

Show and Guide: Instructional-Plan Grounded Vision and Language Model

Diogo Glória-Silva, David Semedo, João Magalhães
NOVA LINCS, NOVA School of Science and Technology, Portugal
dmgc.silva@campus.fct.unl.pt
{df.semedo, jmag}@fct.unl.pt

Abstract

Guiding users through complex procedural plans is an inherently multimodal task in which having visually illustrated plan steps is crucial to deliver an effective plan guidance. However, existing works on plan-following language models (LMs) often are not capable of multimodal input and output. In this work, we present MM-PlanLLM, the first multimodal LLM designed to assist users in executing instructional tasks by leveraging both textual plans and visual information. Specifically, we bring cross-modality through two key tasks: Conversational Video Moment Retrieval, where the model retrieves relevant step-video segments based on user queries, and Visually-Informed Step Generation, where the model generates the next step in a plan, conditioned on an image of the user’s current progress. MM-PlanLLM is trained using a novel multitask-multistage approach, designed to gradually expose the model to multimodal instructional-plans semantic layers, achieving strong performance on both multimodal and textual dialogue in a plan-grounded setting. Furthermore, we show that the model delivers cross-modal temporal and plan-structure representations aligned between textual plan steps and instructional video moments.¹

1 Introduction

The research of Large Language Models (LLMs) in assisting users with complex procedural plans, such as cooking or DIY projects, presents an exciting new frontier in NLP research (Choi et al., 2022). However, while LLMs can excel at text-based conversational interactions (Chiang et al., 2023), procedural plans are inherently multimodal, often accompanied by illustrative images or instructional videos (Sener and Yao, 2019; Marin et al.,

¹The model, code, and non-personal data will be made publicly available at <https://github.com/dmgcsilva/mmplanllm>

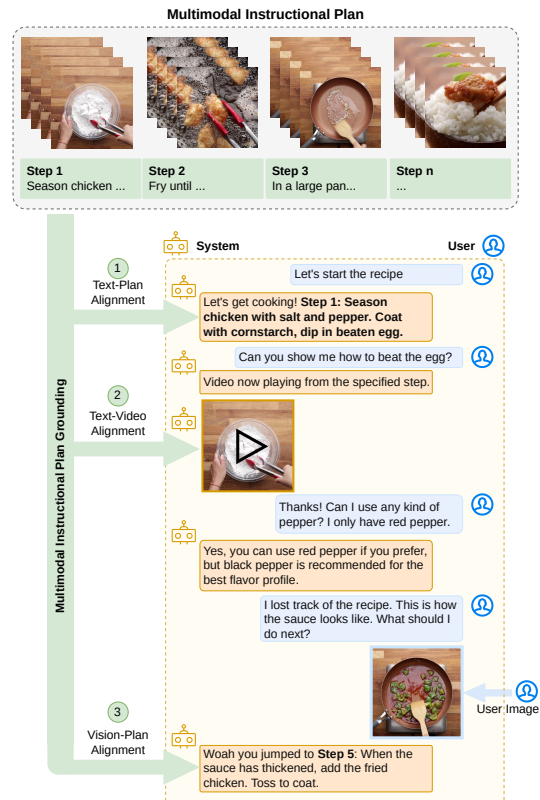


Figure 1: Example of a plan-grounded multimodal dialogue. The proposed model has the ability to understand and respond to multimodal input, provide relevant information from multiple knowledge sources, and guide the user through a complex task while adhering to a structured plan.

2019). Thus, to interact in a reliable and trustworthy manner, it is crucial that these models not only comprehend procedural plans, but also ground the dialogue on these plans and align them with the visual domain, understanding images and videos to accurately assess progress and provide helpful guidance (Figure 1).

In this work, we tackle this challenge and propose a multimodal LLM that is deeply grounded in both procedural text-plan and the accompanying visual-plan. Specifically, we focus on jointly

learning three fundamental tasks: **Plan-following** capabilities where the LLM can generate and skip steps of the plan (label 1 of Figure 1), **Conversational Video Moment Retrieval** to retrieve a relevant step-video moment that accurately describes the current plan step (label 2 of Figure 1), and **Visually-Informed Step Generation**, where, the goal is to, based on visual user input describing their current progress, generates the appropriate follow-up plan step (label 3 of Figure 1). To address these challenges, we propose **MultiModal Plan LLM** (MM-PlanLLM), a dedicated model architecture capable of guiding users through a complex task plan, while supporting textual and visual plan information, both as input and output. In particular, we extend an LLM backbone with task-specific projection layers. These allow capturing video semantic and temporal information and supporting flexible decode-time multimodal retrieval, conditioned on task plans. For training, we devise a novel multitask, multistage training approach designed to progressively instill the desired multimodal capabilities while preserving or improving on previously learned ones.

MM-PlanLLM, the main contribution of this paper, is a model capable of guiding users through complex tasks, while adhering to the user requests, grounding the plan progress on user-uploaded images through visually-informed step retrieval, and performing conversational step-video moment retrieval. In particular, its groundbreaking multimodal plan-guiding capabilities, lets it align image inputs with the correct step of the instructional plan, perform step-video moment retrieval, producing step-aligned cross-modal representations, with limited performance drop on text-only requests.

A thorough evaluation shows MM-PlanLLM’s competitive performance on text-only tasks against task-specific baselines, and substantial improvements over existing approaches on multimodal tasks.

2 Related Work

In recent years, with the release of large open source foundational models such as OPT (Zhang et al., 2022), Llama (Touvron et al., 2023a) and others (Radford et al., 2019; Brown et al., 2020a; Jiang et al., 2023a), the field of Large Language Models (LLMs) for conversational settings has received significant attention. Due to this, the contributions have been diverse, with work focusing on

improving training data (Chiang et al., 2023; Touvron et al., 2023b), scaling model size (Chowdhery et al., 2022), and adopting a Mixture of Experts (MoE) architecture (Jiang et al., 2024; Shen et al., 2023). The applications of these models are varied, such as instruction following (Brown et al., 2020b; Taori et al., 2023; Mishra et al., 2022), conversational dialogue (Zhang et al., 2020c; Chiang et al., 2023), and other task-specific applications (Raffel et al., 2020; Hosseini-Asl et al., 2020).

Researching models capable of understanding multimodal input has also been a topic of great interest. A common approach has been the usage of pretrained LLMs and Visual Encoders to achieve efficient and effective Large Vision-Language Models (LVLMs) with limited resources; however, the way these models interface has been varied. Some approaches, such as the LLaVa models (Liu et al., 2023; Sun et al., 2023) and FROMAGe (Koh et al., 2023) have found that linear projections are enough. Others deploy larger "interpretation" modules such as the Q-Former in BLIP (Li et al., 2023; Dai et al., 2023), the Visual Abstractor in mPLUG-Owl (Ye et al., 2023a), or the Perceiver (Jae-gle et al., 2021) employed in Flamingo (Alayrac et al., 2022). Another interesting approach is the modification of the internal Transformer (Vaswani et al., 2017) attention blocks such as the visual expert in CogVLM (Wang et al., 2023a) and the Modality-Adaptive Module in mPLUG-Owl2 (Ye et al., 2023b). Some work has also been done on training multimodal models from scratch such as PaLi (Chen et al., 2023), Gemini (Team et al., 2023), and Large World Model (Liu et al., 2024).

Video Moment Retrieval (VMR) is the task of, given a video and textual prompt that describes an action or event that occurs in a video, retrieving a video clip from within said video that best matches the provided textual prompt. Proposal-driven approaches focus on identifying candidate proposals and then ranking them to find the most relevant one (Gao and Xu, 2021; Wang et al., 2022; Liu et al., 2018). In contrast, others opt for a proposal-free approach that predicts the target moment directly from the video-prompt feature mappings (Wang et al., 2023b; Zhang et al., 2020a; Yuan et al., 2018) often relying on cross-modal attention modules or on learnable query embeddings such as EaTR (Jang et al., 2023) or MH-DETR (Xu et al., 2023). A common problem in VMR is the need to do extensive and expensive temporal an-

notations, an alternative is point-level VMR where the annotation is a single frame point (Jiang et al., 2023b) or a small segment (Ji et al., 2023). Recently, several approaches have been adopting a Detection Transformers (Carion et al., 2020), as it does away with the need for many hand-designed components, and tackling the problem as a direct set prediction (Lei et al., 2021a; Sun et al., 2024; Moon et al., 2023; Lei et al., 2021b).

3 Multimodal Plan-Grounded LM

In this section, we present the main elements of this work: we start by formalizing the problem, then we describe MM-PlanLLM, its architecture, and the multi-stage training process used. We end by detailing how the supporting synthetic training dataset is generated.

3.1 Problem Definition

Let $D = \langle P, T, V \rangle$ be a dialogue that consists of a procedural plan P composed of k sequential steps $P = \{s_1, \dots, s_k\}$, and a set of n user-system interaction turns $T = \{t_1, \dots, t_n\}$, where a turn $t_i = \langle U_i, R_i, I_i^* \rangle$ is composed of a user request U_i , a system response R_i and, optionally, a user-uploaded image I_i , and V a video that demonstrates how to follow the plan P . V is composed of l frames $V = \{f_1, \dots, f_l\}$. Here, a plan step is a sequence of words $s = \{w_1, w_2, \dots\}$, and a video moment m_V is the sequence of video frames denoted by its starting and ending frame $m_V = \{f_s, f_e\}$ with $f_s, f_e \in V$ and $s, e \leq l$. Each video moment represents a plan step or part of it.

Based on the user-request type, our approach simultaneously adapts and performs interleaved multimodal plan-grounded tasks. In particular, three key features are supported: general plan-grounded answer generation, conversational video moment retrieval, and visually-informed step generation. These key features are delivered by extending a vision and language model, in a multi-task setting, through a multi-stage training scheme.

3.2 MM-PlanLLM Learning

Plan-Grounded Answer Generation (PGAG).

In this task, given a dialogue D_j and the latest user request U_{i+1} the objective is to generate $R_{i+1} = \{w_1^r, \dots, w_n^r\}$ that adequately answers the user request, while conditioning on the previous turns $T_{i-c:i}$, with c being the context size and $1 \leq i < i + 1$. The objective is formulated as a

plan-grounded cross-entropy loss,

$$\mathcal{L}_{pgag} = - \sum_{t=1}^T \log P(w_t^r | w_{1:t-1}^r, U_{i+1}, D_j)$$

Conversational Video Moment Retrieval (CVMR).

This task seeks to retrieve a video moment that illustrates the current step of the task plan. Namely, given a textual user video request U_{i+1} , it seeks to retrieve the relevant video moment from a video V , given a dialogue D_j , considering only the previous turns $T_{i-c:i}$, with c being the context size. To formulate the retrieval problem, MM-PlanLLM generates a system response R_{i+1} and locates the corresponding video moment m_V within V . For tractability, we focus on retrieving a single keyframe f_m that represents moment m_v . We define f_m as the relevant segment’s middle frame, with $m = \lfloor \frac{e-s}{2} \rfloor$. Recognizing the high similarity between consecutive frames (see Appendix C), we formulate a video moment retrieval task by relaxing the retrieval target to consider a bidirectional context window of N adjacent frames. This translates to retrieving any frame in a window of $2N + 1$ frames centered around f_m . Specifically, the two-component loss is formulated as follows:

$$\mathcal{L}_{ret} = - \sum_{k=m-N}^{m+N} \log \left(\frac{P(f_k | D_j, U_{i+1})}{2N + 1} \right)$$

$$\mathcal{L}_{cvmr} = \mathcal{L}_{ret} + \mathcal{L}_{pgag}$$

Visually-Informed Step Generation (VSG).

In this last task, given a user request U_{i+1} and a user-uploaded image I_{i+1} , that visually depicts their current progress on the task being executed, the goal is to generate an appropriate system response R_{i+1} , that accurately copies the relevant plan step s , while accounting for the conversational history D_j , considering only the previous turns $T_{i-c:i}$, with c being the context size. The loss is formulated as a visually conditioned cross-entropy loss,

$$\mathcal{L}_{vsg} = - \sum_{t=1}^T \log P(w_t^r | w_{t-1:1}^r, I_{i+1}, D_j, U_{i+1}).$$

3.3 Model Architecture

The architecture of the proposed model, MM-PlanLLM, expands on the framework presented in FROMAGe (Koh et al., 2023) and is composed of three main component groups: *a) a language*

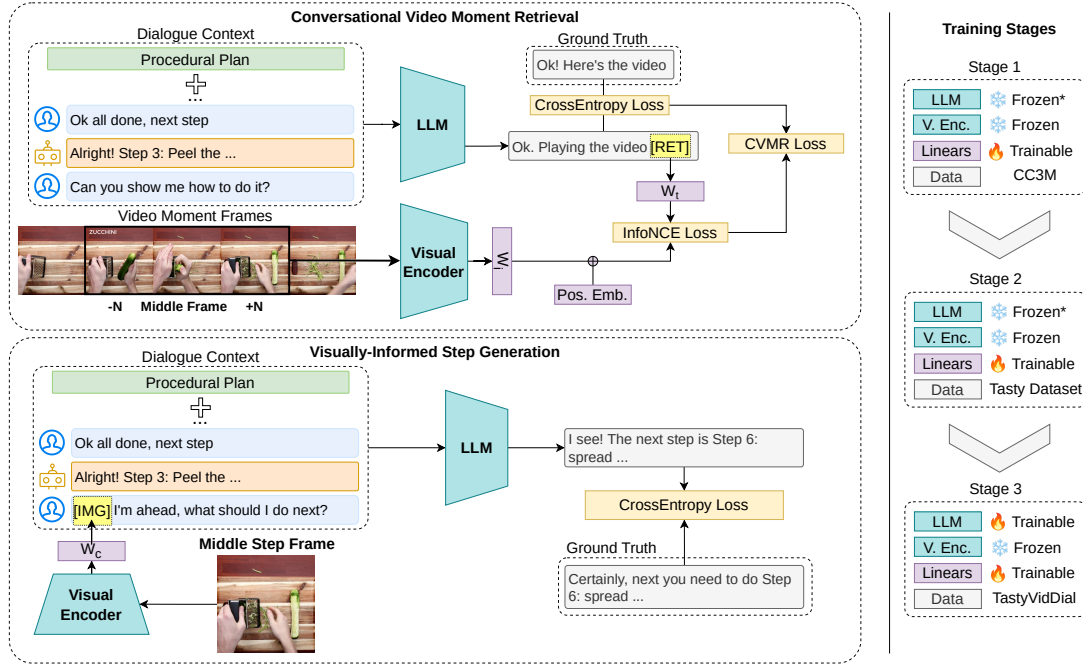


Figure 2: Comprehensive illustration of the MM-PlanLLM architecture, including the 3 training stages employed for model training. *Denotes the [RET] token embedding representations and the Language Modeling Head of the LLM remain trainable.

model backbone, b) a vision encoder, and c) task-specific projection layers. Each of these layers will be responsible for establishing an interface between the visual encoder and language model representations, while providing an efficient adaptation to new tasks, in a sequential or interleaved manner. Figure 2 provides an overview of this architecture. This section describes these three main component groups:

a) V&L Model Backbone. The vision and language backbone model, takes as input a multi-modal sequence, comprised of a user request U_{n+1} , the conversation history D , and an optional image I_{n+1} , and generates an appropriate system response. For MM-PlanLLM’s backbone model we use a pretrained decoder-only Transformer model. We experiment with different backbone models as detailed in Appendix B. The backbone LM model is trained with cross-entropy loss.

b) Video Encoder. Given a frame I_n with resolution $H \times W$, we leverage the ViT (Dosovitskiy et al., 2021) architecture, such that the video encoder outputs a learnable [CLS] token that attends to the entire frame, and a sequence of N_v visual tokens $v_i \in \mathbb{R}^{d_{ve}}$, with d_{ve} being the visual token embedding dimension. Each token is the result of attending to different non-overlapping patches

of the frame, with $\frac{H}{N_v} \times \frac{W}{N_v}$ resolution. In MM-PlanLLM this encoder remains frozen.

c) Task-specific layers. Support for novel tasks is achieved in MM-PlanLLM through task-specific projection layers:

- **VSG-specific layers.** For VSG, and general Image-to-Text support, we learn a single linear mapping $W_c \in \mathbb{R}^{d_{ve} \times d}$, with d being the LLM hidden dimension, that maps the [CLS], obtained from the visual encoder, token to the embedding space of the language model, the resulting representation used to replace the [IMG] text embedding in the LLM.
- **CVMR-specific layers.** For CVMR, the model needs to be able to retrieve the middle frame of the relevant video clip, for the current moment. In our task, each textual step is annotated with a relevant video segment. The fact that these textual steps are not directly describing the clip visual content, but rather the actions that the user has to perform, poses a greater challenge, compared to traditional VMR datasets where the captions offer a visual description of the clip.

We propose to address this challenge with a multi-stage multimodal plan-grounded

training scheme, designed to close the visual \leftrightarrow plan step semantic gap. Originally, a [RET] token is added to the language model’s vocabulary, and is then appended to the end of each retrieval request. Its decoder-output embedding is then mapped onto a cross-modal retrieval embedding space, using a trained linear mapping $W_t \in \mathbb{R}^{d \times q}$. A second linear layer is trained $W_i \in \mathbb{R}^{d_{ve} \times q}$ to map the visual features onto the retrieval space. We leverage this approach, and use the [RET] token to retrieve the video moment. For the training of these layers, we use the InfoNCE Loss (van den Oord et al., 2018) as the \mathcal{L}_{ret} loss component, where we consider the middle clip frame, plus a bidirectional context-window of N consecutive frames, as targets (i.e. positives). To incorporate temporal information we use fixed Rotary Positional Embeddings (RoPE) (Su et al., 2021) and apply temporal position shifting, where each positional embedding is shifted according to the frame’s position within the video.

3.4 Multi-stage Multimodal Training

The model undergoes a multi-stage training scheme, in two core tasks: *image captioning* and *text-to-image retrieval*. We design a three-stage training approach tailored to our setting:

Stage 1. Visual Projection Layers. This preliminary phase is focused on bootstrapping the model’s linear layers, W_c , W_t , and W_i , by training on the *image-captioning* and *image-text retrieval* tasks. For both tasks, we use the CC3M (Sharma et al., 2018) dataset, while the LLM and Visual Encoder are kept static. Only the embedding for the introduced [RET] token and the language modeling head are subject to training.

Stage 2. Task Data Specialization. The subsequent stage seeks to specialize the model in the target domain. The same previous two proxy tasks are considered, but instead of general-domain data, we use domain-specific videos and captions. Specifically, we leverage the annotations present in the Tasty Dataset (Sener and Yao, 2019). In this dataset, recipes are broken into actions, and these actions are then annotated with the start and end frame of the relevant video clip; we use these to create image-text pairs where the text is the action text and the image is the middle frame of the relevant clip.

Stage 3. Multimodal Plan-Grounded Dialogue.

The third, and most important, training stage aims to convey the necessary abilities to dialogue in the target plan-grounded dialogue setting on the recipes domain, attending to both uni- and multimodal user requests. To this extent, plan-following multimodal instructional data is used (see section 3.5), covering dialogue interactions, with particular emphasis on the envisioned multimodal interactions. To facilitate training, we start with text-only samples and then move to multimodal ones, for the latter we alternate between CVMR and VSG batches. During this phase, the LLM is fully trained, along with all of the additional linear layers.

3.5 Synthetic Multimodal Plan-oriented Training Data

To prepare the model to cope with the wide range of user requests in plan-grounded dialogues, we resort to synthetic data generation. Namely, we build upon the methodology of PlanLLM, and further incorporate multimodal queries. This methodology follows a pipeline that utilizes real user-agent dialogues and, using an intent classifier, extracts a user policy and user utterances. To generate dialogues, user intents are selected for each turn, and a combination of templates, external knowledge bases, and generative models are used to create accurate system responses. The incorporation of multimodal requests is accomplished by exploiting the Tasty Videos Dataset (Sener and Yao, 2019), which comprises culinary recipes, each accompanied by a video and annotations delineating each step into individual actions and signaling the start and end of the said actions within the video. Herein, we detail how these annotations were leveraged to integrate multimodal user requests into the pre-existing data generation pipeline.

CVMR Requests. For the retrieval of specific video moments, the target clip corresponds to the one annotated for the current action. In instances where a plan step is composed of multiple actions, the first action is considered.

VSG Requests. Regarding VSG queries, a step subsequent to the user’s current progress within the recipe is selected (e.g., if the user is at step 3, any step from 4 onwards is eligible), biasing to closer steps. The middle frame of the selected step is then used as the user-uploaded image that showcases the user’s progress, at that point in the dialogue.

For both request types, the textual user requests and system responses are sampled from handwritten template lists. To improve diversity, an external generative model is prompted to extend the lists of possible user and system utterances. To this dataset we call **Tasty Video Dialogue** (TastyVidDial).

4 Experimental Setup

4.1 Instructional Tasks Datasets

TastyVidDial. To conduct our experiments, we propose a novel dataset for conversational multi-modal dialogue over complex tasks. We create a dataset of 50k generated dialogues, between a user and a multimodal agent, while following complex tasks, resulting in $\approx 500k$ dialogue turns. We utilize a set of 1500 illustrated recipes obtained from the Tasty Videos Dataset (Sener and Yao, 2019) to ground the generated dialogues. To maximize dialogue quality, we only consider recipes with 5 to 10 steps, at least 6 ingredients, no more than 300 tokens, and at least 1 annotated video action for every step. To reduce frame count, we consider 1 for every 20 frames in the video. For training, validation, and testing we use a 90/5/5 split.

Simulated Alexa TaskBot. For the evaluation of text-only requests, we use the PlanLLM dataset as described in Glória-Silva et al. (2024). We use this dataset version to avoid dialogue turns where the user request is text-only but one of the previous turns, present in the context, is multimodal.

4.2 Methodology

Backbone Models. For the LM Backbone we use Llama2 (Touvron et al., 2023b) (results with more models in Appendix B). The visual encoder is CLIP ViT-L/14 (Radford et al., 2021). See Appendix A for more implementation details.

Metrics. For evaluation of CVMR turns, we follow recent works (Fang et al., 2023; Diwan et al., 2022; Wang et al., 2022) and use $R@n$, mean Average Precision (mAP), Step Accuracy to measure if the retrieved frame is inside the video moment for the relevant step, and Mean Normalized Frame Distance (MNFD) $MNFD = \frac{1}{N} \sum_{i=1}^N \frac{|f_{retrieved,i} - f_{target,i}|}{F_i}$, where $f_{retrieved,i}$ and $f_{target,i}$ are the retrieved and target frame respectively.

For the automatic evaluation of answer generation, we consider BERTScore(BS) (Zhang et al., 2020b) and ROUGE-L (Lin, 2004). To measure the

Model	Answer Gen.		Plan-Navigation	
	ROUGE	BS	Explicit	Implicit
FROMAGe	29.98	63.55	—	—
PlanLLM	75.58	88.66	0.895	0.480
MM-PlanLLM	66.58	83.28	0.855	0.440

Table 1: Instructional plan following generation results, on automatic metrics. PlanLLM results as reported in (Glória-Silva et al., 2024)

VSG performance, apart from ROUGE-L, we use Exact Match, which measures whether the target step is contained, or not, in the system’s response.

Protocol. Across all dialogue-based evaluations, we consider a context window of the 4 previous turns and pass the model the recipe steps along with the current step the user is on. The steps are included in the prompt in a numbered manner (eg, "Step 1 ..., Step 2 ..."). For CVMR, the model also sees the candidate system response and we extract the output embeddings for the position immediately after the generated [RET] token, if the model fails to generate the [RET] token we use the output embeddings of the first generated token. We set $N = 2$. For VSG and Answer Generation the model does not see any additional context. When evaluating a specific task, the model is not provided with any marker or information indicating the type of response wanted. For CVMR the candidate pool is composed of all of the frames of the recipe instructional video and the negative frames are the target frames for other samples in the same batch.

Baselines. As a baseline, we compare our approach with FROMAGe on a zero-shot setting, as it was not fine-tuned for our domain specifically. We also compare against a random baseline that randomly retrieves a frame from the video for CVMR, and, for VSG, it randomly selects a plan step from the ones not yet completed by the user. For textual requests, we compare against the PlanLLM model, with no further fine-tuning, to gauge performance variance on text-only dialogue turns.

5 Results and Discussion

5.1 Plan Grounding

Plan-Grounded Answer Generation. A key property of MM-PlanLLM is its strong plan-following capabilities. To assess this, we evaluated MM-PlanLLM on text-only dialogues, mirroring the main evaluation setting of PlanLLM. We uti-

Method	LLM Backbone (# Params)	Conversational Video Moment Retrieval						VSG	
		R@1	R@5	R@10	mAP	Step Acc.	MNFD↓	Ex. Match	ROUGE
Random	—	1.65	8.66	17.10	7.70	16.12	32.21	28.02	37.51
FROMAGe	OPT (7B)	3.08	11.75	22.76	10.17	25.09	26.11	0.34	7.31
MM-PlanLLM	Llama2 (7B)	5.50	38.53	53.82	21.52	54.10	13.26	38.16	42.62

Table 2: Evaluation results of our best-performing model MM-PlanLLM-Llama2, on multimodal tasks, against the baselines. For the CVMR and VSG tasks we used the TastyVidDial dataset.

lized the original PlanLLM dataset, which exclusively comprises text-based conversations.

As shown in Table 1, MM-PlanLLM achieves a BERTScore of 83.28, approximately 94% of PlanLLM’s performance (88.66), on answer generation in a text-only plan-grounded setting. In contrast, FROMAGe demonstrates notably weaker performance in this setting. To understand if this performance differential also reflects on MM-PlanLLM’s ability to guide users through tasks, we replicated the GPT-4-based Plan Navigation evaluation from Glória-Silva et al. (2024). The results in Table 1 indicate that MM-PlanLLM remains competitive on this task having 85.5 accuracy on explicit navigational requests, a 4.0 accuracy loss over PlanLLM, reinforcing that it retained the ability to effectively follow instructional plans.

Conversational Video Moment Retrieval. To assess CMVR performance, we evaluated MM-PlanLLM on all video moment retrieval requests within the TastyVidDial test set. We benchmarked MM-PlanLLM against FROMAGe, which is capable of general conversational image retrieval, and the random baseline (described in Section 4.1), to quantify the gains achieved through our task-specific training approach.

The results shown in Table 2 highlight the efficacy of our focused training. MM-PlanLLM significantly outperforms FROMAGe across all metrics, demonstrating over 100% improvement in most cases. Whereas FROMAGe, in turn, demonstrates minimal improvement over the random baseline. The performance gap between R@1 and R@5, coupled with a high Step Accuracy, suggests that while MM-PlanLLM consistently identifies the relevant video moment (evidenced by high Step Accuracy), the high visual similarity between adjacent frames within the same video moment proves a challenge for R@1 scores. This is explored in Section 5.2.

Visually-Informed Step Generation. To evaluate MM-PlanLLM’s ability to interpret visual input and align it with instructional plans, we also eval-

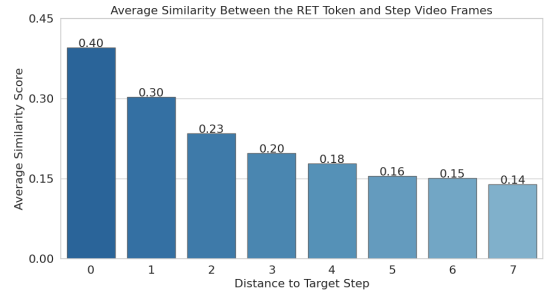


Figure 3: Text-query to visual plan alignment. MM-PlanLLM effectively learns to align textual [RET] token representations with that of the target step frames. We remove outliers for clarity.

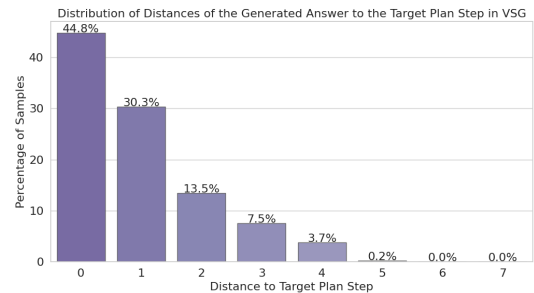


Figure 4: Image-query to text plan alignment. Most similar plan step to the provided visual input, as measured by BS using the generated answer.

uated solely in the VSG requests. Results for this task are shown in the second column group of Table 2. There is a stark contrast in performance between MM-PlanLLM and FROMAGe, with the latter rarely preserving the step text verbatim. Conversely, MM-PlanLLM achieves an Exact Match score of 38%.

5.2 Multimodal Plan Alignment

Text to Visual Plan. In Figure 3 we plot the average similarity between the [RET] token and all frames of each plan step in the plan video, ordered by their absolute distance to the target step. The results demonstrate that MM-PlanLLM effectively learns to produce a representation that aligns closely with the video moment relevant to the target step. This is supported by the significantly higher

	Conversational Video Moment Retrieval						VSG		Answer Gen.	
	R@1	R@5	R@10	mAP	Step Acc.	MNFD↓	Ex. Match	ROUGE	ROUGE	BS
LLaMa2 - Phase 1	2.05	12.13	19.02	9.01	16.42	29.25	0.00	6.48	31.78	63.44
+ Phase 2	3.45	15.21	23.04	10.68	16.98	29.46	7.33	18.87	23.68	55.30
+ Phase 3	6.72	35.26	48.69	20.80	52.52	14.03	37.14	42.84	66.11	83.03
+ Adj. Frames	7.46	30.50	48.60	20.43	52.05	14.14	34.58	40.59	65.78	82.95
+ Pos. Embs.	5.50	38.53	53.82	21.52	54.10	13.26	38.16	42.62	66.58	83.28

Table 3: Impact of the several training stages on model performance on the three main tasks.

similarity scores observed for frames within the target step (distance 0) compared to frames from other steps. Additionally, the gradual decline in similarity as the distance from the target step increases, further confirms the model’s ability to discriminate between relevant and irrelevant video moment frames based on the textual plan.

Image to Text Plan. To assess MM-PlanLLM’s ability to align visual representations of steps with the corresponding textual descriptions, we used BERTScore to measure the similarity between generated answers and plan steps in the VSG task. Then, for each VSG instance, we identified the plan step with the highest BERTScore similarity to the generated answer and plotted its distance from the actual target step in Figure 4.

The distribution in Figure 4 reinforces that MM-PlanLLM demonstrates a substantial capacity for aligning visual input with the corresponding textual step, achieving a success rate of 44.8% on the test set. Moreover, 30.3% generated answers are most similar to steps immediately preceding or following the target step, highlighting the model’s ability to capture the sequential nature of instructions and identify steps closely related to the visual input.

5.3 Ablation Study

We conducted ablation studies to investigate the impact of each training stage and architectural choices on model performance across all tasks.

Training Stages. To train MM-PlanLLM, we devised a multi-stage approach to maximize performance gains and minimize catastrophic forgetting. This analysis can be seen in Table 3. These results show that each of the three training stages contributed incrementally to improving the targeted capabilities. Stage 1, which focused on general image understanding, established a foundation for the model to outperform the random baseline in CVMR. Stage 2, which aimed to instill domain-specific multimodal understanding, further

enhanced performance on both CVMR and Step Generation tasks, even before explicitly training on these tasks. Finally, Stage 3, where we integrated conversational abilities, led to substantial improvements across all three tasks, highlighting the importance of end-to-end task-specific training.

Within the last stage, we also report the improvements provided both by the usage of adjacent frames as candidates and usage of positional embeddings for CVMR training. Surprisingly, utilizing multiple candidate frames for CVMR training yielded minimal benefits. We hypothesize that this is due to the high similarity between consecutive frames in video moments. However, the addition of positional embeddings, which incorporate temporal information, significantly improved performance across the board, underscoring the model’s ability to leverage this additional context.

LLM Backbone. To understand how different LLM Backbones affect model performance we evaluated 8 LLMs on CVMR, VSG, and PGAG. The evaluated models ranged in size from 1.8B (Qwen-1.5 (Bai et al., 2023)) to 7B parameters (Vicuna1.5 (Chiang et al., 2023)). We consider PlanLLM as a backbone but skip the text-only data training in Stage 3. On the CVMR evaluation Llama2 achieves the best performance across most metrics, with the exception of R@1 that is led by PlanLLM with 6.16 R@1. For VSG, PlanLLM exhibited a substantial lead over Vicuna in Exact Match (42.76 vs. 40.37), while Mistral (Jiang et al., 2023a) achieved the highest ROUGE score (46.60).

In the answer generation task, most models demonstrated similar performance. Our approach generalizes well for smaller LLMs, such as Qwen-1.5 (1.8B Param.) and Phi-2 (2.7B Param.) which achieved, on average, 88% and 95% of Llama2’s performance, respectively. The complete results are shown in the Appendix B.

6 Conclusion

We propose MM-PlanLLM, a multimodal architecture that enables multimodal comprehension for LMs in plan-grounded conversational settings. We follow a multistage training paradigm, coupled with task-specific synthetic data creation, that enables the model to slowly acquire the necessary abilities to understand multimodal input and generate multimodal outputs.

Experimental results demonstrates that MM-PlanLLM outperforms task-specific baselines, showcasing minimal performance loss in text-only dialogues, while being capable of aligning textual steps with video moments and user images with the plan steps. The ablation study further highlights the effectiveness of the multi-phase training methodology and the value of incorporating temporal information.

Limitations

While MM-PlanLLM addresses two key multimodal request types (CVMR and VSG) crucial for plan-grounded dialogue, we acknowledge that a complete system would need broader multimodal support, including visual question answering. Furthermore, long-term dialogue dependencies remain a challenge due to the limited context window of 4 turns during training (limited by the available hardware), hindering the model’s ability to effectively recall and utilize information from earlier turns in the conversation. This limitation may impact the model’s performance in extended interactions where maintaining context is essential. We plan to address these limitations in future work.

Acknowledgements

This work was supported by the FCT Ph.D. scholarship grant Ref. PRT/BD/152810/2021 awarded by CMU Portugal Affiliated Ph.D. program, and by the FCT project NOVA LINC Ref. (UIDB/04516/2020). Data collection was possible under the Alexa Prize Taskbot Challenge organized by Amazon Science.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei,

Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. [Language models are few-shot learners](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20, Red Hook, NY, USA*. Curran Associates Inc.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. [End-to-end object detection with transformers](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Good-

- man, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023. [Pali: A jointly-scaled multilingual language-image model](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. [Wizard of tasks: A novel conversational dataset for solving real-world tasks in conversational settings](#). In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3514–3529. International Committee on Computational Linguistics.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Anuj Diwan, Puyuan Peng, and Raymond J Mooney. 2022. Zero-shot video moment retrieval with off-the-shelf models. *TL4NLP*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. 2023. [Multi-modal cross-domain alignment network for video moment retrieval](#). *IEEE Transactions on Multimedia*, 25:7517–7532.
- Junyu Gao and Changsheng Xu. 2021. Fast video moment retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1503–1512. IEEE.
- Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and Joao Magalhaes. 2024. [Plan-grounded large language models for dual goal conversational settings](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1271–1292, St. Julian’s, Malta. Association for Computational Linguistics.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. [A simple language model for task-oriented dialogue](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. 2021. [Perceiver: General perception with iterative attention](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.
- Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13846–13856.
- Wei Ji, Renjie Liang, Lizi Liao, Hao Fei, and Fuli Feng. 2023. [Partial annotation-based video moment retrieval via iterative learning](#). In *Proceedings of the 31st ACM International Conference on Multimedia, MM ’23*, page 4330–4339, New York, NY, USA. Association for Computing Machinery.

- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023a. [Mistral 7b](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. [Mistral of experts](#).
- Xun Jiang, Zailei Zhou, Xing Xu, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023b. [Faster video moment retrieval with point-level supervision](#). In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 1334–1342, New York, NY, USA. Association for Computing Machinery.
- Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. [Grounding language models to images for multimodal inputs and outputs](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 17283–17300. PMLR.
- Jie Lei, Tamara L Berg, and Mohit Bansal. 2021a. [Detecting moments and highlights in videos via natural language queries](#). *Advances in Neural Information Processing Systems*, 34:11846–11858.
- Jie Lei, Tamara L. Berg, and Mohit Bansal. 2021b. [Detecting moments and highlights in videos via natural language queries](#). In *Neural Information Processing Systems*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024. [World model on Million-Length video and language with RingAttention](#). *arXiv [cs.LG]*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#).
- Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. [Cross-modal moment localization in videos](#). In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 843–851, New York, NY, USA. Association for Computing Machinery.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. [Recipe1m+](#): A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 3470–3487. Association for Computational Linguistics.
- WonJun Moon, Sangeek Hyun, Sang shin Paldal-gu Suwon-city Park, Dongchan Park, and Jae-Pil Heo. 2023. [Query - dependent video representation for moment retrieval and highlight detection](#). *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23023–23033.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Fadime Sener and Angela Yao. 2019. [Zero-shot anticipation for instructional activities](#). In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 862–871.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning](#). In *Proceedings of ACL*.

- Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. [Mixture-of-experts meets instruction tuning: a winning combination for large language models](#).
- Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. [Roformer: Enhanced transformer with rotary position embedding](#).
- Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. 2024. [Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4998–5007.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. [Aligning large multimodal models with factually augmented RLHF](#).
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Stanford alpaca: An instruction-following llama model](#). https://github.com/tatsu-lab/stanford_alpaca.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruiho Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Ad-danki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pi-dong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Deendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona

Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobonkerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaime Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm

Levskeya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Gianoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xi-hui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Cave-ness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phueng Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-

- Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivièrè, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fjeldland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. [Gemini: A family of highly capable multimodal models](#).
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivièrè, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, and et al. 2024. [Gemma](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiohu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan. 2022. Prompt-based zero-shot video moment retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia, MM '22*, pages 413–421, New York, NY, USA. Association for Computing Machinery.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023a. [Cogvlm: Visual expert for pretrained language models](#). *CoRR*, abs/2311.03079.

Yunxiao Wang, Meng Liu, Yinwei Wei, Zhiyong Cheng, Yinglong Wang, and Liqiang Nie. 2023b. [Siamese alignment network for weakly supervised video moment retrieval](#). *IEEE Transactions on Multimedia*, 25:3921–3933.

Yifang Xu, Yunzhuo Sun, Yang Li, Yilei Shi, Xiaoxia Zhu, and Sidan Du. 2023. [Mh-detr: Video moment and highlight detection with cross-modal transformer](#). *ArXiv*, abs/2305.00355.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. [mPLUG-Owl: Modularization empowers large language models with multi-modality](#).

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. [mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration](#).

Yitian Yuan, Tao Mei, and Wenwu Zhu. 2018. [To find where you talk: Temporal sentence localization in video with attention based location regression](#). In *AAAI Conference on Artificial Intelligence*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. [Span-based localizing network for natural language video localization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [Opt: Open pre-trained transformer language models](#). *ArXiv*, abs/2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020c. [Dialogpt: Large-scale generative pre-training for conversational response generation](#). In *ACL, system demonstration*.

Stage	1	2	3
Batch Size	64	48	1/4
Grad. Acc.	64	1	64/4
Train Steps	10000	5000	2000
Val. Freq.	1000	1000	500
GPU #	1	1	1
Seq. Max Len.	24	45	800
DType	BF16	BF16	BF16
Learning Rate	$5 * 10^{-4}$	$1 * 10^{-4}$	$5 * 10^{-4}$
Scheduler	Constant	Constant	Constant
Optimizer	AdamW	AdamW	AdamW
T. Emb. Dropout	0.1	0.1	0.1
Ret. Dimension	512	512	512
LoRa DType	—	—	16 bits
LoRa Rank	—	—	4
LoRa α	—	—	8
LoRa Dropout	—	—	0.1

Table 4: Hyperparameters used to train MM-PlanLLM models across all three stages.

A Implementation Details

Table 4 details some of the hyperparameters used. Each model is trained for 10k, 5k, and 2k steps for each phase, using a batch size of 64 (and 16 on multimodal batches in phase 3) on a single A100 40GB GPU. For optimization, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1 * 10^{-5}$ for all runs. We used a constant learning rate of $1 * 10^{-5}$, for the third stage, with no warmup steps. All images are resized to fit a 224x224 image resolution. For phase 3, text-only training was separated from the multimodal training with the first 1k steps being text-only and the later 1k being multimodal. The visual encoder used was CLIP ViT-L/14 (Radford et al., 2021), the retrieval embedding dimension was set to 512, and the embedding dimension was kept the same as the LM Backbone so it varied from model to model. For the 3rd phase we use LoRa (Hu et al., 2022), when training the LM Backbone, with a $r = 4$ and $\alpha = 8$ to reduce memory requirements.

For BERTScore calculations we utilize microsoft/deberta-xlarge-mnli.

B LM Backbone Ablation

We consider a comprehensive array of language model backbones in order to assess their impact on the overall model and select the best-performing one. In particular, we consider **Qwen-1.5** (Bai

et al., 2023), **Phi-2**, **Gemma2b** (Team et al., 2024), **Mistral-v0.1** (Jiang et al., 2023a), **OPT** (Zhang et al., 2022), **PlanLLM** (Glória-Silva et al., 2024), **LLama2** (Touvron et al., 2023b), and **Vicuna-7B** (Chiang et al., 2023). As such, we cover LM backbones of different sizes, pre-training, and fine-tuning schemes.

While we report our main evaluation results in Section 5 using Llama2 (Touvron et al., 2023b) as the LM backbone, we trained a total of 8 models by varying the LM backbone. This sought to not only assert which was the best-performing model but also understand the impact of scaling the LM backbone on all tasks.

The results from this analysis, shown in Table 5, show a surprisingly low performance differential between models for all three tasks, with the only clear outlier being OPT (Zhang et al., 2022) on the answer generation task. For Conversational Video Moment Retrieval we see a clear lead for Llama2 (Touvron et al., 2023b) for most metrics, particularly for R@5, and a close second for Step Accuracy. For Step Generation, PlanLLM (Glória-Silva et al., 2024) outperforms the other models on Exact Match whereas Mistral holds a small lead on BertScore. On this task Llama2 underperforms indicating that there might be a performance trade-off between this task and the previous. For Answer Generation Vicuna performs the best likely due to its pretrain in a conversational setting, despite this both Llama2 and PlanLLM also perform closely to Vicuna. Focusing on MM-PlanLLM, on phase 3 we skipped training on text-only samples as the model already had been trained on this setting, despite this the model is still competitive across all 3 tasks showing that our training approach seems to be agnostic to the models’ pertaining tasks.

C Frame Similarity

To investigate the degree of visual similarity between frames within recipe videos, we conducted an analysis using a subset of 1446 recipe videos from our dataset, each containing 100 or more frames. For each frame within the first 100 frames of a video, we computed its cosine similarity with all other frames in the same video using a CLIP image encoder, and averaged the similarity for each frame position across all videos.

The resulting averaged similarity matrix, shown in Figure 5, confirms that frames exhibit exceptionally high similarity to their immediate neigh-

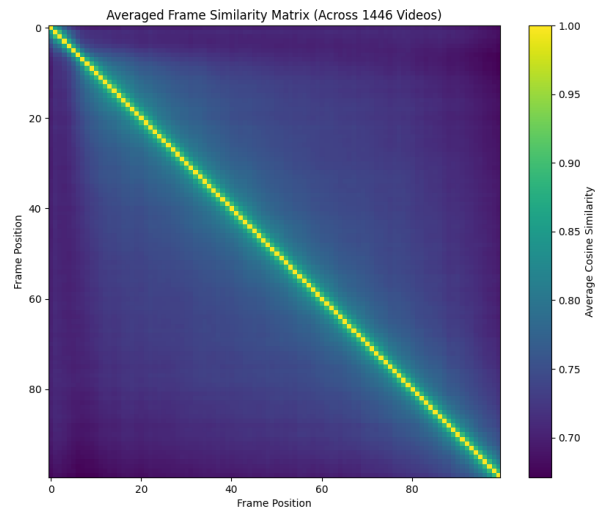


Figure 5: Average similarity of each frame against all other frames from the same video. It shows a clear bidirectional 3-frame window of higher similarity.

bors (mostly on a bidirectional 3 to 5-frame window), with a gradual drop-off in similarity beyond that point. Interestingly, we also note how similar frames from the same video tend to be with most frames having at least 0.7 similarity score to every other frame in the video. This underscores the need of visual encoders capable of differentiating the subtle visual changes that separate frames relevant to different plan steps.

D CVMR and VSG Examples

In this section, we include a few examples of CVMR (Figure 6) and VSG (Table 6) generations, extracted from the dataset test set. Additionally, in Figure 7, we showcase two dialogues collected by having a volunteer interact with the system. These examples and dialogues, demonstrate the model’s performance on both single and multi-turn scenarios, showing that it can accurately answer to a wide range of requests in the target setting.

LM Backbone (# Params)	Conversational Video Moment Retrieval						VSG		Answer Gen.	
	R@1	R@5	R@10	mAP	Step Acc.	MNFD↓	Ex. Match	ROUGE	ROUGE	BS
Qwen-1.5 (1.8B)	4.2	27.52	46.83	17.11	44.68	15.66	39.18	44.19	64.21	81.63
Gemma (2.5B)	3.08	24.44	47.39	15.69	46.46	15.45	37.14	42.30	63.33	82.07
Phi-2 (2.7B)	6.06	31.53	52.05	18.69	53.26	13.73	39.18	43.44	54.35	77.97
OPT (7B)	4.48	31.44	52.61	18.80	50.84	15.34	35.78	43.15	38.52	70.65
Mistral-v0.1 (7B)	5.69	33.40	50.56	19.70	47.95	14.51	39.52	46.60	61.47	80.81
PlanLLM (7B)	6.16	33.30	52.33	20.20	44.68	14.06	42.76	44.67	66.64	83.13
Llama2 (7B)	5.50	38.53	53.82	21.52	54.10	13.26	38.16	42.62	66.58	83.28
Vicuna1.5 (7B)	6.06	32.93	50.84	20.02	53.64	13.66	40.37	43.38	68.13	84.05

Table 5: Evaluation results of different LM Backbones for multimodal requests in the TastyVidDial dataset.



Figure 6: Five examples of CVMR results from the TastyVidDial test set. These examples demonstrate that the model is adept at identifying the key elements that should be characterized in the target frame. However, in some cases, the selected middle frame may not be ideal (3rd and 4th rows), and in others, alternative frames may more accurately represent the target plan step (1st row). The conversational dialogue context has been omitted for readability, and green boxes highlight the target frame amongst the retrieved ones.

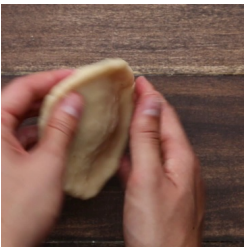
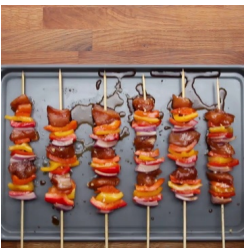
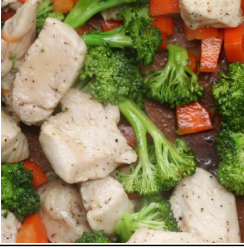
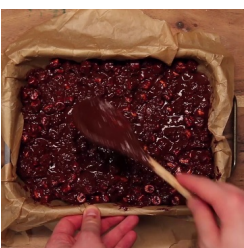

User Image		
	<p>Target:</p> <p>Gen:</p>	<p>Step 5: Pinch a piece of dough, about the size of a large egg, and roll into a ball. Using your hands, flatten the dough and spoon filling into the center. Wrap the dough into a pear shape and make sure there aren't any holes.</p> <p>Continue with Step 5, Pinch a piece of dough, about the size of a large egg, and roll into a ball. Using your hands, flatten the dough and spoon filling into the center. Wrap the dough into a pear shape and make sure there aren't any</p>
	<p>Target:</p> <p>Gen:</p>	<p>Step 8: Bake or grill for 20-30 minutes or until chicken is no longer pink.</p> <p>Continue with Step 8, Bake or grill for 20-30 minutes or until chicken is no longer pink.</p>
	<p>Target:</p> <p>Gen:</p>	<p>Step 4: Add the rice, soy sauce, and peas, and mix thoroughly.</p> <p>Continue with Step 4, Add the rice, soy sauce, and peas, and mix thoroughly.</p>
	<p>Target:</p> <p>Gen:</p>	<p>Step 4: Melt 75g of the white chocolate and stir in some of the yellow food colouring until the desired colour is reached. If it seizes on you, add a little bit of oil and stir it until it reaches a runny consistency again.</p> <p>Continue with Step 4, Melt 75g of the white chocolate and stir in some of the yellow food colouring until the desired colour is reached. If it seizes on you, add a little bit of oil and stir it until it reaches a runny consistency again.</p>
	<p>Target:</p> <p>Gen:</p>	<p>Step 5: Add the shrimp and let them cook for 3 minutes. Flip and cook for another 3 minutes.</p> <p>Continue with Step 4, Add garlic, rosemary, thyme, smoked paprika, salt and pepper. Stir and cook an additional 2 minutes.</p>

Table 6: Five examples showcasing Visually-Informed Step Generation, from the test set. The model successfully identifies the appropriate next step in most cases, leveraging dialogue history and user-uploaded images. However, the difficulty increases when uploaded images lack clarity regarding the user's progress, as seen in the final example where it is unclear if the ingredient from Step 4 have already been added. Dialogue history omitted for readability.

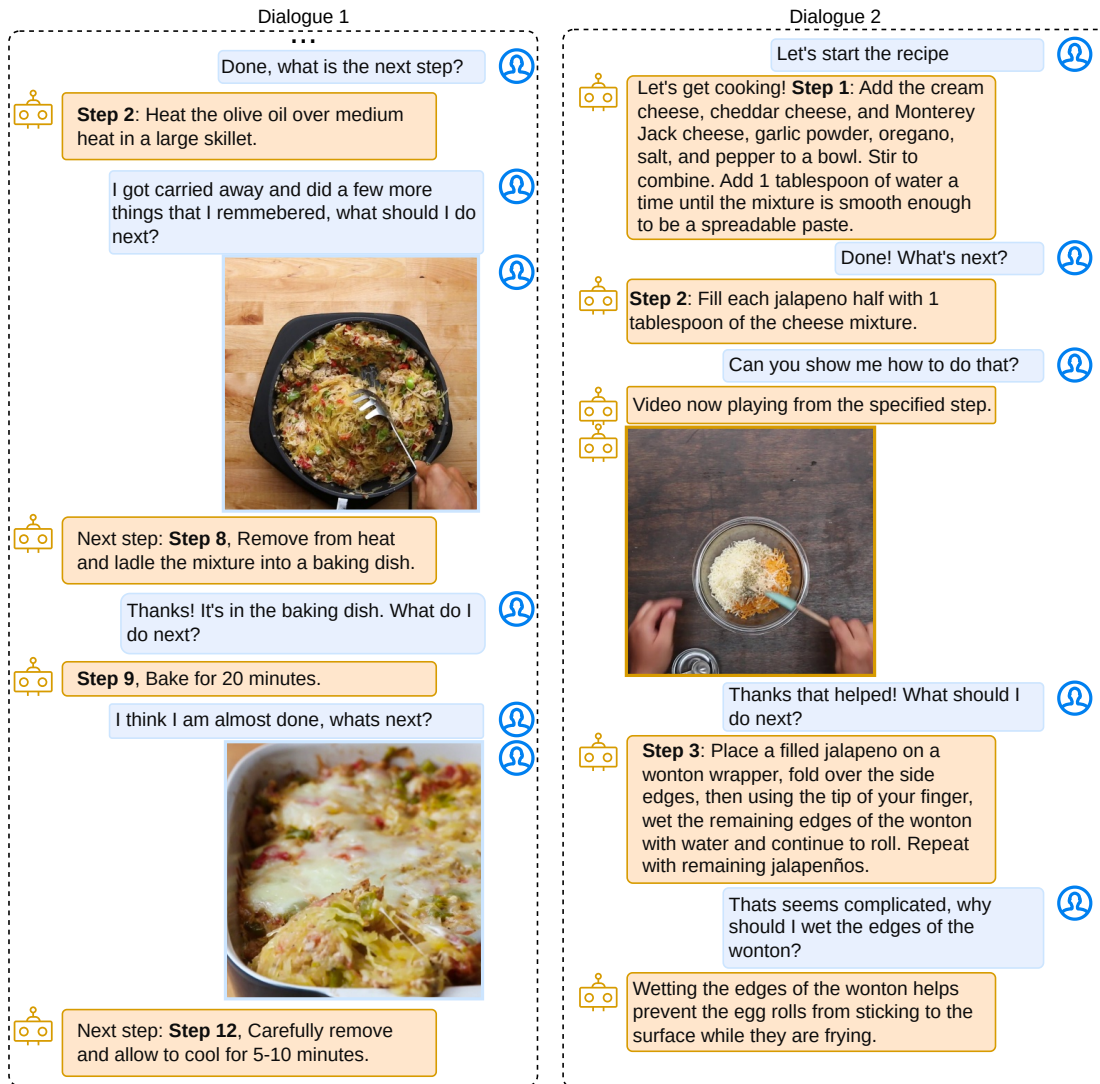


Figure 7: Real multimodal dialogues carried out by a volunteer interacting with MM-PlanLLM. These dialogues showcase the model's ability to carry out full conversations with interleaved multimodal requests, without hindering text-only performance.