta(x_{t+i} \mid x_{1:t}) \tag{2}$$

We consider the language model with multi-token prediction as one shared backbone and multiple output heads. Therefore, we can compute

$$P_\theta(X_{t+i} \mid X_{1:t}) = \text{softmax}(h_i(f(X_{1:t}))) \tag{3}$$

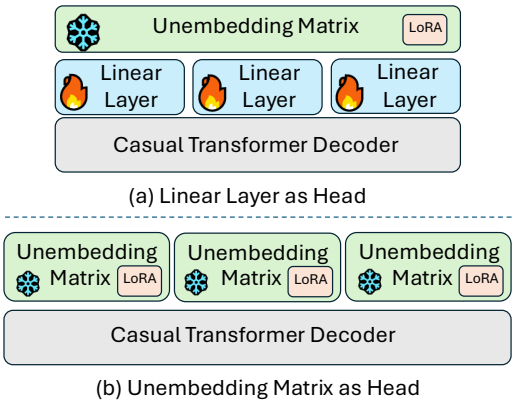


Figure 3. **Different Head Architecture for Multi-token Prediction. Top:** The original MTP [9] introduces additional linear layers as the heads and shares the unembedding matrix on top of each head. **Bottom:** We reuse the unembedding matrix as the heads. During training, we initialize each unembedding matrix with the same pre-trained weights but add different LoRA modules.

where f is the shared backbone and h_i is the i-th output head. During inference, the model keeps only the next token prediction head, disabling other heads. It then generates next tokens autoregressively, as illustrated in Fig. 4(b).

Note that the original MTP [9] is designed for large-scale pre-training on standard NLP tasks, while the visual planning tasks fall within the low-data regime and the output action space is more structured. Therefore, we propose a modified head architecture for MTP. As shown in Figure 3, the original MTP introduces additional linear layers as the heads. All heads share the same unembedding matrix, which is used to map the output embeddings of the LLM to the token indices. As a comparison, our method does not introduce new layers by reusing the unembedding matrix. Specifically, we duplicate the pre-trained unembedding matrix for K times as the heads, where K is the number of future tokens to be predicted. We freeze the weights of the heads and add different LoRA modules for different heads. Section 4.3 show that our novel head architecture achieves better performance than the original MTP, despite having significantly fewer trainable parameters.

3.5. Multi-stage Training

Inspired by prior works [13, 21], we adopt a three-stage training pipeline to maximize the effectiveness of Auxiliary Task Augmentation and Multi-token Prediction (Fig. 2).

Feature Alignment. Following common practice in training MLLMs [17, 18, 29], we freeze both the visual encoder and the LLM, training only the visual adapter. The aligns visual features with the LLM's input embedding space.

Auxiliary Task Pre-Training. We freeze both the visual encoder and the visual adapter, training the LLM on all auxiliary tasks. This stage enables the model to learn task dynamics and understand user intentions, which is critical for goal-oriented planning.

Primary Task Fine-Tuning. Finally, we fine-tune the LLM on the VPA task directly with the visual encoder and the visual adapter frozen. In this stage, we enable Multi-token Prediction to model the structured label space. We do not use MTP in prior stages due to the different label space structure between the auxiliary tasks and the VPA task.

3.6. Implementation Details

Motivated by transparent data curation and privacy policies, we use MetaCLIP [8] with ViT-L-14 [8] as our visual encoder and Llama-2-7B [41] as our LLM, freezing the visual encoder and fine-tuning the LLM with LoRA [12]. Following prior works [17, 23, 29], we use a linear layer as a visual adapter. We uniformly sample 100 frames from the input video at 0.5 FPS. When a video is shorter than 50 seconds, we sample the maximum amount of frames at 0.5 FPS. For feature alignment, we use a subset of LAION dataset with $550K$ image-text pairs. We use 4 additional heads for MTP, predicting 4 future tokens in addition to the next token during training. We provide more implementation details in the appendix.

4. Experiments

4.1. Setup

We evaluate our method on two tasks: Visual Planning for Assistance (VPA) and Long-term Action Anticipation (LTA). For VPA, we evaluate on two widely used instructional video datasets, COIN [40] and CrossTask [55]. For LTA, we evaluate on Ego4D [10]. Both VPA and LTA use untrimmed videos as input. VPA provides a specified goal to the model, whereas LTA focuses on action anticipation without an explicit goal. Additionally, the planning horizon in VPA is set to predict 3 or 4 steps, while LTA requires the model to anticipate 20 actions in the future.

Datasets. The COIN [40] dataset is a large-scale instructional video dataset designed for understanding complex tasks across various domains. It includes over 11,827 videos with 180 different tasks, such as cooking, DIY, and other household activities. The actions are labeled with one natural language description (*e.g.*, *'install sofa leg'*), start time, end time, and the high-level task (*e.g.*, *'assemble sofa'*). On average, each video is 2.4 minutes long and includes about 3.6 labeled actions. The CrossTask [55]. dataset includes 2,750 videos from 18 procedural tasks. Each video is approximately 5 minutes long on average and contains around 7.6 annotated actions. Ego4D [10] contains over 3,600 hours of egocentric video of daily life activity spanning hundreds of scenarios. We focus on the subset for long-term action anticipation, which contains 3,472 annotated clips with a total duration of around 243 hours. The actions are labeled with one verb and one noun. The dataset

Method	Language Model	Visual Encoder	T=3			T=4		
			SR ↑	mAcc ↑	mIoU ↑	SR ↑	mAcc ↑	mIoU ↑
DDN [5]	-	I3D	10.1	22.3	32.2	7.0	21.0	37.3
LLM Baseline [41]	LLama-2-70B	VideoCLIP	10.2	36.6	50.8	6.1	30.5	51.5
LLM Agent Baseline [14]	LLama-2-70B	VideoCLIP	11.1	40.6	52.8	6.8	33.5	53.5
VLaMP [35]	GPT-2	VideoCLIP	18.3	39.2	56.6	9.0	35.2	54.2
VidAssist [15]	LLama-2-70B	VideoCLIP	21.8	44.4	64.4	13.8	38.3	66.3
VideoPlan (ours)	LLama-2-7B	VideoCLIP	25.6	45.3	67.7	18.2	43.3	71.9
VideoPlan (ours)	LLama-2-7B	MetaCLIP	**29.1**	**50.1**	**69.4**	**20.5**	**47.5**	**73.9**

Table 2. **Visual Planning for Assistance on COIN.** Despite not finetuning the visual encoder, VideoPlan achieves the best performance on all metrics. Specifically, our method outperforms the prior best-performing method, VidAssist, by **7.3%** and **6.7%** in Success Rate when predicting the future 3 and 4 actions, respectively.

Method	Language Model	Visual Encoder	T=3			T=4		
			SR ↑	mAcc ↑	mIoU ↑	SR ↑	mAcc ↑	mIoU ↑
DDN [5]	-	I3D	6.8	25.8	35.2	3.6	24.1	37.0
LTA [10]	-	MViT	2.4	24.0	35.2	1.2	21.7	36.8
LLM Baseline [41]	LLama-2-70B	VideoCLIP	4.6	29.7	35.6	1.1	22.2	41.3
LLM Agent Baseline [14]	LLama-2-70B	VideoCLIP	5.8	31.3	39.6	2.1	24.7	44.2
VLaMP [35]	GPT-2	VideoCLIP	10.3	35.3	44.0	4.4	31.7	43.4
VidAssist [15]	LLama-2-70B	VideoCLIP	12.0	36.7	48.9	7.4	31.9	51.6
VideoPlan (ours)	LLama-2-7B	VideoCLIP	14.4	37.4	**52.0**	8.4	34.9	53.8
VideoPlan (ours)	LLama-2-7B	MetaCLIP	**15.4**	**39.4**	51.4	**9.9**	**37.4**	**54.3**

Table 3. **Visual Planning for Assistance on CrossTask.** Our method achieves the best results across all metrics without finetuning the visual encoder. Specifically, when predicting the next 3 and 4 actions, VideoPlan outperforms the previous state-of-the-art method, VidAssist, by **3.4%** and **2.5%** in Success Rate, respectively.

has 117 types of verbs and 521 types of nouns in total.

Metrics. For VPA, we evaluate our model on three metrics. (1) Success Rate (SR) is the strictest metric. It considers the predicted sequence of actions to be success only if every predicted action is correct. (2) mean Accuracy (mAcc) calculates the average accuracy of the predicted actions at every step. (3) Mean Intersection over Union (mIoU) treats the predicted and ground truth actions as two sets. It calculates the average Intersection over Union between the predicted action set and the ground action set across the test set. For the LTA task, we calculate the edit-distance (ED) following the Ego4D LTA setup. Specifically, we report the minimum edit distance among 5 predicted verb, noun, and action sequences, for a horizon of 20 actions.

4.2. Main Results

VPA. We compare our method with prior methods on the COIN and CrossTask datasets. The results are presented in Table 2 and Table 3, respectively. We observe that VideoPlan outperforms the previous methods by a large margin in all metrics. Specifically, on COIN dataset, VideoPlan achieves $+7.4\%$ and $+5.7\%$ higher Success Rate than VidAssist [15] for predicting $T = 3$ and $T = 4$ future steps. Similarly, on the CrossTask dataset, our method outperforms the prior state-of-the-art methods by $+3.4\%$ and $+2.5\%$ in Success Rate, respectively. For fair comparison with prior methods, we also include a variant of our method with VideoCLIP as the visual encoder. From Table 2 and Table 3 we can see that with the same visual encoder, our method still outperforms VidAssist [15] on both COIN and CrossTask dataset across all metrics. Specifically, our method with VideoCLIP outperforms VidAssist [15] by $+3.8\%$ and $+4.4\%$ in Success Rate when planning for the future $T = 3$ and $T = 4$ future steps. Similarly, on the CrossTask dataset, our method outperforms VidAssist [15] by $+2.4\%$ and $+1.0\%$ in Success Rate, respectively. These results indicate that the strong performance of our model come from the proposed modules (i.e., ATA and MTP) rather than the visual encoder alone. Additionally, our language model only has 7B parameters compared to prior LLM-based methods which use language models with more than 70B parameters. These results highlight the superior video-based planning ability of our proposed method.

LTA. In Table 4, we compare our method with prior methods on the Ego4D LTA benchmark. Many existing methods [16, 36] leverage large-scale egocentric pretraining data for action anticipation. In contrast, our method does not rely on such egocentric pretraining. Despite this difference, our model still achieves the lowest edit distance on verb prediction and competitive results on noun and action prediction.

Method	Language Model	Visual Encoder	ED (verb) ↓	ED (noun) ↓	ED (action) ↓
ObjectPrompt [48]	-	CLIP	0.7004	0.7092	0.9142
PlausiVL [28]	LLama-2-7B	CLIP	0.679	0.681	-
AntGPT [51]	LLama-2-70B	CLIP	0.6531	0.6446	0.8748
Vamos [45]	LLama-2-7B	CLIP	0.643	0.650	0.868
EgoVideo [36]	Vicuna-7B	EgoVideo	0.6354	0.6367	0.8504
PALM [16]	LLama-2-13B	EgoVLP	0.6471	**0.6117**	**0.8503**
VideoPlan (ours)	LLama-2-7B	MetaCLIP	0.6491	0.6504	0.8746
VideoPlan (ours)	Vicuna-7B	CLIP	**0.6340**	0.6395	0.8649

Table 4. **Long-term Action Anticipation on Ego4D.** We report our model's performance on the test set, and de-emphasize the methods that are pretrained on large-scale egocentric data. Our method achieves the lowest edit-distance on verb prediction and competitive results on noun and action prediction. Compared with prior methods that are not pretrained on egocentric data, our model achieves the best performance across all metrics.

ATA	MTP	T=3			T=4		
		SR	mAcc	mIoU	SR	mAcc	mIoU
✗	✗	25.7	46.4	67.4	17.5	43.6	71.9
✓	✗	27.9	48.6	67.3	18.8	45.1	70.8
✗	✓	27.7	48.2	68.7	19.2	45.1	73.3
✓	✓	**29.1**	**50.1**	**69.4**	**20.5**	**47.5**	**73.9**

Table 5. **Effects of Auxiliary Task Augmentation and Multi-token Prediction on COIN dataset.** Both Auxiliary Task Augmentation (ATA) and Multi-token Prediction (MTP) improve planning ability. Our best model (ATA + MTP) outperforms the baseline model (without ATA and MTP) by **3.4%** and **3.0%** on Success Rate when predicting the next 3 and 4 actions respectively.

Aux. Tasks	T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU
All	**29.1**	50.1	**69.4**	**20.5**	**47.5**	**73.9**
w.o. GMA	28.1	48.8	69.0	19.6	46.3	73.0
w.o. GP	28.9	**50.3**	69.2	19.6	46.9	72.9
w.o. All	27.7	48.2	68.7	19.2	45.1	73.3

Table 6. **Breakdown of Auxiliary Tasks on COIN dataset.** GMA: Goal Modality Augmentation. GP: Goal Prediction. Our model with all auxiliary tasks performs the best in SR for both $T = 3$ and $T = 4$. Removing GMA and removing GP both lead to a drop in SR, indicating that each auxiliary task improves the model's planning ability.

Notably, when compared with prior methods that are also not pretrained on egocentric data, our model outperforms all others across all metrics. These results validate the generalizability of our proposed method.

4.3. Ablations and Discussion

In this section, we perform various ablation studies on Auxiliary Task Augmentation and Multi-token Prediction. We conduct all experiments on the COIN dataset.

Auxiliary Task Augmentation. We study the effects of Auxiliary Task Augmentation (ATA) in Table 5. From the results, we can observe that ATA consistently improves the model's performance across most of the metrics. Specifically, when predicting 3 and 4 future actions, ATA improves our baseline model (without MTP and ATA) by 2.2% and 1.3% in success rate, respectively. Additionally, ATA enhances the success rate of the MTP-only model by 1.4% and 1.3%, respectively. These results show the effectiveness of ATA in enhancing the model's visual planning ability.

Multi-token Prediction. Table 5 shows the impact of Multi-token Prediction (MTP) on model performance. The results show that MTP consistently boosts model performance across all metrics. Specifically, when predicting 3 and 4 future actions, MTP enhances the success rate of the our baseline model (without MTP and ATA) by 2.0% and 1.7%, respectively. Additionally, MTP improves the ATA-only model by 1.2% and 1.7% in success rate, respectively. These results demonstrate the effectiveness of MTP in improving the model's visual planning abilities.

Auxiliary Tasks Analysis. We analyze the effect of each auxiliary task type by removing them from our best-performing model variant. As shown in Table 6, our model with all auxiliary tasks performs best in SR on the COIN dataset when planning for the both T=3 and T=4 future action steps. Specifically, when removing Goal Modality Augmentation (GMA), we observe 1.0% and 0.9% drop in Success Rate for T=3 and T=4. Meanwhile, removing Goal Prediction (GP) leads to 0.2% and 0.9% drop in Success Rate, respectively. These results indicate that each auxiliary task improves the model's planning ability. We hypothesize that GMA brings the most significant performance improvement because introducing goal specifications (text, image, or video) allows the model learn cross-modal goal dependencies via step actions, leading to better generalization, e.g., wash carrots is a good first step for a goal make carrot cake or any goal of a dish with carrots.

Head Architecture for Multi-token Prediction. We compare the performance of our head design with the original MTP [9] in Table 7. Our head design achieves the best results across all metrics, indicating the effectiveness of our

Head Type	Head Params	T=3 SR	T=3 mAcc	T=3 mIoU
Linear Layer [9]	80M	28.2	49.4	68.5
Unembedding Matrix (ours)	11M	**29.1**	**50.1**	**69.4**

Table 7. **Performance of Different Head Architectures on COIN Dataset.** The head parameters are computed with 4 extra heads. Ours outperforms the original MTP [9] across all metrics, showing the effectiveness of our head architecture.

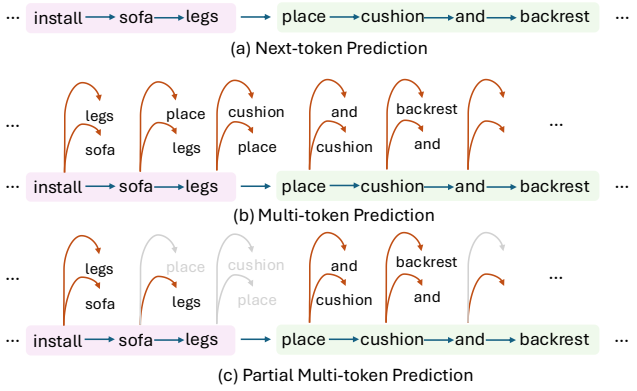


Figure 4. Token prediction schemes operating on a sequence of actions. (a) **Next-token Prediction (NTP)** only considers the following token (blue arrows). (b) **Multi-token Prediction (MTP)** also reasons about future tokens via parallel token heads (red arrows). (c) **Partial MTP** is an ablation on MTP where additional heads are constrained to not extend beyond action boundaries (gray arrows are inactive heads). Our experiments show MTP outperforms both NTP and partial MTP, thus more effectively capturing the temporal dependency in the label space for visual planning.

method. We hypothesize this is because our head design leads to less trainable parameters (11M) compared to the original MTP [9] (80M). The extra heads will be discarded during inference. We hypothesize that our lightweight head design forces the model to update the shared LLM backbone more than the extra heads, thus enabling it to better capture the structured action space.

Role of MTP in goal-based visual planning. Gloeckle et al. [9] posit that Multi-token Prediction (MTP) assigns higher implicit weights to *consequential token transitions*–transitions that are more ambiguous to predict but significantly influence the subsequent generation–thus improving the quality of overall generation. Visual planning tasks, by virtue of the temporal structure in the target action sequences, inherits many such consequential token transitions. Consider the sequence of actions shown in Fig. 4, in particular the transition between two actions: *legs* $\rightarrow$ *place*. Predicting the leading verb *place* incurs a larger ambiguity due to nature of visual planning task compared to the subsequent nouns *cushion and backrest* given rest of the context.

Method	T=3 SR	T=3 mAcc	T=3 mIoU	T=4 SR	T=4 mAcc	T=4 mIoU
NTP	27.9	48.6	67.3	18.8	45.1	70.8
partial-MTP	28.1	48.7	69.0	20.0	45.7	72.3
MTP	**29.1**	**50.1**	**69.4**	**20.5**	**47.5**	**73.9**

Table 8. **Comparison between Multi-token Prediction (MTP), Next-token Prediction (NTP) and partial-MTP.** Partial-MTP allows only the standard next-token prediction to handle consequential transitions while disabling multi-token prediction. When planning for the next 3 and 4 steps, MTP outperforms partial-MTP by 3.9% and 1.2% in Success Rate, respectively. These results show that MTP is effective in modeling the consequential transitions in the structured label space of the planning task.

Additionally, correctly predicting *place* is more critical for the fidelity of the action in the plan compared to remainder tokens within the action, thus making it a consequential transition. We hypothesize that MTP captures these inter-action transitions better by utilizing the structure in the label space. To verify this, we design an ablated version of MTP where the additional heads operate only for tokens within a given action, and remain inactive across the action boundary (shown as gray arrows in Fig.4(c)). This partial-MTP setting allows only the standard next-token prediction to handle consequential transitions while disabling multi-token prediction. From Tab. 8, we observe that the performance of partial-MTP significantly drops compared to MTP. Specifically, when predicting for the future 3 actions, the Success Rate drops by 1.0%. When predicting for the future 4 actions, the mAcc and mIoU drops by 2.2% and 1.6%, respectively. These findings indicate that partial-MTP is less effective in modeling consequential token transitions, confirming our hypothesis that MTP plays a critical role in capturing temporal dependencies within the structured label space of our task.

5. Conclusion

We introduce VideoPlan, a multimodal large language model optimized for long-horizon visual planning. To tackle the data scarcity issue of procedural annotations, we introduce Auxiliary Task Augmentation (ATA). To more explicitly leverage the structured action space unique to visual planning tasks, we adapt Multi-token Prediction (MTP) for visual planning, extending traditional next-token prediction by using multiple heads to predict multiple future tokens. Extensive experiments show that both ATA and MTP improves the model's planning ability. Our approach achieves state-of-the-art results on the COIN and CrossTask datasets for the VPA task. On the Ego4D LTA task, our method achieve the SOTA results in verb prediction and competitive results on both noun and action prediction without any large-scale egocentric pretraining, demonstrating its strong generalizability.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 3

[2] 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. 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[4] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 3

[5] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 2, 3, 6

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2

[7] 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 preprint arXiv:2312.06722*, 2023. 3

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[9] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024. 2, 4, 5, 7, 8

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 3, 5, 6

[11] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023. 2

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[13] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 2, 5

[14] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 2, 6

[15] Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Fu-Jen Chu, Kris Kitani, Gedas Bertasius, and Xitong Yang. Propose, assess, search: Harnessing llms for goal-oriented planning in instructional videos. *arXiv preprint arXiv:2409.20557*, 2024. 2, 3, 6

[16] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *European Conference on Computer Vision*, pages 140–158. Springer, 2025. 3, 6, 7

[17] 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, 5

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 5

[19] 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

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 4

[21] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 2, 5

[22] Zhiheng Li, Wenjia Geng, Muheng Li, Lei Chen, Yansong Tang, Jiwen Lu, and Jie Zhou. Skip-plan: Procedure planning in instructional videos via condensed action space learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10297–10306, 2023. 3

[23] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2, 5

[24] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3, 4

[26] Jiateng Liu, Sha Li, Zhenhailong Wang, Manling Li, and Heng Ji. A language-first approach for procedure planning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1941–1954, 2023. 3

[27] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 3

[28] Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18580–18590, 2024. 3, 7

[29] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Anymal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1314–1332, 2024. 3, 5

[30] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[31] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 3

[32] Kumaranage Ravindu Yasas Nagasinghe, Honglu Zhou, Malitha Gunawardhana, Martin Renqiang Min, Daniel Harari, and Muhammad Haris Khan. Why not use your textbook? knowledge-enhanced procedure planning of instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826, 2024. 3

[33] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *European Conference on Computer Vision*, pages 558–576. Springer, 2022. 3

[34] Yulei Niu, Wenliang Guo, Long Chen, Xudong Lin, and Shih-Fu Chang. Schema: State changes matter for procedure planning in instructional videos. *arXiv preprint arXiv:2403.01599*, 2024. 3

[35] Dhruvesh Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, and Ruta Desai. Pretrained language models as visual planners for human assistance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15302–15314, 2023. 1, 2, 3, 6

[36] Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, et al. Egovideo: Exploring egocentric foundation model and downstream adaptation. *arXiv preprint arXiv:2406.18070*, 2024. 6, 7

[37] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2

[38] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 2

[39] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 2

[40] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2, 5

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 5, 6

[42] Mrinal Verghese, Brian Chen, Hamid Eghbalzadeh, Tushar Nagarajan, and Ruta Desai. User-in-the-loop evaluation of multimodal llms for activity assistance. *arXiv preprint arXiv:2408.03160*, 2024. 2

[43] An-Lan Wang, Kun-Yu Lin, Jia-Run Du, Jingke Meng, and Wei-Shi Zheng. Event-guided procedure planning from instructional videos with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13565–13575, 2023. 3

[44] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. Pdpp: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14836–14845, 2023. 3

[45] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*, 2023. 3, 7

[46] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Octopus: Embodied vision-language programmer from environmental feedback, 2023. 2

[47] Ali Zare, Yulei Niu, Hammad Ayyubi, and Shih-fu Chang. Rap: Retrieval-augmented planner for adaptive procedure planning in instructional videos. *arXiv preprint arXiv:2403.18600*, 2024. 3

[48] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6751–6761, 2024. 3, 7

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

[50] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2938–2948, 2022. 3

[51] 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? *arXiv preprint arXiv:2307.16368*, 2023. 3, 7

[52] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634, 2024. 2

[53] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 2

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

[55] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 2, 5