VIGiA: Instructional Video Guidance via Dialogue Reasoning and Retrieval

Anonymous ACL submission

Abstract

We introduce VIGiA, a novel multimodal dialogue model designed to understand and reason over complex, multi-step instructional video action plans. Unlike prior work which focuses mainly on text-only guidance, or treats vision and language in isolation, VIGiA supports grounded, plan-aware dialogue that requires reasoning over visual inputs, instructional plans, and interleaved user interactions. To this end, VIGiA incorporates two key capabilities: (1) multimodal plan reasoning, enabling the model to align uni- and multimodal queries with the current task plan and respond accurately; and (2) plan-based retrieval, allowing it to retrieve relevant plan steps in either textual or visual representations. Experiments were done on a novel dataset with rich Instructional Video Dialogues aligned with Cooking and DIY plans. Our evaluation shows that VIGiA outperforms existing state-of-the-art models on all tasks in a conversational plan guidance setting, reaching over 90% accuracy on plan-aware VQA. [1]

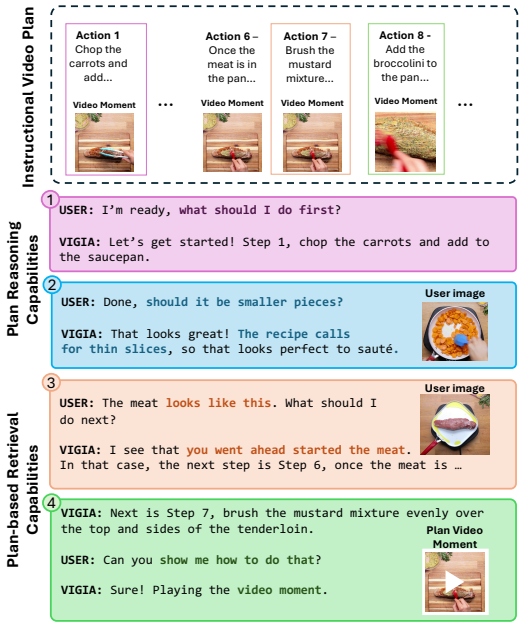


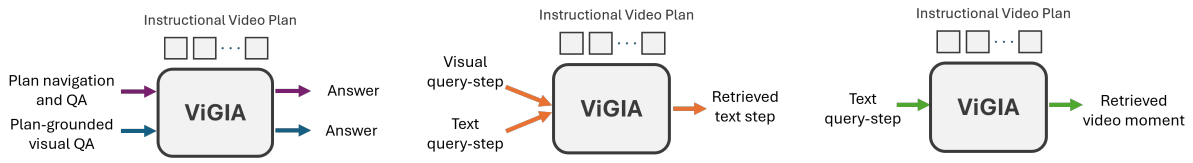
Figure 1: VIGiA is grounded on complex multimodal instructional plans, delivering unified multimodal alignment over dialogue turns by providing text-based guidance (Goal 1), plan-aware visual question answering (Goal 2) aligning plan actions and visual context (Goal 3), and retrieving relevant video moments (Goal 4).

1 Introduction

Assisting in the execution of complex instructional plans (e.g. cooking and DIY projects), in a conversational manner, is a recent task (Choi et al., 2022) that requires Large Vision and Language Models (LVLMs) to understand and reason about logical sequences of actions. As illustrated in Figure 1, this task presents a unique challenge due to the nature of the setting, where models need to be able to attend to various and complex instructions (Choi et al., 2022; Fischer et al., 2024) arbitrarily interleaved throughout the dialogue, while remaining grounded in the instructional plan and tracking explicit and implicit plan progress. Although some works address text-only plan guidance (Zhu et al., 2022; Glória-Silva et al., 2024a; Zhu et al., 2025), only a few consider multimodality (Glória-Silva et al., 2024b) and yet image-grounded reasoning tasks remain largely unexplored, a particularly challenging scenario, as it requires the model to ground its answer on the provided image, the instructional plan, the dialogue history, and its encoded knowledge.

In this paper, we address the challenge in a comprehensive manner and propose **VIGiA** (**V**isual and **I**nstructional **G**uiding **A**ssistant), a complete conversational plan guidance model with strong multimodal support. To allow for a wide array of request types, VIGiA has two key capability types: (1) **multimodal plan reasoning** capabilities and (2) **plan-based retrieval** capabilities. The first enables it to reason about uni- and multimodal requests,

[1] Model, code and dataset will be public upon acceptance.



(a) Plan reasoning capabilities.

(b) Plan-based retrieval capabilities.

Figure 2: VIGiA is an LVLM model that processes instructional video plans to navigate through steps of the plan, perform QA, VQA and retrieval of arbitrarily random text steps or video moment steps.

ground and align them on a multimodal plan, and answer accurately. The second allows it to align multimodal input queries against the instructional plan and perform both in-context retrieval of a plan action text or retrieval of the video moment of that action.

To address both plan reasoning and retrieval capabilities, VIGiA's multi-goal architecture is purpose-built to support the multiple output and input modalities, and trained in a multi-objective learning paradigm using a novel dataset, InstructionVidDial. This dataset incorporates all of these supported goals via semi-automatic dialogue generation and expands the scope by considering multiple plan domains. A multi-stage training approach progressively delivers VIGiA capacity to follow complex plan-oriented multimodal instructions and support for multi-goal unified plan-guidance (Figure 2), generalizable to multiple domains. Our main contributions are:

- **VIGiA**: A novel model that dialogues and reasons about multi-step instructional plans, while allowing for arbitrary retrieval of plan actions both in text and visual domain.
- **InstructionVidDial**: We built a novel multimodal dialogue dataset that pushes plan-guidance into the complex and diverse domains of Cooking and DIY tasks.

Our evaluation compared VIGiA to existing state-of-the-art models and results showed that it was the best model on plan-grounded reasoning QA and VQA tasks, while also delivering top performance across all plan-grounded retrieval tasks.

2 Related Work

LVLMs are models that are able to process not only text but also visual input (images and videos). These models combine the strengths of language decoders with visual encoders, enabling multimodal input and, in some cases, multimodal output. LVLMs have benefited from the advances in pretrained Large Language Models (Dubey et al., 2024; DeepSeek-AI, 2024; Team et al., 2024) as they can be paired with pretrained visual encoders such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023) to form an LVLM connected via a connector module.Several architectures have been proposed for the connector, ranging in complexity. While examples like the LLaVA family of models (Liu et al., 2023; Sun et al., 2024; Liu et al., 2024a; Zhang et al., 2023) use a linear or MLP module, others used more complex approaches, an example is the QFormer in BLIP models (Li et al., 2023; Dai et al., 2023) or the Perceiver (Jaegle et al., 2021) in Flamingo (Alayrac et al., 2022).

The need for modality equilibrium in LVLMs is key to their success, as performance in text-only settings needs to be maintained while also being able to attend to rich multimodal inputs. This led to increasingly more complex training approaches, with modern LVLMs training in a multistage paradigm (Chen et al., 2024c; Yang et al., 2024; Liu et al., 2024a,b; Laurençon et al., 2024b,a), allowing to progressively build up capabilities in a structured manner, while preventing catastrophic forgetting (McCloskey and Cohen, 1989) of previously learned tasks. The increased complexity of this has also led some works (Laurençon et al., 2024a; Tong et al., 2024; Zohar et al., 2024) to measure and verify the benefit of multistage training.

Instructional plan guidance is the multifaceted task of guiding users through complex manual instructional plans in a user-led conversational manner. Although early work has focused on a text-only setting (Choi et al., 2022; Mo et al., 2023; Fischer et al., 2024; Glória-Silva et al., 2024a), these plans are inherently multimodal and often accompanied by instructional videos. The availability of large instructional video datasets (Sener and Yao, 2019; Zhukov et al., 2019; Tang et al., 2019) makes the application of LVLMs particularly suitable for this task. This was the case of MM-PlanLLM (Glória-

Silva et al., 2024b), however, it lacked the ability to answer image-grounded questions. While VQA has been studied and applied to several domains (Mathew et al., 2020; Lobry et al., 2020; Ren and Zhou, 2020; Tapaswi et al., 2015), its application in instructional plan guidance remains unexplored.

3 Definitions and Problem Formulation

Let $P_k = \langle (a_{k,1}, vm_{k,1}), \cdots, (a_{k,m}, vm_{k,m}) \rangle$ be the instructional plan k, composed of actions $a_{k,j}$ with the text action description, which is also illustrated by the video moment $vm_{k,m}$. An instructional plan has an associated sequence of video moments $V_k = \langle vm_{k,1}, \cdots, vm_{k,m} \rangle$ illustrating the execution of the plan P_k. Each video moment $vm_{k,m} = \langle f_{k,\text{start}}, \ldots, f_{k,\text{end}} \rangle \subseteq V_k$ range from a starting frame to an ending frame. Text actions and video moments are ordered in a logical sequence, where $a_{k,j}$ or $vm_{k,j}$ must be completed before $a_{k,j+1}$ or $vm_{k,j+1}$.

Let D_k be a dialogue that is composed of multimodal interactions and associated with a specific instructional plan, and an instructional video that covers all actions. Formally, each dialogue instance is defined as $D_k = (P_k, T_k)$ where $T_k = \langle t_{k,1}, t_{k,2}, \cdots, t_{k,n} \rangle$ is the sequence of turns that constitutes the dialogue interaction for d_k. Each turn $t_{k,j}$ is defined as $t_{k,j} = (u_{k,j}, r_{k,j}, I_{k,j})$, where $u_{k,j}$ is the human input (question or request) during the j-th turn, $r_{k,j}$ is the system response, and $I_{k,j} \in \mathcal{I} \cup \{\texttt{None}\}$ is an optional image provided by the user during that turn.

This problem setting requires a model that aligns the instructional plan with the dialogue context and accurately perform complex reasoning and plan retrieval operations.

4 Methodology

Addressing the problem outlined above, requires multimodal, context-aware, and plan-grounded reasoning, which will be integrated by design into VIGiA's architecture and training strategies, as detailed in the next sections. In addition, we will also detail the supporting InstructionVidDial dataset.

4.1 Multimodal Plan Guidance LVLMs

Due to the required set of capabilities and input diversity, optimizing a single learning objective is infeasible. We address this by defining dedicated learning objectives for each type of dialogue capability.

4.1.1 Plan-grounded reasoning capabilities

The first set of dialogue capabilities support the core requests to navigate and interact with the instructional plan.

Plan-Grounded Answer Generation. The task of textual plan-grounded answer generation encompasses every textual request that requires the generation of a response based on the dialogue and the instructional plan, such as plan navigation requests and questions. Here, a textual request $u_{k,j+1}$, a instructional plan P_k and the previous conversation turns $T_{k,1:j} \in D_k$, the model needs to generate $r_{j+1} = \{w_1, w_2, \cdots, w_t\}$ that accurately answers the request. In VIGiA, the loss function for this task is the negative log-likelihood of the generated response tokens:

$$\mathcal{L}_{PGAG} = -\sum_{t=1} \log P(w_t | T_{k,1:j}, u_{k,j+1}, P_k) \quad (1)$$

Plan-aware Visual Question Answering (pVQA). This task extends the previous one to include requests that require the comprehension of user-provided images. When provided with an image $I_{k,j+1}$ and a request $u_{k,j+1}$, the model must reason about the dialogue history $T_{k,1:j}$, the instructional plan P_k, and the given image, to generate a text response. For this task, we extend the same objective as in PGAG, now also conditioned on the provided image $I_{k,j+1}$. The loss function is defined as:

$$\mathcal{L}_{\text{pVQA}} = -\sum_{t=1} \log P(w_t | T_{k,1:j}, u_{k,j+1}, I_{k,j+1}, P_k) \quad (2)$$

It is important to note that, to support pVQA, the model develops a disambiguation capability that grounds its answers in the visual question rather than the instructional video.

4.1.2 Plan-step retrieval capabilities

The second set of model capabilities aim at retrieving an arbitrary plan step: a text step or a video moment with a text only or image query.

Visually-Informed Step Generation (VSG). Given an unseen user-query-image, the model needs to align it with the plan to generate the corresponding action $a_j \in P_k$. This generation occurs with an in-context retrieval strategy that copies the step from the plan that is in the context of the model.

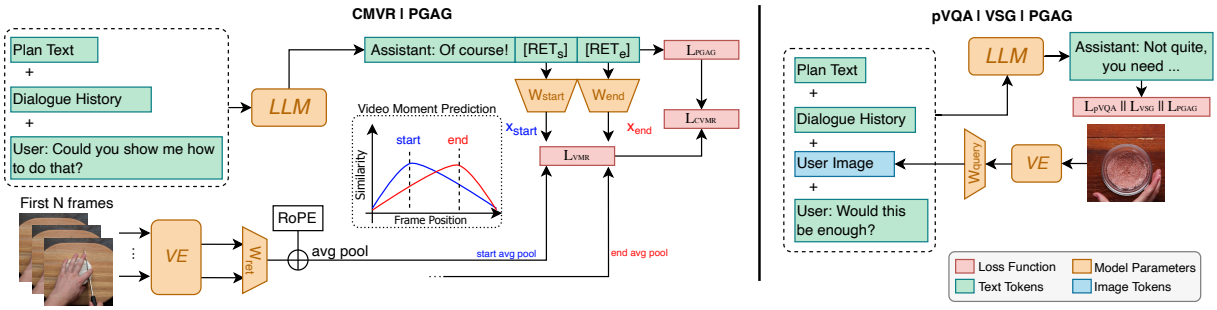


Figure 3: Global view of VIGiA's architecture. To handle multimodal inputs VIGiA combines a visual encoder and an LLM using an MLP as a connector module. For conversational video moment retrieval, VIGiA outputs a dedicated video moment start and end representation that can be used for start and end frame retrieval.

The objective function is similar to the previous one:

$$\mathcal{L}_{\text{VSG}} = -\sum_{t=1} \log P(w_t | T_{k,1:j}, u_{k,j+1}, I_{k,j+1}, P_k) \quad (3)$$

Again, the model acquires novel reasoning capabilities that allow it to navigate to a specific instructional plan action and not answer a visual question.

Conversational Video Moment Retrieval. The CVMR task involves retrieving a relevant video moment $vm = \langle f_{k,\text{start}}, \ldots, f_{k,\text{end}} \rangle \in V_k$ that corresponds to the current plan step, addressed in the request $u_{k,j+1}$. We tackle this in a principled manner by retrieving the start and end frames of the relevant video moment. Specifically, we introduce two dedicated textual tokens, `[RETs]` and `[RETe]`, whose output embeddings are trained to approximate the average representation of the first and last N frames of the relevant video moment, respectively. This is optimized using a contrastive loss, specifically InfoNCE (van den Oord et al., 2018), which tries to maximize the similarity between the `[RETs]` and `[RETe]` tokens output representations and the target representations and uses in-batch samples as negative examples. We add the PGAG loss to regularize the model to generate coherent responses to these requests. The loss function is defined as:

$$\mathcal{L}_{\text{RET}} = \text{InfoNCE}\left(g\left(\texttt{[RETs]}\right), \frac{1}{N}\sum_{i=0}^{N-1} v_{frame}(f_{\text{start}+i})\right)$$
$$+\text{InfoNCE}\left(g\left(\texttt{[RETe]}\right), \frac{1}{N}\sum_{i=0}^{N-1} v_{frame}(f_{\text{end}-i})\right)$$
$$\mathcal{L}_{\text{CVMR}} = \frac{\mathcal{L}_{\text{RET}}}{2} + \mathcal{L}_{\text{PGAG}} \quad (4)$$

where $g(x)$ denotes the projected retrieval token x embedding and $v_{frame}(f)$ the retrieval representation of f (both defined later in Section 4.2).

4.2 Model Architecture

We design a custom multi-goal architecture (see Figure 3) designed to effectively integrate multimodal information, while being flexible enough to accommodate the defined tasks:

Language Model (LM) & Vision Encoder (VE). For the LM backbone, we adopt a pretrained LLM to leverage the knowledge and capabilities learned during pretraining and instruction tuning. We use a pretrained vision encoder, $ve(\cdot)$, to obtain region-level visual features for target retrieval video frames and uploaded images. We follow Chen et al. (2024b); Laurençon et al. (2024b) and use all visual features as vision tokens. To bridge the modalities, we follow Liu et al. (2024a) and connect these 2 backbones using a 2-layer projector, W_{query}.

Retrieval Projectors. To deliver retrieval capabilities, two linear projection layers are learned $W_{start} \in \mathbb{R}^{d_{lm} \times d}$ and $W_{end} \in \mathbb{R}^{d_{lm} \times d}$, where d is the dimension of the shared retrieval space and d_{lm} is the output dimension of the LM backbone. A third layer $W_{ret} \in \mathbb{R}^{d_{f_V} \times d}$ is learned to map the visual encoder `[CLS]` token to the retrieval space. Following MM-PlanLLM, video temporal information is encoded by applying RoPE (Su et al., 2024) to the W_{ret} output features, using each frame's relative position in the video. Formally, our set of generative retrieval features are learned as:

$$\text{RET Start Tok.: } g\left(\texttt{[RETs]}\right) = W_{\text{start}} \cdot h\left(\texttt{[RETs]}\right)$$
$$\text{RET End Tok.: } g\left(\texttt{[RETe]}\right) = W_{\text{end}} \cdot h\left(\texttt{[RETe]}\right)$$

where h is the last hidden layer embedding of the LM backbone. Target and candidate video frames are mapped to the same retrieval space as:

$$v_{\text{frame}} = \text{RoPE}\left(W_{\text{ret}} \cdot ve\left(f_{k,pos}\right), pos\right) \quad (5)$$

here pos is an integer with the position of the frame in the video and $ve(\cdot)$ is the vision encoder.

4.3 Training

Multistage LVLM training has been the predominant training paradigm, with several works highlighting its benefit (Tong et al., 2024; Zohar et al., 2024; Chen et al., 2024a). VIGiA is trained with the following structure (shown in Table 5): [2]

Stage 1: Initialization. The initial stage targets bridging the modality gap between modalities, by aligning the visual features extracted by the VE with the LM representation space, with both remaining frozen. The W_{query} connector is trained on a captioning task, and the retrieval-specific projection layers, W_{ret}, W_{start}, and W_{end}, are initialized with a text-to-image retrieval task.

Stage 2: Visual Instruction Tuning. This stage is aimed at introducing strong general image understanding, in an effort to improve the final model's performance on unseen tasks. For pVQA and VSG, the model is trained end-to-end on general visual instruction datasets, whereas for retrieval captioning datasets are repurposed for image retrieval.

Stage 3: Domain-specific Training. The third stage focuses on domain specialization using datasets with in-domain dialogues or instructions to start to expose the model to the target domain. For video datasets, such as YouCook2 (Zhou et al., 2018) and CrossTask (Zhukov et al., 2019), we use instructional video annotations to construct instructional training samples where the model needs to identify the plan step to which 4 provided frames belong. For retrieval, the model is trained to retrieve the relevant video moment provided a plan step and title. The FoodDialogues (Yin et al., 2023) and FoodReasonSeg (Yin et al., 2023) datasets are used for their dialogues with domain-specific knowledge.

Stage 4: Task-specific Training. The final stage trains the model for conversational instructional plan guidance, exposing the model to the domain-specific conversational patterns, CVMR, and VSG. The visual encoder $ve(\cdot)$ is frozen to avoid inserting a plan-specific bias that hinders the model's generalization capabilities, as the tasks are limited. Here, we train solely on InstructionVidDial, described in Section 4.4.

4.4 InstructionVidDial Dataset

One of our core contributions is the introduction of a conversational plan-guiding dataset that extends TastyVidDial (Glória-Silva et al., 2024b), including both cooking and DIY plans, and adding plan-grounded visual question answering. Thus greatly widening the dataset scope and allowing for a more comprehensive evaluation on instructional plan guidance. Below we outline the details of each step of the dataset construction process:

DIY Plans. To improve plan diversity and promote more generalizable learning, we leverage the COIN (Tang et al., 2019) dataset to include DIY instructional plan dialogues. To do this, for each type of task in COIN, we randomly select 4 videos and then replicate the dialogue generation process from TastyVidDial. As COIN videos are not paired with instructional plans, we produce a step-by-step plan from the textual video annotations by prompting Claude 3.5 [3] to rewrite the annotation sequence into a instructional plan. This results in 700 additional dialogues.

InstructionVidDial Curation. TastyVidDial contains 1.5k illustrated recipes, with a total of 50k generated dialogues, with an average of 33 dialogues per recipe. This led to redundancy within the dataset, as many of the dialogues share the same plan. To mitigate this, we trim the dataset keeping only the 4 dialogues per plan that contained the most multimodal turns.

Visual Question Answering Augmentation. To introduce pVQA requests into the already generated dialogues, we navigate the dialogues and introduce a 30% chance of inserting a pVQA request after each time the user progresses to a new action in the plan. For each, we select the middle frame of the relevant video moment to mimic a user-uploaded image, and prompt Claude 3.5 to generate a (question, answer) pair based on the image, the provided dialogue context, and the instructional plan. This results in dialogues where pVQA requests are naturally woven into the conversation (Appendix F).

5 Experimental Setup

5.1 Multitask Training Details

As VIGiA is trained in a multitask setup, we use a custom trainer, where each minibatch is sampled from multiple dataloaders to ensure a balanced task mix. Training batches are limited, per epoch, to the size of the smaller dataloader, ensuring a 1:1 sample ratio. As retrieval samples tend to have less tokens, task-specific batch sizes are used, which is

[2] For more details on each stage see Appendix A

[3] https://www.anthropic.com/news/claude-3-5-sonnet

beneficial for the InfoNCE retrieval loss as it uses as in-batch negatives. For retrieval, we set $d = 512$, and $N = 5$. [4]

5.2 Datasets

During our multistage training, the training data size progressively decreases at each stage as specialization occurs. For stage 4, InstructionVidDial has 6760 dialogues, which leads to a total of $\approx 114k$ dialogue turns, with a mix of multimodal requests (pVQA, VSG, CVMR) and textual plan-grounded answer generation. During training, we consider the 4 previous turns, and split the dataset in 90/5/5 for train/dev/test, with no plan overlap between splits. To reduce the frame count, for retrieval, only 1 in every 20 frames of the instructional video is kept. [4]

5.3 LM & Vision Backbones

For the LM we adopt LLaMa 3 8B (Dubey et al., 2024), but we compare different model sizes in Appendix D For the VE, following previous findings (Tong et al., 2024; Zohar et al., 2024; Laurençon et al., 2024b,a), we use SigLIP SO400M (Zhai et al., 2023) as the VE with an image size of 224x224 and a patch size of 14, for a total of 257 patches per image.

5.4 Metrics

For **PGAG** we use text generation metrics: METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020). For **VSG** we use ROUGE-L and Exact Match, where a match is if the generated text contains the exact step text. For **pVQA** we complement ROUGE-L with Accuracy as measured by the majority vote of 3 LLMs (Claude 3.5 Sonnet (Anthropic, 2024), LLaMa Vision 90B (Dubey et al., 2024), and Mistral Large (Team, 2024)), which are prompted with the image, dialogue context, instructional plan, and the system response, and asked to evaluate the accuracy of the response, more details in Appendix G. For **CVMR**, we use Recall@k, where $k \in \{1, 5\}$ with Intersection over Union (IoU) with $m \in \{0.5, 0.7\}$.

5.5 Baselines

Due to the limited amount of domain-specific models and models capable of performing CVMR, we complement our baseline models with general-purpose models. Specifically, we consider the following baselines:

MM-PlanLLM, the current in-domain SotA model. A 7B model capable of generating plan-grounded responses and retrieving video moments, but with no pVQA training. For CVMR, as it only retrieves the middle frame, we leverage their observation that the produced retrieval embedding has high similarity with the target video moment frames so we obtain candidate video moments by considering all frames adjacent to the retrieved frame with a similarity greater than 0.3.

TRACE (Guo et al., 2025), a VMR-specific model that has strong performance in the QVHighlights (Lei et al., 2021) dataset. This is a generative model, so we can only report Recall@1 results.

We also consider four recent ≈7B general-purpose baselines that have shown strong performance across multiple computer vision tasks and domains: **LLaVA-1.5** (Liu et al., 2024a), **Idefics2** (Laurençon et al., 2024b), **LLaVA-OV** (Li et al., 2024), and **Qwen2.5VL** (Bai et al., 2025).

5.6 Dataset Quality Annotation

As we add pVQA turns to the InstructionVidDial-dataset, we perform data annotation using volunteer annotators to obtain a quality estimation. In this annotation, we asked five annotators to, once shown the complete plan, the reference image, the question, and the answer, rate (from 1 to 3) the turn on three independent measures: **Relevance** of the question to both the image and instructional plan, **Plausibility** of the question in a real setting, and **Accuracy** of the provided answer. This annotation was carried out using 250 pVQA turns.

Our results show that the generated turns are composed of very relevant questions and accurate answers with a score of 2.67 and 2.75 respectively. Plausibility was the lowest rated measure at 2.22, although still strong. Based on annotator feedback and sample observation, we determine that the lower Plausibility score is linked to turns where, occasionally, the question is phrased as if the system is the one providing the image, not the user. [5]

6 Results and Discussion

In this section, we present and discuss the results of VIGiA on the various tasks considered. We start by evaluating on in-domain tasks using the InstructionVidDial test set. This is followed by an evaluation of general-purpose image understanding using MME.

[4] Appendix B includes more details on the training setup.

[5] Additional details are provided in Appendix C.

Table 1: Results on the in-domain tasks: PGAG, pVQA, and VSG. Underlined scores are the best among the baselines. *Accuracy is measured using the majority vote of three independent LLMs (See Section 5).

	Plan Reasoning					Plan-based Retrieval	
	PGAG			pVQA		VSG	
Model	*ROUGE-L*	*METEOR*	*BERTScore*	*ROUGE-L*	*Accuracy**	*ROUGE-L*	*EM*
Idefics2	37.16	45.09	67.55	21.22	47.48	20.16	5.85
LLaVA-1.5	42.47	40.84	70.55	26.77	71.03	16.07	4.21
Qwen 2.5 VL	31.61	41.51	63.91	26.39	81.68	16.65	9.47
LLaVA-OV	41.18	40.85	70.70	27.02	81.28	32.28	17.89
MM-PlanLLM	58.85	59.34	80.03	7.49	0.37	44.70	40.00
VIGiA	**75.30**	**76.67**	**88.72**	**33.65**	**90.28**	**55.66**	**57.37**

Table 2: Results on the CVMR task.

	Model	CVMR			
		Recall@1		Recall@5	
		m=0.5	*m=0.7*	*m=0.5*	*m=0.7*
firm	TRACE	5.97	1.63	—	—
	VIGiA	10.49	2.89	30.38	11.75
adj.	MM-PlanLLM	25.32	12.48	31.46	16.46
	VIGiA	**30.74**	**15.37**	**51.18**	**26.94**

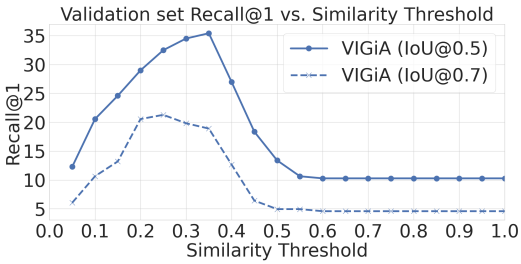


Figure 4: Comparing R@1 performance with varying values for the similarity threshold on InstructionVidDial's dev set.

6.1 Textual Plan-Grounded Answer Generation

Textual Plan-Grounded Answer Generation covers the majority of the requests seen in the InstructionVidDial dataset, and is a good benchmark of the model's ability to generate coherent and contextually relevant responses based on the instructional plan dialogue. The evaluation of this task has all text-only requests from the test set, which includes a mix of questions, navigation requests, and more. The results are summarized in Table 1.

These results show that VIGiA outperforms all baselines and also reinforces the need for domain-specific training for instructional plans, as MM-PlanLLM, the only in-domain baseline, outperforms general purpose baselines.

6.2 Multimodal Answer Generation

Regarding multimodal answer generation, we test the model on tasks where direct image understanding is needed to respond accurately. Specifically, we evaluate VIGiA's performance, against the baselines, on the VSG and pVQA turns of the test set and present the results in Table 1. The **pVQA** results attest the benefit of in-domain pVQA training as VIGiA has a 90.28% accuracy, a meaningful leap over Qwen with 81.68, while MM-PlanLLM, which was not trained for any type of VQA, has near-zero performance, highlighting the critical gap that we are the first to address. **VSG** results are similar with VIGiA strongly outperforming all baselines. General purpose baselines struggle as this task demands aligning the provided image with the textual plan, requiring a strong understanding of instructional plan structure on both modalities. We also hypothesize that the inclusion of in-domain pVQA in the training data elicits richer plan understanding as we outperform MM-PlanLLM by over 25% despite training on 5x less in-domain data.

6.3 Conversational Video Moment Retrieval

The **CVMR** task requires retrieving a specific video moment from an instructional video, based on the dialogue context. VIGiA produces a start and end frame embedding thus it can be used for CVMR in two distinct ways: a *"firm"* setting where candidate video moments are determined by the most similar frames to the `[RETs]` and `[RETe]` token embeddings, and an *"adjusted"* setting where candidate video moments are retrieved by finding the start and end frame where the similarity with the produced `[RETs]` and `[RETe]` tokens representation drops below a certain threshold, respectively. The *"adjusted"* setting leverages the model's ability to retrieve the start and end regions, while also allowing for flexibility in the candidate moment's limits. In Figure 4 we compare different thresholds on the dev set and see that a 0.35 seems to be optimal for

Table 3: Ablation of model training stages. Underlined scores represent improvements over the previous stage/row.

		Plan Reasoning				Plan-based Retrieval				General	
		PGAG		pVQA		VSG		CVMR R@1		MME	
		R-L	*METEOR*	*R-L*	*Acc.*	*R-L*	*EM*	*m=0.5*	*m=0.7*	*Cog.*	*Per.*
Stages	4 Only	74.75	76.05	33.56	89.16	54.27	56.84	11.39	5.97	**234**	503
	1+4	74.67	76.61	33.89	**91.59**	54.17	56.32	20.43	11.39	211	501
	1+2+4	74.99	76.39	**34.20**	89.35	54.12	56.84	9.58	5.06	223	708
	1+2+3+4	**75.30**	**76.67**	33.65	90.28	**55.66**	**57.37**	**30.74**	**15.37**	229	**717**

Table 4: MME performance across stages.

Model	Stage 2		Stage 4 / Final	
	Cog.	*Per.*	*Cog.*	*Per.*
MM-PlanLLM	—	—	34	1
Idefics2	—	—	328	1474
LLaVA-1.5	—	—	314	1483
LLaVA-OV	—	—	418	1580
Qwen 2.5 VL	—	—	**638**	**1685**
VIGiA	263	1344	229	717

recall performance.

Table 2 shows the results of this evaluation. VIGiA is able to substantially improve on the previous SotA for this task, with a decent improvement over MM-PlanLLM. Comparing the *firm* and *adjusted* settings, we note that the latter effectively leverages the representations learned by our retrieval-specific layers and highlights the benefit of learning start/end region representations, with much higher R@1 scores. TRACE while a strong baseline for general VMR struggles in our domain-specific conversational setting showing that off-the-self models are not viable for this task.

6.4 General Task Performance

Domain-specific models still benefit from rich general image understanding to better address unseen plans and domains. We use MME (Fu et al., 2023), a benchmark of 14 image understanding sub-tasks, as a proxy to measure flexibility and the impact of domain-specific training.

From Table 4 we see the expected performance plot with general purpose models excelling. Regarding VIGiA, we confirm that Stage 2 training instills rich image understanding but, after Stage 4, there is a performance drop, a trade-off of as the model begins to be specialized. In general, VIGiA retains general image understanding capabilities even after instructional plan guidance specialization, particularly compared to MM-PlanLLM.

6.5 Ablation Studies

To evaluate and measure the impact of each training stage on the different tasks, we performed an ablation study by progressively adding additional training stages. Table 3 shows our ablation of multistage training. We see an incremental benefit of each added training stage. While the single-stage approach does present a strong baseline, it is lacking CVMR and MME performance, which are then addressed in the latter stages. The first and third stage collectively improve CVMR performance, up 20 points on R@1 with m=0.5. The 2nd stage improves model flexibility with an MME Perception increase of 200 points, without a significant loss on in-domain tasks. We also ablated RoPE usage for frame position encoding and found that not using it degraded performance, with R@1 m=0.5 dropping 13%, from 30.74 to 26.76 and R@1 m=0.7 dropping 29% from 15.37 to 11.03.

7 Conclusion

This paper proposes VIGiA, the first LVLM model for conversational instructional plan guidance, that jointly addresses the key goals of the task: 1) rich image understanding for Plan-grounded pVQA, 2) conversational video moment retrieval, 3) multimodal plan alignment and 4) plan-grounded answer generation. An extended cross-modal architecture is designed, supporting unified plan-grounded dialogue and generative video moment retrieval.

In our detailed evaluation, we compare VIGiA against both domain-specific and general-purpose LVLMs and find that it consistently and significantly outperforms both types of baselines, with over 90% accuracy in plan-grounded in-context pVQA. We complement our analysis with an evaluation on a general image understanding benchmark and show that VIGiA remains far more capable than the in-domain state-of-the-art.

8 Limitations

The very limited amount of instructional plans paired with an annotated instructional video lead to a dataset with not enough training examples diverse enough to perform a full training, without needing auxiliary datasets. As presented in Table 5, VIGiA is trained with different datasets over different phases. These are intended to equip the model with increasingly complex capabilities, to mitigate the knowledge gaps of the latter stages.

The datasets and instructional plans used for domain-specific training (Stage 3 & 4) might have a western bias, as they were sourced from English sources. Because of this, we posit that for other cultural contexts model performance can vary as the model is not as familiar with region-specific terminology and techniques. However, we do not believe that core functionalities, such as navigation and context-dependent QA, are affected because these functionalities depend on task structure rather than external knowledge.

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, and 8 others. 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.

Anthropic. 2024. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. Accessed: 2025-04-03.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. 2024. LoRA learns less and forgets less. *Transactions on Machine Learning Research*. Featured Certification.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, volume 15075 of *Lecture Notes in Computer Science*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

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.

Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. 2024. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o.

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*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana

Ehsani, Huong Ngo, Yen-Sung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 32 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *CoRR*, abs/2409.17146.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Sophie Fischer, Carlos Gemmell, Niklas Tecklenburg, Iain Mackie, Federico Rossetto, and Jeffrey Dalton. 2024. Grillbot in practice: Lessons and tradeoffs deploying large language models for adaptable conversational task assistants. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 4951–4961. ACM.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and Joao Magalhaes. 2024a. 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.

Diogo Glória-Silva, David Semedo, and Joao Magalhaes. 2024b. Show and guide: Instructional-plan grounded vision and language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21371–21389, Miami, Florida, USA. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. 2025. TRACE: Temporal grounding video LLM via causal event modeling. In *The Thirteenth International Conference on Learning Representations*.

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*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

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.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *CoRR*, abs/2408.12637.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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 *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual 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.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, and 8 others. 2024b. Nvila: Efficient frontier visual language models. *Preprint*, arXiv:2412.04468.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. RSVQA: visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote. Sens.*, 58(12):8555–8566.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Lingbo Mo, Shijie Chen, Ziru Chen, Xiang Deng, Ashley Lewis, Sunit Singh, Samuel Stevens, Chang-You Tai, Zhen Wang, Xiang Yue, Tianshu Zhang, Yu Su, and Huan Sun. 2023. Roll up your sleeves: Working with a collaborative and engaging task-oriented dialogue system. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 197–201, Prague, Czechia. Association for Computational Linguistics.

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.

F. Ren and Yangyang Zhou. 2020. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

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.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13088–13110. Association for Computational Linguistics.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1207–1216. Computer Vision Foundation / IEEE.

Makarand Tapaswi, Yukun Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. 2015. Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, and et al. 2024. Gemma.

Mistral AI Team. 2024. Au large. Mistral AI Blog Post. Accessed: 2025-04-03.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356. Curran Associates, Inc.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2023. Foodlmm: A versatile food assistant using large multi-modal model. *CoRR*, abs/2312.14991.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

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

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. LLaVAR: Enhanced visual instruction tuning for Text-Rich image understanding.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press.

Qingxiaoyang Zhu, Yi-Chieh Lee, and Hao-Chuan Wang. 2022. Action-a-bot: Exploring human-chatbot conversations for actionable instruction giving and following. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'22 Companion, page 145–149, New York, NY, USA. Association for Computing Machinery.

Qingxiaoyang Zhu, Yi-Chieh Lee, and Hao-Chuan Wang. 2025. Actionabot: Structuring metacognitive conversations towards in-situ awareness in how-to instruction following. In *International Conference on Conversational User Interfaces*.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David F. Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3537–3545. Computer Vision Foundation / IEEE.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. 2024. Apollo: An exploration of video understanding in large multimodal models. *CoRR*, abs/2412.10360.

A Training Dataset Details

In this section, we provide additional details on the datasets used to train VIGiA. We include the specific number of samples used in each stage, as shown in Table 6.

A.1 Multitask Trainer

As there are 2 losses being optimized in our model, passing through task-specific components, we need to employ a dual dataloader trainer that allows us to train on both tasks at the same time without needing to format our batches in a multitask format, something that is not doable for pVQA, Visual Instruction Tuning, and CMVR. This is achieved by alternating the dataloaders when sampling batches, allowing for an even distribution of training batches.

A.2 Dataset Transformations

As we target instructional plans in a conversational manner, we needed to adapt and transform some of the datasets to train our model. In this section, we detail the transformations applied to each dataset:

LAION400M - For LAION400M we use the image-caption pairs to train the model in a Image-to-Text task and Text-to-Image retrieval.

ShareGPT4o - This dataset is used as is, with no alterations.

MAmmoTH-VL - This dataset is used as is, with no alterations. We only consider the first 1M multimodal samples.

VQAv2 & GQA - These datasets are framed as an instruction following task using templates to convert the question into an instruction, the answer is kept as is.

ShareGPT4v, LLaVA-Pretrain, Pixmo-cap - These datasets are used as is, with no alterations. Pixmo is framed as a user request for a detailed image caption.

FoodDialogues - This dataset includes special tags, such as `[CARB]`, to indicate nutritional info estimation in system answers, we remove these tags and preserve the rest of the text.

FoodReasonSeg - This dataset includes requests for segmentation mask along with ingredient identification, we focus on the latter and remove the segmentation mask requests. We also rewrite the user utterances using a Llama 3 model, to improve naturalness and fluency.

YouCook2, Crosstask, COIN - These datasets are made up of instructional videos with annotated video segments. To adapt them to out needs, for each annotated video segment, we extract 4 frames and train the model to predict the annotated text. For retrieval, we instead provide the model with the task name and annotated video segment text and train the model to retrieve the start and end frames of the video segment. In this case we adopt the strategy of picking N frames from the start and end, and the model is trained to approximate the average of the embeddings of these frames. This allows us to train the model to retrieve a single frame, while also allowing for some flexibility in the candidate video moment limits.

B Training Details

In this section we provide additional details on the training setup and hyperparameters used to train VIGiA. The hyperparameters are shared across all models and are shown in Table 7, for each training stage. The training was performed using NVIDIA A100 GPUs with 80GB of memory each, connected with SXM4.

On stage 2 and 3 we leverage our dual dataloader training approach to utilize larger batch sizes for retrieval, this was possible as the retrieval samples are smaller than the captioning samples. This approach benefits retrieval training as we utilize in-batch samples as negatives for the contrastive loss calculation.

For stage 4 we use a single dataloader with a batch size of 32, and the different mode samples are uniformly sampled from the dataset, with this sampling logic being handled by the sampler instead.

In Table 7, CAP refers to the connector between the Vision Encoder and the LLM, RET refers to the retrieval-specific layers used in the model ($W_{start}, W_{end}, W_{ret}$).

C pVQA Data Quality Annotation

As LLMs were used to generate the pVQA turns in InstructionVidDial we find it important to conduct a data quality annotation study to ascertain the quality of the added turns on several metrics. To this end, we conduct an annotation study with five volunteer annotators and tasked them with annotating the following 3 criteria:

- **Plausibility** - Here we asked annotators to rate how plausible each turn is when considering the provided dialogue context in a real scenario. With this criteria, we aimed to measure how realistic the generated turns are and how well they pass for real user inputs.

Table 5: Overview of the multistage training process for VIGiA. Status indicators: 🔥= Trainable, ❄️= Frozen.

Stage	Specialized Layers	LLM	VE	Datasets
Initialization	🔥	❄️	❄️	Laion400M (Schuhmann et al., 2021)
Visual Instruction Tuning	🔥	🔥	🔥	ShareGPT4o (Cui et al., 2024), MAmmoTH-VL (Guo et al., 2024), VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), ShareGPT4v (Chen et al., 2024a), PixMo-Cap (Deitke et al., 2024), LLaVA-Pretrain (Liu et al., 2023)
Domain-specific Tuning	🔥	🔥	❄️	FoodDialogues (Yin et al., 2023), FoodReasonSeg (Yin et al., 2023), Youcook2 (Zhou et al., 2018), COIN (Tang et al., 2019), Crosstask (Zhukov et al., 2019)
Task-specific Training	🔥	🔥	❄️	InstructionVidDial (ours)

Table 6: Overview of the datasets used to train VIGiA.

Stage	Dataset	Type	Samples #
1	LAION400M	Cap + Ret	10M+10M
2	ShareGPT4o	Cap	≈57k
	MAmmoTH-VL	Reasoning	1M
	VQAv2	pVQA	≈444k
	GQA	pVQA	943k
	ShareGPT4v	Ret	≈1.25M
	Pixmo-cap	Ret	≈698k
	LLaVA-Pretrain	Ret	≈558k
3	FoodDialogues	Domain	≈26k
	FoodReasonSeg	Domain	≈14k
	YouCook2	Cap+Ret	≈10k+≈10k
	Crosstask	Cap+Ret	≈17k+≈16k
	COIN	Cap+Ret	≈43k+≈39k
4	InstructionVidDial	All	≈110k

- **Relevance** - For this criteria, we wanted to assess how relevant the generate user visual question was to both the selected image, and the instructional plan being executed. This is critical to understand if the generated turns are closely related to all of the surrounding context.

- **Accuracy** - It was also important to ensure that the generated turns were made up of questions that were adequately answered and did not contain inaccurate answers instead. In this criteria, we asked annotators to, considering the dialogue, instructional plan, and their own knowledge, rate if the provided answers were accurate to the generated user questions.

The selected annotators were volunteers all with higher education in the computer science field, all fluent in English, and familiar with data annotation tasks and user studies. In total 250 pVQA samples were randomly selected from InstructionVidDial's

Table 7: Shared hyperparameters used to train all models. For Stage 1, LLM-specific hyperparameters only affect the `[RET]` tokens input and output embeddings, along with the language modeling head.

Stage	1	2	3	4
Batch Size	512	64	128	32
RET Batch Size	512	96	192	—
Acc. Steps	4/4/8	4/8/8	4/8/8	4/8/16
Train Steps	19500	15000	866	3252
Parallel	DDP	DDP	DDP	DDP
GPU #	4	8	4	1
Model DType	BF16	BF16	BF16	BF16
LLM LR	$1*10^{-3}$	$5*10^{-5}$	$2*10^{-6}$	$5*10^{-5}$
VE LR	—	$5*10^{-5}$	—	—
CAP LR	$5*10^{-4}$	$5*10^{-4}$	$5*10^{-4}$	$2*10^{-4}$
RET LR	$1*10^{-3}$	$1*10^{-3}$	$1*10^{-3}$	$1*10^{-3}$
Scheduler	Constant	Constant	Constant	Constant
Optimizer	AdamW	AdamW	AdamW	AdamW
LLM Decay	0.03	0.01	0.0	0.0
VE Decay	—	0.0	—	—
CAP Decay	0.03	0.03	0.01	0.01
RET Decay	0.03	0.03	0.01	0.01
LoRA (Only for 8B)				
LoRa DType	—	FP32	FP32	FP32
LoRa Rank	—	32	32	32
LoRa α	—	128	128	128
LoRa Dropout	—	0.03	0.03	0.03

Table 8: Results of the data quality annotation conducted on 250 pVQA turns from the test set of InstructionVidDial.

Plausibility	Relevance	Accuracy
2.22	2.67	2.75

test set. Annotators were shown the instructional plan, provided user image, generated user question, and generated system response. The specific instructions provided to the volunteers are shown in Table 9.

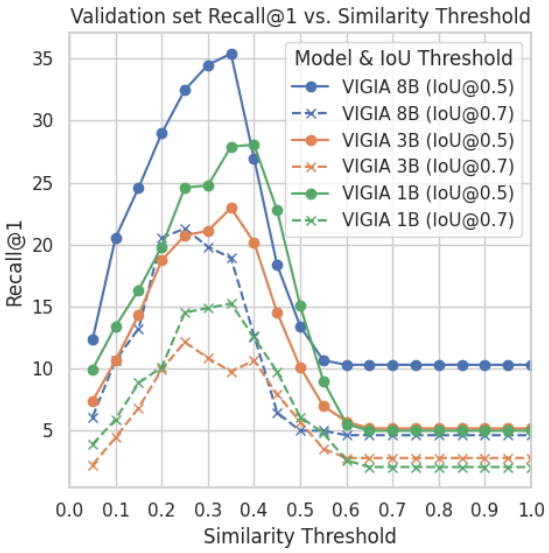
The results, shown in Table 8, highlight the quality of the generated pVQA turns with very high scores for both Relevance and Accuracy, meaning that the turns were closely related to the context they were inserted in and provided reliable answers. Plausibility scores were also very positive asserting that the generated user questions often fell into the scope of possible questions a user might ask in a realistic setting, a key factor to ensure high data quality.

D LM Backbone Size Variance

To measure how model performance varies across LM backbone sizes, we ran some experiments with other Llama 3 model sizes. Specifically, we tested with Llama 3 1B and 3B, the results for all tasks are shown in Table 10.

Here we see that, with some exceptions, the smaller models have very strong performances, considering their reduced size, whilst also being clear the benefit of increase LM backbone size. For **CVMR** VIGiA 1B outperforms 3B, we hypothesize that it is caused by 3B training with a smaller batch size, which slows down convergence as InfoNCE loss uses in-batch negatives, meaning that smaller batches can slow down convergence. Focusing on **pVQA** the benefit of higher parameter count is very accentuated with significant improvements as we scale up the LM. We believe that as this is a task that requires the models to encode external knowledge acquired during training, smaller models have more limited external knowledge remembrance capacity. Whereas for **PGAG**, where most requests are context-dependent, the models all have impressively similar performance. **MME** emerges as an outlier, having similar performance to the 3B variant, despite the increased LLM size. We hypothesize that this is caused by training the 8B LM backbone using LoRA (Hu et al., 2022), due to memory constraints, from stage 2 onward, which means that only a small number of parameters are being updated, thus suffering from decreased learning capacity, as studied in Biderman et al. (2024).

Figure 5: Recall@1 performance of VIGiA on the CVMR task with varying values for the similarity threshold used to define the limits of the candidate video moments.



E CVMR Evaluation

While the task of CVMR is formulated by retrieving the start and end frame of a relevant video moment, realizing the harshness of such an approach, specifically in a setting where adjacent images are largely similar, we train our model to instead produce the representation of the averaged N frames around the target frame. Based on this, it is fair to say that our model is trained to predict a region representation rather than only retrieving a single frame.

To account for this, during evaluation, we evaluate our models and MM-PlanLLM in what we call an "adjusted" CVMR evaluation setting where the limits of the candidate video moment are determined by the point/frame where the frame's similarity to the produced representation drops below a given threshold. This leverages the observations of Glória-Silva et al. (2024b) that the model's retrieval features show a decrease in similarity to steps further away from the target step, suggesting that the model is able to effectively identify the target plan step and produce an adequate video region representation. This threshold is determined using the validation set, as seen in Figure 5. Here we see that there is a clear peak where performance is maximized that is followed by a significant performance drop, confirming that the produced representations are effectively aligned with the target visual plan step. Based on this, we set the threshold to 0.35,

where most models are at or close to their performance peak.

F pVQA Generation Details

Of the two crucial extensions that we make to TastyVidDial is the addition of pVQA user requests. To achieve this, we used Claude 3.5 Sonnetv2 to generate user questions relevant to a provided image, the ongoing dialogue, and the instructional plan being executed. To determine whether a pVQA turn would be added we set a 30% chance of that after any time the user moved forward in the task (using the NextStepIntent annotation provided in TastyVidDial). Table 17 shows the prompt used to generate the visual question-answer pairs, and Table 11 shows an example from InstructionVidDial's test set.

G pVQA Evaluation Details

For the evaluation of pVQA turns on the InstructionVidDial dataset, we utilize accuracy as measured by three LLMs. For this evaluation, we consider three large-scale LVLMs to maximize the quality of the annotations. Specifically, we utilize Claude 3.5 Sonnet v2, LLaMa Vision Instruct 90B, and Mistral Large.

For each model, we provided it with the relevant image, the previous 4 dialogue turns, and the instructional plan being followed. The prompt structure is present in Table 18. The majority vote of the three responses is then used to provide the final accuracy score.

H Qualitative Examples

In Tables 12, 13, and 14 we share some examples of generated answers to plan-grounded pVQA turns from InstructionVidDial test set. In Tables 15 and 16 we compare how different model sizes answer the same user questions, highligthing their capabilities and limitations.

Table 9: Annotation instructions provided to annotate in-domain pVQA turns.

```
Consider a dataset of dialogues
between users and a system designed
to assist with manual tasks (recipes
or DIY projects). The objective of
these annotations is to quantify the
quality of automatically generated
VQA (Visual Question Answering) turns
within that dataset.

Each line contains an example of a
question-answer pair, along with the
image that originates the question and
the task to be performed.
For each line, on a scale of 1 to 3
(from worst to best), annotate the
following 3 criteria:

Relevance: The relevance of the
question concerning the image and
the task in question.
1 - The question has nothing to do
with either the image or the task.
3 - The question is relevant to both.

Plausibility: How plausible the
question is, considering a person
trying to perform the task.
1 - A question that no one would ask
in the context of that task.
3 - A completely plausible question to
be asked during the execution of the
task.

Accuracy: Based on the task and your
own knowledge, how correct the given
answer to the question is.
1 - An incorrect answer.
3 - A correct answer.

Consider each criterion separately. A
question may not be at all relevant or
plausible (in the context of the task)
but still have a correct answer.
```

Table 10: Ablation of LM backbone sizes on model performance across all considered tasks.

	Plan Reasoning				**Plan-based Retrieval**				**General**	
	PGAG		**pVQA**		**VSG**		**CVMR R@1**		**MME**	
	R-L	*METEOR*	*R-L*	*Acc.*	*R-L*	*EM*	*m=0.5*	*m=0.7*	*Cog.*	*Per.*
VIGiA 1B	75.44	76.67	29.35	59.81	51.53	50.53	26.22	13.92	203	632
VIGiA 3B	**76.51**	**77.72**	30.16	73.80	52.15	53.68	23.33	**15.91**	214	**808**
VIGiA 8B	75.30	76.67	**33.65**	**90.28**	**55.66**	**57.37**	**30.74**	15.37	**229**	717

Table 11: pVQA test set examples

User Image		
	Question:	I'm adding the pasta to the boiling milk mixture now - is it normal that the liquid looks this white and creamy already, or did I add too much milk?
	GT Answer:	The liquid consistency shown in the image looks perfect! The white, creamy appearance is exactly what you want at this stage. The milk should be bubbling and have this rich, creamy texture before adding the farfalle (bow-tie) pasta. Just remember to keep stirring frequently while the pasta cooks for about 10 minutes at medium heat to prevent it from sticking to the bottom of the pot and to ensure even cooking.
	Question:	Based on the color of the chicken pieces in the pan, are they browned and cooked through enough to be removed from the pan?
	GT Answer:	Yes, from what I can see in the image, the chicken pieces appear to have a nice golden-brown color and look properly cooked through. You can see they've lost their raw pink color and have taken on a light golden hue. They're being removed from the pan with a black spoon, which is exactly what you should be doing at this step. The bits of browned fond left in the pan will add great flavor to the next steps of the recipe.
	Question:	Is the green and black cap shown in the image the correct type of cap for sealing a volumetric flask?
	GT Answer:	Yes, that appears to be the correct type of cap for a volumetric flask. The cap shown in the image has a green band and black top which is a standard laboratory flask cap design. It's designed to create an airtight seal when properly attached to the volumetric flask. Make sure to screw it on firmly but not too tight to avoid damaging the threads.

Table 12: Examples of answers provided by VIGiA and some baselines to pVQA samples. In this example LLaVA-OV provided a very dangerous suggestion to the user, whereas both VIGiA and Qwen2.5 VL were able to provide adequate and safe answers.

User Image	
Question:	I've added the oil and garlic to the pan as shown in the image, but how do I know if the heat is at the right temperature? I can see some garlic starting to sizzle but I don't want to burn it.
VIGIA 8B:	The heat appears to be at a good medium temperature as shown in the image. You can see the garlic is starting to sizzle and release its oil, which is the right reaction. If you're seeing this gentle sizzling and the garlic isn't browning too quickly, the heat is appropriate. Just keep an eye on it and stir occasionally to prevent burning. The garlic should be fragrant but not browned when you move on to the next step.
Qwen2.5VL:	To ensure the heat is at the right temperature, you can use a few methods: 1. **Sizzling Garlic:** As you've noticed, the garlic is sizzling in the pan. This is a good sign because it indicates the oil is hot enough to start cooking the garlic without burning it. However, if the sizzling becomes too intense or the garlic starts to brown quickly, reduce the heat slightly. 2. **Visual Check:** If you can see the oil moving around the pan and the garlic pieces are moving as well, it's likely the oil is at a good temperature.
LLaVA-OV:	To check if the heat is at the right temperature, you can use the "tongue test." Place the tip of your tongue close to the pan, but not touching it. If it feels warm but not uncomfortable, the heat is likely medium. If it feels too hot, reduce the heat. If it feels cool, increase the heat. Remember, it's better to start with a lower heat and adjust as needed to prevent burning.

Table 13: Examples of answers provided by VIGiA and some baselines to pVQA samples. In this example, all three models provided adequate answers but general-purpose models tend to be more verbose.

User Image	
Question:	How do I know if I've rolled the pastry out to the correct dimensions? Should I measure it or is eyeballing it okay?
VIGIA 8B:	For best results, you should measure your rolled pastry to ensure it's exactly 24 inches long and 8 inches wide. This is crucial because puff pastry needs to be rolled to the correct size to properly fold and layer. While eyeballing might work for some recipes, for puff pastry, precise measurements are essential to achieve the right texture and layers. You can use a measuring tape or ruler to check the dimensions while rolling.
Qwen2.5VL:	When rolling out the puff pastry, it's important to ensure that it reaches the desired dimensions to achieve the best results. Here's how you can check: ### Measuring: - **Use a Ruler:** Place a ruler next to the pastry and measure the length and width. This ensures accuracy and consistency. - **Mark the Dimensions:** If you're using a recipe, mark the dimensions on the pastry sheet with a pen or a small piece of paper to guide your rolling. ### Eyeballing: - **Experience:** If you have experience with baking and are comfortable with your eye, you might be able to eyeball it.
LLaVA-OV:	When rolling out puff pastry, it's important to ensure that the dimensions are correct for the recipe you're following. Here are a few ways to check if you've rolled the pastry out to the correct dimensions: 1. **Measurement**: Yes, you should measure the pastry. Use a ruler or a measuring tape to ensure that the length and width of the rolled-out pastry are as specified in the recipe. This is crucial for puff pastry, as the thickness and size can affect the final texture and rise of the pastry. 2. **Visual Inspection**: While it's not as accurate as measuring, you can also visually inspect the pastry.

Table 14: Examples of answers provided by VIGiA and some baselines to pVQA samples. In this example, all three models provided adequate answers but general-purpose models tend to be more verbose.

User Image	
Question:	I see the balance pointer is not quite centered at zero - should I adjust the calibration knob before weighing anything?
VIGIA 8B:	Yes, you should definitely calibrate the balance first. The image shows the balance is currently not zeroed, which means you'll need to adjust the calibration knob to ensure the indicator needle is centered at zero before proceeding to weigh any objects. This is a crucial first step to ensure accurate measurements throughout the process.
Qwen2.5VL:	Yes, it's important to ensure that the balance is properly calibrated before using it for any measurements. If the pointer is not centered at zero, it means the balance is not calibrated correctly, which could lead to inaccurate readings. Here's what you should do: 1. **Locate the Calibration Knob**: Typically, there will be a small knob or screw near the base of the balance, often labeled "Calibration" or something similar. This is used to adjust the balance. 2. **Adjust the Calibration Knob**: Turn the knob slowly in the direction needed to bring the pointer back to the zero mark. Usually, turning clockwise
LLaVA-OV:	Yes, you should adjust the calibration knob to ensure the balance is properly calibrated and reading zero before weighing anything. This is a crucial step to ensure the accuracy of the weight measurement.

Table 15: Examples of answers provided by VIGiA and its variants. All 3 model sizes available are capable of providing an accurate and coherent answer to the user question, while also providing a brief explanation of the importance of the thickness of the cinnamon rolls.

User Image	
Question:	Are my cinnamon roll discs being cut at the right thickness based on what's shown in the image?
VIGIA 8B:	Looking at the image, the cinnamon roll discs appear to be cut at a good thickness. You can see they're being cut into uniform, bite-sized pieces that are approximately 1-2 inches in diameter. This size is ideal for mini cinnamon roll bites as it ensures they'll cook evenly and be easy to eat. Just make sure to keep an eye on them while baking as the thickness may vary slightly depending on how you cut them.
VIGIA 3B:	From what I can see in the image, the thickness appears to be appropriate. The discs are being cut into small, uniform pieces that are about 1-2 inches in diameter, which is ideal for mini cinnamon roll bites. This size will allow them to cook evenly and be easy to handle. When baked, these will become crispy on the outside while remaining soft on the inside.
VIGIA 1B:	Yes, the thickness shown in the image appears to be appropriate for the recipe. The discs should be cut to be about 1/4 inch (6mm) thick, which is ideal for baking. If you cut them too thick, they might burn easily, and if too thin, they might burn too quickly. The uniform, flat shape shown in the image is also a good sign that you're working on the right material.

Table 16: Examples of answers provided by VIGiA and its variants. The smaller model variants fail to accurately answer the user question, claiming that only a jar lid can be used whereas VIGIA 8B accurately clarifies that other objects can be used as long as they are dry and clean.

User Image	
Question:	The pie crust in the image looks quite round - is a jar lid the only tool I can use to cut these circles, or can I use something else of similar size?
VIGIA 8B:	While a jar lid is the traditional tool for cutting pie crust circles, you can use other round objects of similar size as a substitute. Some alternatives include a clean, dry glass or a small bowl. Just make sure the object you use is completely dry and clean to prevent sticking. The most important thing is to have a consistent size for your circles so they'll bake evenly.
VIGIA 3B:	A jar lid is the only tool that will work for cutting the circular pie crust. The size of the jar lid (4 inches) is perfect for creating uniform, bite-sized pieces. Using a different tool could lead to unevenly sized pieces, which would be difficult to work with. Make sure you're using a clean, sharp jar lid and press it into the center of the pie crust to ensure a clean cut.
VIGIA 1B:	A jar lid is the correct tool for cutting the 12 4-inch (10 cm) circles. While the image shows a circular shape, a jar lid is specifically designed for this purpose. Using a jar lid will give you the most precise and uniform cuts needed for the pie crust. Make sure to press firmly into the crust to ensure the circles are firmly attached. The circular shape will help the crust hold its shape when baked.

Table 17: Prompt used to generate the in-domain pVQA question-answer pairs.

```
You are a helpful assistant simulating
a user interacting with a visual task
guidance system. You will be provided
with a task, a conversation history,
the current step's description, and an
image related to that step.

Based on the conversation context, the
current step, and the image, generate
one question-answer pair. The question
should mimic a question a user might
ask after uploading the provided
image while seeking guidance on the
current step, this is VERY important.
The questions should be DIRECTLY
related to the image and current
task step, and should be relevant
to the ongoing conversation. Make sure
that these questions sound natural
as if a real user had asked them. Use
the conversation history, the task
instructions and your knowledge to
answer as accurately as possible.
Some examples of questions you might
generate include asking if the
mixture in the image is the correct
consistency, if the oven temperature
is correct, or if the bolt in the
image needs to be tightened, etc.

Output the question-answer pair in the
following format:

Q: <Question>
A: <Answer>

---
Task:
{task_text}

Conversation History:
{conv_history}

---

Current Step: {step_number} -
{step_text}
Image:
```

Table 18: Prompt used to evaluate the in-domain pVQA responses.

```
You will be provided with an image
and a bit of a dialogue. Your task
is to tell me if the last system
response accurately answers the last
user question based on the image.
Answer solely regarding the LAST
user question and the last assistant
response in the dialogue.
As I am parsing this automatically,
please write your answer (YES or NO)
immediately after the words "FINAL
ANSWER:" in your response. Feel free
to think and reason out loud in your
response before you give the final
answer.

Task Context:

{task_text}

Dialogue Context (up to 4 turns):
{conv_history}

Image: (provided in the request)
```