Deduce and Select Evidences with Language Models for Training-Free Video Goal Inference

Yeo Keat Ee Hao Zhang Alexander Matyasko* Basura Fernando
Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore
Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore
1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore
{ee_yeo_keat, zhang_hao, fernando_basura}@cfar.a-star.edu.sg

Abstract

*We introduce ViDSE, a **Vi**deo framework that **D**educe and **S**elects visual **E**vidence for training-free video goal inference using language models. Unlike approaches that directly apply vision-language models (VLM) or combine VLM+LLM to process dense video visuals, ViDSE explicitly selects relevant visual evidence (e.g., frames) based on the hypothesis deduced by the LLM. This approach not only improves accuracy but also reveals the logical process behind the model's decisions, enhancing explainability. Our experiments demonstrate that this selection process significantly reduces ambiguity in the subsequent inference reasoning stage and outperforms VLM-only and VLM+LLM models on goal inference tasks such as CrossTask and COIN. We further validate ViDSE's generalizability and robustness on action recognition tasks, such as ActivityNet and UCF-101, under training-free and open-vocabulary conditions. We observe that ViDSE easily generalizes to other video tasks (e.g., action recognition) requiring filtering of redundant and irrelevant information.*

1. Introduction

Video understanding benefits from the "*Scaling Law*" [18], which suggests that progress can be achieved by increasing data scales [13, 16, 19], model complexities [3, 31, 44, 53], and computational resources [1]. The recent emergence of large language models (LLMs) and their visual extensions (VLMs) further verified the effectiveness of scaling law. These language-based models (LMs) [5, 25, 41, 51] have shown strong generalization abilities across various vision-language tasks. This has attracted a growing research interest in effectively leveraging the LMs to transfer knowledge to out-of-domain and novel tasks.

*This work was done while Matyasko was at A*STAR.

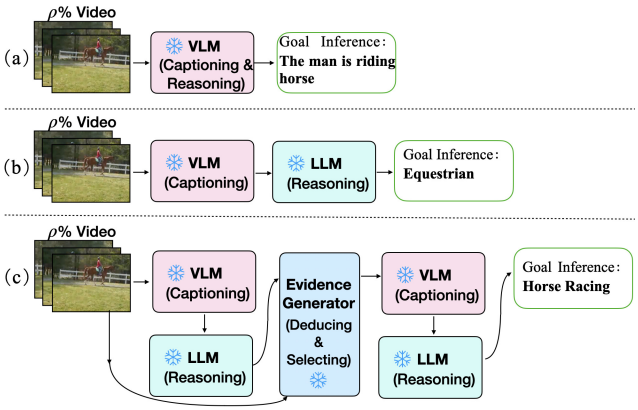


Figure 1. **Comparison of ViDSE with VLM-only and VLM+LLM Integrations** for the Video Goal Inference Task: (a) VLM consists of a vision encoder and an LLM, jointly tuned with video instructional data (e.g., BLIP-2, Video-ChatGPT, mPLUG-Owl); (b) VLM+LLM integrations separate VLM as the perception module and LLM as the reasoning module (e.g., mPLUG-Owl + Vicuna-13B); (c) ViDSE enhances the (b) by introducing an Evidence Generator to deduce and select relevant visual frames ($\rho = [10, 30, 50]$ denotes the ratio of temporally observed video).

However, language models [23, 25, 54] and other foundational models [21, 31] typically require large-scale annotated datasets and sufficient computational resources for training. This makes fine-tuning these models for every downstream task impractical, especially for goal inference with partially observed video sequences and limited annotations. Thereby, a new research direction has emerged: exploring training-free and open-vocabulary vision-language understanding [42, 47, 48], which enables models to handle new tasks without task-specific training or large amounts of annotated data.

Besides the training challenges of LLMs and VLMs, videos often contain redundant and irrelevant visual information, leading to ambiguity in goal inference. Current methods generally follow two approaches: using VLM-

only [20, 26, 51] or combining VLM with LLM [22, 45, 52] (see Figure 1a-b). The former integrates perception and reasoning with a combo of vision encoder and LLM, while the latter separates perception (VLM) and reasoning (LLM). While most methods show promising performance, there is a clear need for explicit filtering of ambiguous frames. This is particularly important during the goal inference perception and reasoning stages, a key focus of our research.

To tackle these challenges, we developed **ViDSE** (see Figure 1c, 2), a video framework for goal inference that reduces ambiguity by deducing and selecting only relevant visual evidence through an evidence generator. ViDSE operates within the VLM+LLM paradigm, enhancing perception and reasoning by focusing on selected evidence under traning-free open-vocabulary conditions. Specifically, it leverages frozen models, including Large Language Models and Vision Language Models, alongside the Vision foundational model without task-specific tuning. First, VLM (BLIP-2 [20]) generates textual captions for each video frame, while LLM (Vicuna [57]) deduces hypotheses (scripts) relevant to these captions. Next, the "evidence generator" uses CLIP [31] and the LLM-deduced hypotheses to dynamically select evidential frames. Finally, these selected frames are fed into the VLM+LLM pipeline to perform perception and reasoning, predicting the final goals.

We tested ViDSE on multiple video datasets for open-vocabulary tasks like goal inference and action recognition. The results show that ViDSE performs better than VLM-only and VLM+LLM. While VLMs excel at describing visual content, they often falter when it comes to reasoning, underscoring the necessity of ViDSE's LLM reasoning. By adding LLM reasoning, ViDSE handles complex video tasks without extra training. Our main contributions are:

- **Training-Free ViDSE**: We introduce ViDSE, which combines VLM, LLM, and Evidence Generator for open-vocabulary video tasks without training. VLM acts as the "*eye*", LLM as the "*reasoner*", and Evidence Generator as the "*selector*", communicating through language to conduct video deductive inferences.

- **Evidence Generator**: We propose a training-free module for deducing and selecting the evidential frames to support video inference. LLM generates hypotheses and scripts (sub-steps), while the vision encoder (e.g., CLIP) matches them to the frames as supporting evidence for final inferences.

- **Generalization on Video Tasks**: ViDSE is tested on four datasets covering goal inference and action recognition tasks. It performs on par with or better than state-of-the-art VLMs and VLM+LLMs, proving its generalizability.

2. Related Work

Supervised learning for video understanding has been extensively studied in the era of foundational models. With the success of foundational models on static images (e.g., CLIP [31]), numerous video models have been proposed to learn visual video representations from large-scale data. Representative works included [29, 32, 39, 44, 46, 50]. ViFi-CLIP [32] shows that fine-tuning CLIP with large-scale video data leads to better video classification. Whereas, with less data, prompt tuning CLIP can help reduce the risk of overfitting. Similarly, Vita-CLIP [46] proposes learnable prompts at different temporal levels to align video-text pairs. While [17] add learnable prompt vectors to the CLIP text encoder to create action classifiers. AIM [50] plugs adapters into backbones to reduce training computations and alleviate overfitting. These methods require supervised training with substantial video annotation data. In contrast, our ViDSE leverages off-the-shelf language and foundational models with a new deduction-then-selection module for action inference and recognition without training, extending the adaptability of foundational models.

Instructional tuning of videos uses both large language models and vision foundational models [23, 26, 49, 51, 54, 56]. These models are adapted using large-scale VQA datasets. They show robust zero-shot and open-vocabulary generation capability on unseen downstream video tasks. Specifically, Video-LLaMA [54] uses frozen VLM (ViT [15]) and LLM (e.g., Vicuna [10], LLaMA), and only learn the Q-Fromer [20]. Similarly, Video-LLaVA [23] combines LanguageBind [58] and Vicuna for video encoding and language processing and includes a projection layer to link visual and text tokens together. VideoChat [22] uses two separate VLMs to create visual captions and visual embeddings. These are then combined and fed into a LLM for question-and-answer processing. The mPLUG-Owl [51] model adopts a cross-attention mechanism with learnable queries to project visual tokens into textual space. LongVILA [49] is VLM that scaled up from [24] for longer context video understanding and uses multiple stages of supervised training. ViDSE differs by not needing to fine-tune (e.g., Q-Former or linear projection) and insert the training-free "evidence generator" between VLM+LLMs for video inferences.

Training-free open-vocabulary image understanding gaining extensive research interests by treating large-scale pre-trained models as tools. Many studies, like [27, 30], utilize strong zero-shot capabilities of pre-trained CLIP and combine it with ChatGPT-3.5 for open-vocabulary image classification. Other research efforts focus on solely enhancing CLIP's ability to understand different vocabularies without additional training, as in [42, 48]. Specifically, VisDesc [27] expands unseen categories using detailed text descriptions by inquiring ChatGPT and then pairs images

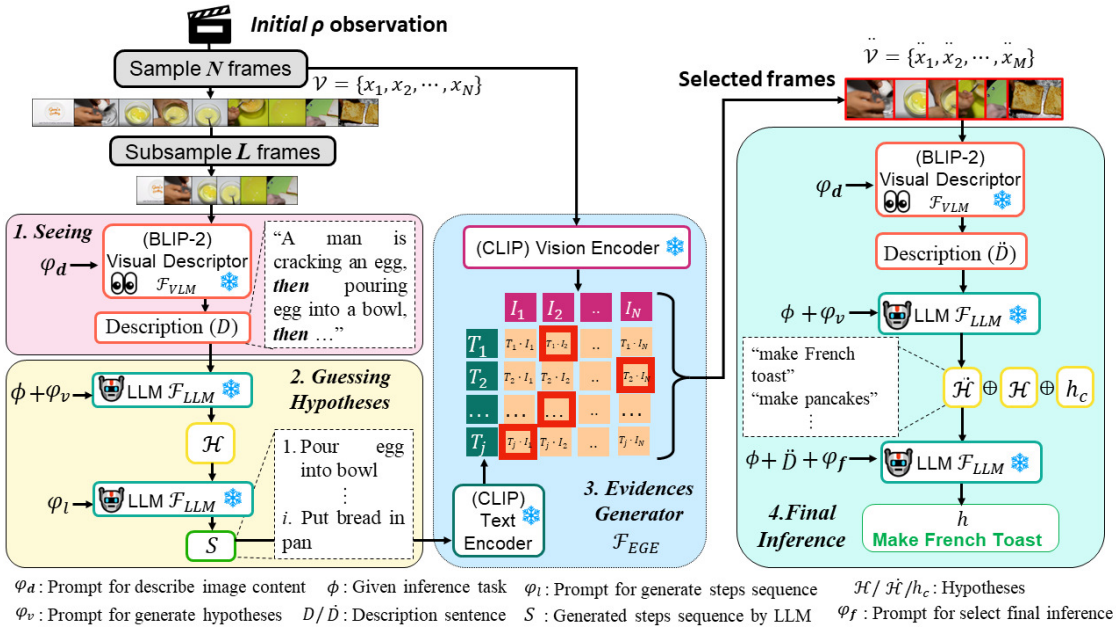


Figure 2. ViDSE contains four stages: *See*, *Guess*, *Select*, and *Infer*. (1). Seeing through Visual Descriptor (i.e., BLIP-2) translates visual frames into dense textual descriptions. (2). Guessing by LLM generate hypotheses ($\mathcal{H}$) and corresponding sub-events (steps). (3). Selecting frames using CLIP reduce irrelevant frames. (4). Inferring final answer by using selected frames with the "see" & "guess" process again. Best viewed on computer full screen.

with these descriptions using a frozen CLIP model. Similarly, the CHiLS [30] replaces coarse-defined categories with more specific sub-categories. These sub-categories are created using label hierarchies or consulting ChatGPT and matched with visual content using CLIP. Besides, SuS-X [42] creates a support set that includes open categories by stable-diffusion [34] or retrieval methods. Using CLIP models, it then measures the distance between a query image and the support set, broadcasting labels from the support set to the query. Xu et al. [48] utilize off-the-shelf mask generators and frozen CLIP for open-vocabulary semantic segmentation. ViDSE also employs ready-to-use BLIP-2, CLIP, and Vicuna, but it differs in handling dynamic video inputs and introduces training-free frame selection module for narrowing down evidence using foundational models.

Training-free open-vocabulary video understanding also uses pre-trained foundational models' perception and language models' reasoning abilities to tackle new video tasks. Example works like [9,52] involve using several large pre-trained models as tools. These models function in roles of perception and reasoning and interact with each other through language. Specifically, the Socratic Models [52] introduce a technique of multimodal prompting VLM+LLM models. This involves a combination of a vision-language model (like CLIP with BERT/GPT), an Audio Large Model [4], and a Large Language Model (LLM). This approach exchanges information between these large models through text and can handle new video tasks. Similarly, VideoChat-Captioner [9] set up a conversation between ChatGPT and BLIP-2, with ChatGPT asking questions and BLIP-2 answers based on the input video. The video's description is progressively enhanced through multiple rounds of automated conversation. Our ViDSE also aligns with this direction, focusing on interactions between LLMs and VLMs. Our main difference from existing methods is that we have found that focusing on the most important parts of a video using an evidence generator improves open-vocabulary video inference performance.

3. Method

Our ViDSE framework $\mathcal{F}$ solve the open-vocabulary inference task ϕ (e.g., "goal inference") by processing natural untrimmed video $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, which consists of N uniformly sampled frames. We infer the most likely hypothesis h based on the video observation without training or fine-tuning.

$$h = \mathcal{F}(\mathcal{V}, \phi) \tag{1}$$

Examples of the hypothesis include "*make French toast*" for inferring goals in cooking videos, and "*baby crawling*" for recognising actions in videos.

An overview of ViDSE modular framework is shown in Figure 2, it uses three frozen foundational models: VLM (BLIP-2) as the visual descriptor $\mathcal{F}_{\text{VLM}}$, LLM (Vicuna) as the reasoner $\mathcal{F}_{\text{LLM}}$, and evidence generator (CLIP) as the selector $\mathcal{F}_{\text{EGE}}$. Given target task ϕ and video $\mathcal{V}$, these models

work together in four stages: *See*, *Guess*, *Select*, and *Infer* . Details of each stage are as follows.

***Seeing* through Visual Descriptor** : We uniformly subsample L out of N selected frames. The visual descriptor $\mathcal{F}_{\text{VLM}}$ takes each sampled frame x_i as input and outputs a caption (text description) sentence c_i. We use BLIP-2 [20] (FLanT5-XXL) model as a visual descriptor and use a prompt φ_d to obtain the frame description, for example, φ_d =*what is the content of the image?*. The sequence of all frame captions is denoted by $\mathcal{C} = \{c_1, c_2, \cdots, c_L\}$ and there are a total of L captions. Next, we concatenate the captions in $\mathcal{C}$ into a single continuous description paragraph $\mathcal{D}$ using the word "*then*" to link them up so that $\mathcal{D}$ follows the form of "`<caption 1>, then, <catpion 2>, then, ... <caption L>`".

***Guessing* Hypotheses with LLM** : We use a LLM ($\mathcal{F}_{\text{LLM}}$) to guess the top-k initial hypotheses, $\mathcal{H} = \{h_1, h_2, \cdots, h_k\}$ for the given inference task, ϕ (Eq. (2)), with an instructional prompt φ_v. Here φ_v is "`I want to perform <task>, generate top-<k> hypotheses, given <text>`".

$$\mathcal{H} = \mathcal{F}_{\text{LLM}}(\mathcal{D}, \varphi_v) \tag{2}$$

Hereby, `<task>` is the task definition name (e.g. ϕ = goal inference) and `<text>` is description paragraph $\mathcal{D}$. Notably, we only show a simplified prompt version for quick reference and put the full instructional prompt in the supplementary section. We employ Vicuna [57] as the $\mathcal{F}_{\text{LLM}}$. An example of guessed hypotheses $\mathcal{H}$=["*make French toast*","*make pancakes*", $\cdots$] –see also Figure 2. We further expand each candidate hypothesis in $\mathcal{H}$ into a sequence of detailed events or steps, $\mathcal{S}$. We achieve this by using prompt φ_l in the form of "`List the steps to perform <hypotheses>`".

$$\mathcal{S} = \mathcal{F}_{\text{LLM}}(\mathcal{H}, \varphi_l) \tag{3}$$

Since there are k potential hypotheses, we eventually have k number of different step sequences. We gather all these sequences into $\mathcal{S} = \{[s_1^{h_1}, \cdots,], \cdots, [s_1^{h_k}, \cdots, s_i^{h_k}]\}$, re-flatten it into $\mathcal{S} = \{s_1, s_2, \cdots, s_j\}$ of j total steps. The reasons for expanding from $\mathcal{H} \rightarrow \mathcal{S}$ lies in two aspects. Firstly, steps contain more fine-grained information than the hypothesis, as a hypothesis is the outcome of executing a script containing a list of steps [37]. A specific step often corresponds directly to visual details, whereas a hypothesis may lack visual representation. Conversely, video inference tasks like goal inference encompass multiple sub-steps essential for inference based on deductive reasoning. By aligning the relevant frames with corresponding steps in a hypothesis, we can deduce that the hypothesis is a correct answer from the candidate set $\mathcal{H}$.

***Deducing* and *Selecting* by Evidence Generator** : The evidence generator select M out of N frames accroding to hypotheses $\mathcal{H}$ deduced steps $\mathcal{S}$, creating a subset of frames $\ddot{\mathcal{V}}$ where $\ddot{\mathcal{V}} \subset \mathcal{V}$ that are relevant to the inference task. This mechanism finds the most relevant frame $\ddot{x}_i$ ($\ddot{x}_i \in \ddot{\mathcal{V}}$) for each hypothesized step s_i in $\mathcal{S}$. We use frozen CLIP [31], a two-tower vision-language encoder to implement $\mathcal{F}_{\text{EGE}}$.

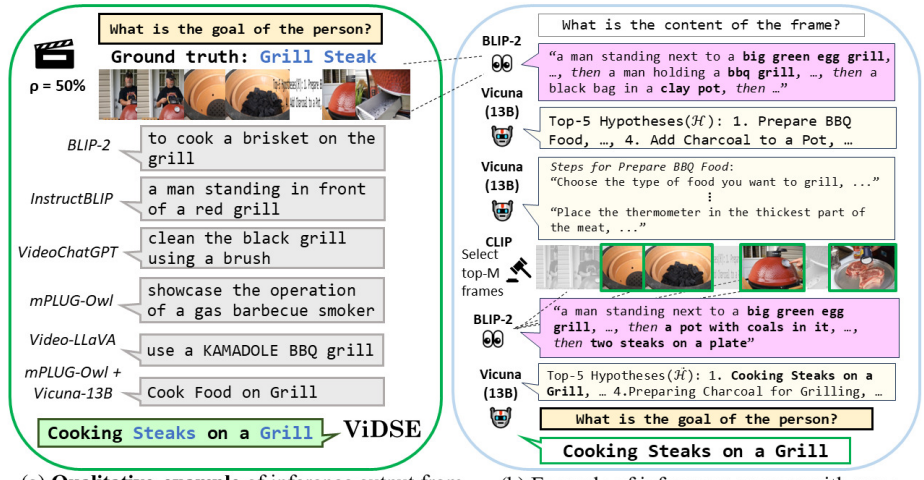
$$\ddot{\mathcal{V}} = \mathcal{F}_{\text{EGE}}(\mathcal{V}, \mathcal{S}) = \{\ddot{x}_1, \ddot{x}_2, \cdots, \ddot{x}_M\} \qquad s.t.\ M < N \tag{4}$$

Specifically, we extract features for N visual frames and S steps (text) with CLIP vision/text encoders. We then calculate the cosine similarity between each $\langle step, frame \rangle$ pair in Figure 2 ($\mathcal{F}_{\text{EGE}}$). Afterwards, we select the top M frames with the highest similarities into a set of evidence frames $\ddot{\mathcal{V}}$. We limit $M \leq 16$ and post-process $\ddot{\mathcal{V}}$ with duplicate filtering. With the evidence generator, we make sure that selected frames have diverse levels of information relevant to the task.

***Final Inferencing* by LLM** : We use the selected frames $\ddot{\mathcal{V}}$ to make inferences and generate the final hypothesis h by LLM in an open-vocabulary manner. We repeat the process from "Seeing through visual descriptor" and "Guessing hypothesis with LLM" except that we do not require the LLM to generate the steps again. Instead, we infer a second set of top-k hypotheses $\ddot{\mathcal{H}}$. Furthermore, we use the CLIP model to infer a single CLIP-based hypothesis using $\ddot{\mathcal{V}}$ and $\mathcal{H} \oplus \ddot{\mathcal{H}}$ which we denote as h_c. The h_c is selected from the candidate hypotheses ($\mathcal{H} \oplus \ddot{\mathcal{H}}$) by finding the best-matched hypotheses to the mean-pooled visual features of those selected frames using CLIP visual and textual embeddings. Then we take the hypothesis combination (operator denoted as $\oplus$) of all generated hypotheses, e.g., $\mathcal{H}$, $\ddot{\mathcal{H}}$ and h_c as the candidates and let LLM infer the final hypothesis h using the selected frame description $\ddot{\mathcal{D}}$ and final inference prompt φ_f as follows:

$$h = \mathcal{F}_{\text{LLM}\{\mathcal{H} \oplus \ddot{\mathcal{H}} \oplus h_c\}}(\ddot{\mathcal{D}}, \varphi_f). \tag{5}$$

Here $\ddot{\mathcal{D}}$ is obtained from the BLIP-2 model after processing $\ddot{\mathcal{V}}$. As before, we use the term "*then*" to form a coherent description of selected frames. The final inference prompt φ_f follows the form of "`I want to perform <task>, only select one answer from options <hypotheses>, given <text>`". The full prompts format is provided in supplementary. Notably, we fill the `<hypotheses>` with $\mathcal{H} \oplus \ddot{\mathcal{H}} \oplus h_c$, and `<text>` with $\ddot{\mathcal{D}}$. For operator $\oplus$, we ablate choices of union operator $\cup$ and concatenation as shown in supplementary and choose the latter one.



(a) **Qualitative example** of inference output from baselines and **ViDSE** on a CrossTask video (ρ = 50%)

(b) Example of inferences process with some intermediate outputs in ViDSE framework

Figure 3. Qualitative comparison of ViDSE with SOTAs on the CrossTask goal inference task. More examples are in the supplementary.

4. Experiments

We evaluate the proposed ViDSE on goal inference tasks using two video datasets under training-free, open-vocabulary conditions. The evaluation metrics, as presented in Section 4.2, include SPICE(S) [2], CIDEr(C) [43], METEOR(M) [14], and BERTScore(B) [55], SBERT(SB) [33]) as in [36] to measure semantic similarity between ground-truth answers and open-vocabulary inferences. We also evaluate open-ended inferences by using *LLM-as-judge*, as in [7, 8]. Dense ablations are conducted, with results reported in Section 4.3 and the supplementary material. Furthermore, we demonstrate ViDSE's generalizability on action recognition tasks in Section 8.

4.1. Datasets

CrossTask [59] consists of 4,700 instructional videos (avg. 5 minutes long) about daily life. We evaluate the goal inference task using the validation set (360 untrimmed videos) that covered 18 primary tasks and only use the task labels as our ground truth goal labels during evaluation.

COIN [40] contains 11,827 instructional videos (avg. 2.36 minutes long) with 180 distinct tasks. We evaluate the test set of 2,797 untrimmed videos and only use the corresponding task label as the goal label.

4.2. Goal Inference

For ϕ = goal inference task, we evaluate the ViDSE on CrossTask and COIN datasets. Specifically, we infer the human's goal with open vocabulary when chronologically observing the initial ρ=10%, 30%, and 50% parts of the videos.

As shown in Table 1, the ViDSE outperforms the current SOTA multimodal language models (MLMs) on most evaluation metrics under a training-free open-vocabulary setting. Notably, the ViDSE surpassed pre-trained end-to-end MLMs, including the BLIP models, Video-LLaVA, and mPLUG-Owl. Compared with mPLUG-Owl + Vicuna which uses LLM to make inferences by using video-level description from MLM, ViDSE is outperformed it by +5.1 (47.6 *vs* 42.5) at ρ=10%, +4.9 (50.9 *vs* 46.0) at ρ=30% and +4.7 (50.2 *vs* 45.5) at ρ=50% respectively. This trend indicates that with the help of the frame selection module, ViDSE can make better inferences. On the COIN dataset, with shorter inputs ρ=10%, ViDSE fall behind Video-LLaVA by -3.7 (45.0 *vs* 48.7) on SBERT; when ρ=30%, ViDSE surpass the Video-LLaVA by +0.8 (49.6 *vs* 48.8); and ρ=50%, ViDSE broaden the gap by +3.4 (51.5 *vs* 48.1). The proposed ViDSE shows overall improvement across the ρ on the goal inference task. The reason is that ViDSE can select relevant frames, thus effectively keeping the necessary information in long-duration untrimmed videos, whereas the other methods lack this flexibility.

4.3. Ablation Study

Ablation Evidence Generator Component. We compare the performance of the ViDSE framework against a simple counterpart without an evidence generator. This baseline uses BLIP-2 as a visual descriptor to generate captions, allowing Vicuna13B to directly infer the goal based on the given frame captions. The baseline does not generate steps (Eq. (3)), and there is no Evidence Generator Component. As in Table 2, the performance drops when the evidence generator is absent on video goal inference. The evidence generator helps to provide relevant frames that support accurate goal inference. We put qualitative results of selected frames in the supplementary material, demonstrating the correctness of the Evidence Generator. In addition, we extend this ablation to additional datasets and report results in supplementary since we also show the generaliza-

CrossTask	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
BLIP-2 [20]	13.3	27.2	11.6	15.9	32.2	11.7	24.2	11.6	16.7	33.1	12.6	24.8	12.2	17.5	34.5
InstructBLIP [12]	6.2	6.6	5.5	-0.2	23.6	4.9	4.6	4.7	-0.4	22.4	4.8	4.2	4.5	-0.3	22.8
Video-ChatGPT [26]	9.0	14.9	10.5	11.9	35.4	10.0	18.1	12.1	15.2	38.4	9.7	23.1	12.5	16.6	39.6
mPLUG-Owl [51]	9.4	13.2	10.2	7.3	35.1	10.1	12.5	10.2	8.9	38.2	10.5	21.3	10.5	10.3	39.4
Video-LLaVA [23]	15.6	39.6	10.6	22.6	43.1	15.3	42.4	10.7	24.0	45.0	17.6	41.1	10.7	25.9	47.2
mPLUG-Owl+V13B	15.7	54.5	11.2	26.9	42.5	16.0	62.3	12.6	28.6	46.0	17.0	50.7	12.8	28.4	45.5
ViDSE (V13B)	**23.0**	**80.1**	**15.4**	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	**80.8**	**16.3**	**34.5**	**50.2**
COIN	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
BLIP-2 [20]	14.4	27.1	9.4	14.8	34.5	14.2	27.7	9.4	15.8	36.0	14.8	28.9	9.7	16.4	37.2
InstructBLIP [12]	7.0	11.6	6.4	3.7	27.6	6.8	9.4	6.0	4.0	27.7	7.6	10.6	6.5	4.2	28.3
Video-ChatGPT [26]	13.2	29.4	10.7	14.8	41.5	13.3	29.1	10.6	14.8	41.8	12.5	28.0	10.5	14.7	41.0
mPLUG-Owl [51]	10.8	15.4	8.7	7.6	35.7	11.8	18.9	9.7	9.4	40.0	12.8	21.4	10.5	10.3	42.2
Video-LLaVA [23]	**21.0**	45.2	12.1	19.9	**48.7**	21.3	44.5	12.0	20.2	48.8	20.4	43.5	11.9	19.8	48.1
mPLUG-Owl+V13B	19.3	60.3	11.9	**28.6**	47.3	18.9	61.2	12.0	29.0	47.5	20.1	63.7	12.1	29.3	47.7
ViDSE (V13B)	20.4	**62.6**	**12.5**	27.2	45.0	**23.0**	**71.4**	**13.7**	**30.4**	**49.6**	**25.1**	**76.7**	**14.3**	**31.6**	**51.5**

Table 1. Open-vocabulary goal inferences results on CrossTask and COIN datasets. We report following metrics in %: SPICE (S), CIDEr (C), METEOR (M), BERTScore (B), and SBERT (SB). Best and second best results are highlighted by bold and underline.

CrossTask	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
w/o ES	18.3	61.3	12.7	25.0	42.9	19.7	72.0	14.0	27.5	46.8	22.1	**83.0**	15.1	30.3	48.8
with ES	**23.0**	**80.1**	**15.4**	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	80.8	**16.3**	**34.5**	**50.2**
COIN	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
w/o ES	18.3	52.8	11.4	23.0	41.9	21.0	63.2	12.7	26.7	46.1	22.0	68.2	13.2	27.8	47.7
with ES	**20.4**	**62.6**	**12.5**	**27.2**	**45.0**	**23.0**	**71.4**	**13.7**	**30.4**	**49.6**	**25.1**	**76.7**	**14.3**	**31.6**	**51.5**

Table 2. Ablation study of the evidence generator component across CrossTask and COIN datasets.

tion ability of ViDSE as discussed in later Section 8.

Select Evidences from Visual Frames *vs* Frame Captions We also investigate the effect of choosing relevant frames based on the original frame captions $\mathcal{C}$ and hypothesis steps using text-to-text matching. We compare the steps $\mathcal{S}$ with frame-captions $\mathcal{C}$ using text-to-text similarity using SBERT model-based text embeddings. Then, those frames (captions) with the highest similarity to the steps are selected. We compare the step-to-caption approach vs the step-to-visual-frame similarity-based approach that uses CLIP visual features. Results in Table 3 show that using the CLIP to select visual frames is better than using SBERT-based text matching. More ablations in supplementary.

Select Evidence using Frame Captions versus Hypothesized Steps by LLM We also compare with counterparts that use frame captions $\mathcal{C}$ generated by visual descriptor (e.g. BLIP-2), and then use CLIP to select the relevant frames from the N sampled frames. Table 4 shows that using LLM-generated steps to find the evidence frames is better for inference performance. The LLM-generated steps capture more details of contextual information and better align the selected frames with the underlying task. This results in a more coherent and relevant selection of evidence frames, enhancing overall performance.

Select Evidence using Hypotheses *vs* Expanded Hypothesized Steps by LLM We compare with counterparts that directly use top-k hypotheses, $\mathcal{H}$, to select the relevant frames from the N sampled frames. Table 5 shows that using LLM-generated steps, which consist of more "fine-grained" information to find the evidence frames, is better for inference performance.

5. Analysis on Impact of Evidence Generator

To further evaluate the effectiveness of our evidence generator, we measured how well the selected frames matched with the ground truth label. We use different frame sampling methods for frame selection; we then use CLIP [31] to measure the similarity between the selected visual frame and text label features. We obtained the visual features by

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
Steps-to-caption(text)	21.8	75.1	15.3	**32.8**	47.2	22.3	**96.7**	16.8	**35.3**	50.6	23.3	**81.3**	15.8	34.2	49.3
Steps-to-frame(visual)	**23.0**	**80.1**	**15.4**	32.3	**47.6**	**23.1**	91.7	**16.9**	35.0	**50.9**	**24.4**	80.8	**16.3**	**34.5**	**50.2**

Table 3. Comparison between step-to-frame vs step-to-caption matching in the Evidence Selector component on CrossTask dataset for goal inferences.

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
Use captions	21.4	79.7	15.1	31.2	45.9	21.4	83.9	16.7	33.3	48.9	22.2	80.4	15.8	33.3	49.3
Use generated steps	**23.0**	**80.1**	**15.4**	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	**80.8**	**16.3**	**34.5**	**50.2**

Table 4. Comparison between captions-to-frame versus steps-to-frame matching in the Evidence Selector on CrossTask dataset for goal inference.

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
Use hypotheses	21.9	79.9	15.2	31.9	46.9	21.6	84.2	16.4	33.8	49.5	23.6	79.9	16.2	33.6	50.1
Use generated steps	**23.0**	**80.1**	**15.4**	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	**80.8**	**16.3**	**34.5**	**50.2**

Table 5. Comparison between hypotheses-to-frame versus steps-to-frame matching in the Evidence Selector component on CrossTask dataset for goal inferences.

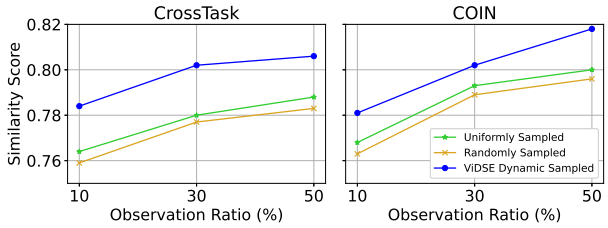


Figure 4. Comparisons of Visual-Textual Similarities w/ and w/o **Deducing and Selecting** by Evidence Generator on goal inference task.

averaging the sampled frames. The results shown in Figure 4 indicate that the frames selected by the ViDSE evidence generator have better similarity with the text features of the ground truth label.

6. New Metric: LLM-as-Judge

Besides conventional evaluation metrics, we use the Llama3-8B[1] model as a "judge" to compare generated inferences with ground truths. This is inspired by recent studies showing that large language models (LLMs) can effectively act as "judges" to evaluate inferencing qualities [7, 8]. With LLM as judge, we ask it to rate a binary output—"yes" or "no"—indicating whether the generated inference and the ground truth have similar meanings. The experimental results in Table 6 demonstrate that ViDSE receives more "yes" ratings from the Llama3 judge than other methods, indicating the effectiveness of ViDSE. The evaluation prompt of Llama3 and full results are in the supplementary materials.

[1]https://llama.meta.com/llama3/

Method	**CrossTask**			**COIN**			**UCF101**	**AN**
Percentage (%)	10	30	50	10	30	50	100	100
BLIP-2 [20]	32.2	34.1	35.8	31.2	31.6	32.2	72.8	53.4
InstructBLIP [12]	11.7	10.0	10.4	16.1	15.1	14.8	<u>74.8</u>	54.1
Video-ChatGPT [26]	22.4	19.8	21.0	24.6	25.0	24.3	64.7	44.7
mPLUG-Owl [51]	27.8	38.8	42.8	26.8	32.1	34.6	65.9	49.0
Video-LLaVA [23]	<u>42.2</u>	<u>43.6</u>	<u>49.0</u>	**42.5**	**43.0**	<u>41.2</u>	63.6	<u>60.4</u>
mPLUG-Owl+V13B	39.1	43.1	44.5	<u>38.7</u>	38.6	30.4	74.1	54.2
ViDSE (V13B)	**51.8**	**58.1**	**63.2**	38.1	<u>42.5</u>	**47.3**	**79.7**	**71.9**

Table 6. Accuracy evaluated by Llama3 judge. Best and second best results are highlighted by bold and underline respectively. (AN : ActivityNet)

7. Inferences Time and Amount of LLM Calls

We record the inference time and number of LLM calls for comparison on a single NVIDIA A100 GPU. The average time taken excludes the time required for loading and pre-processing the videos or visual frame, only start timing when prompting the model to make an inference based on a given inference task ϕ (e.g., "goal inference"). For BLIP-2 and InstructBLIP, we query the language model 16 times as we use them for frame-level inferences. For mPLUG-Owl + Vicuna-13B, we only time the inference after mPLUG-Owl generates the video-level caption. The proposed ViDSE using original Vicuna-13B [10] shows a longer inference time than other MLMs that only need one LLM call. However, the inference time of ViDSE could potentially be shortened through engineering efforts, as shown by using the quantized model, or LLM from [11], which compresses and serves LLM more efficiently, but resulting in degraded inference performance. In addition, the generated intermediate outputs from the LLM are not limited to the primary inference task. These outputs can be leveraged for other downstream tasks, such as video retrieval or summarization,

which could effectively reduce the overall computation required for subsequent tasks where multiple analyses of the same video are required.

Methods	LLM	LLM size	Average Time Taken (s)	Number of LLM calls
BLIP-2 [20]	FlanT5-XXL	11B	7.63	16
InstructBLIP [12]	FlanT5-XXL	11B	10.01	16
Video-ChatGPT [26]	Vicuna7B	7B	1.87	1
mPLUG-Owl [51]	Llama7B	7B	3.92	1
Video-LLaVA [23]	Vicuna7B	7B	2.31	1
mPLUG-Owl+V13B	Vicuna13B	13B	0.50	1
ViDSE	Vicuna13B	13B	15.17	4
	GPT-3.5	*UD*	6.70	4
	Llama3-8B	8B	8.12	4
	Vicuna13B by [11]	13B	4.92	4
	Vicuna13B-8bit	13B	13.25	4

Table 7. Average time taken (seconds) for video inference. "***UD***" is indicating "Undisclosed".

8. Generalizability on Action Recognition

We validate the generalizability of ViDSE on video action inference task (i.e., ϕ = action recognition). We tested ViDSE on UCF101 and ActivityNet datasets using untrimmed video under the same training-free and open-vocabulary settings. **UCF101** [38] is a widely utilized benchmark for action recognition tasks. It contains 13,320 short videos (avg. 7.5 seconds long) and encompasses 101 distinct action classes. We evaluate all three test splits of the dataset. **ActivityNet-v1.3** [16] contains 19,994 untrimmed YouTube videos (avg. 2 minutes long) consisting of 200 action classes. We evaluate the validation set, which consists of 4,926 videos.

As in Table 8, ViDSE outperforms SOTA multimodal LLM on UCF101 and ActivityNet datasets regarding BERTScore and SBERT. This indicates that ViDSE could generate good semantically equivalent inferences as the ground truth categories. On the UCF101, the ViDSE falls behind models like BLIP-2 and InstructBLIP in terms of metrics such as SPICE, CIDEr, and METEOR. The reason is that the BLIPs are pre-trained on image-captioning tasks and excel at generating short image-level captions. Besides, each frame from the short video of UCF101 is more likely to convey similar information about the actions, so frame selection may not be that important in those short videos. In contrast, ViDSE performed better on the ActivityNet, which contained noisy video input, highlighting the advantage of the evidence generator. Since action videos contain fewer sub-events (steps) than long-duration instructional videos (e.g., CrossTask), ViDSE's advantage is lower than that of the goal inference task. However, the ViDSE still achieves comparable performance compared to end-to-end pre-trained multimodal LLM. These findings validate the generalizability of ViDSE and its potential to be extended to other action-relevant tasks without training.

Method	**UCF101**					**ActivityNet**				
	S	C	M	B	SB	S	C	M	B	SB
BLIP-2 [20]	**21.0**	48.9	**16.2**	12.5	60.6	22.3	72.1	13.6	18.4	53.6
InstructBLIP [12]	**21.0**	**87.8**	13.2	21.8	61.9	10.3	42.2	6.5	5.5	46.5
Video-ChatGPT [26]	13.6	27.7	13.2	3.0	54.0	17.9	46.3	13.3	13.0	54.6
mPLUG-Owl [51]	13.4	31.7	13.9	5.8	54.8	14.8	33.0	11.5	11.0	51.1
Video-LLaVA [23]	12.1	24.8	12.7	4.9	50.2	19.8	47.6	**14.9**	16.6	53.7
mPLUG-Owl+V13B	18.2	71.7	12.9	24.5	58.7	22.0	82.2	13.2	25.6	59.2
ViDSE (V13B)	20.7	83.9	15.7	**29.3**	**63.6**	**24.0**	**94.0**	14.7	**28.8**	**61.0**

Table 8. Open-vocabulary action recognition on UCF101 and ActivityNet1.3 datasets.

9. Discussion and Conclusion

In conclusion, we introduce the ViDSE, a training-free modular framework for open-vocabulary video goal inference. The ViDSE uses three frozen large models, BLIP-2, CLIP, and Vicuna, to accomplish a stage video inference process: *See*, *Guess*, *Select*, and *Infer*. We validate that these foundational models could play different roles and interact well with each other through language. We also propose a training-free evidence generator that deduces and selects relevant frames for drawing inference. Our experimental results confirm the ViDSE's effectiveness and its ability to generalize to two distinct video inference tasks, thereby demonstrating its broad applicability. The ViDSE, with its current capabilities, can be further enhanced by integrating more advanced foundational models, promising even better results in the future. Its limitations lie in its reliance on LLMs to draw inferences; thereby, it is difficult to control the generation process and suffers from LLM drawbacks like hallucinations. Besides, LLMs are statistical-based methods and do not contain an explicit logical reasoning process, causing ViDSE to have weak explainability. Despite limitations, the proposed framework is novel for training-free open-vocabulary inference tasks on video data.

Acknowledgments This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001). This research is also supported by funding allocation to B.F. by the Agency for Science, Technology and Research (A*STAR) under its SERC Central Research Fund (CRF), as well as its Centre for Frontier AI Research (CFAR)

References

[1] Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and

vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022. 1

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 5

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1

[4] Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*, 2022. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[6] 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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712, 2023. 5, 7

[8] David Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John F. Canny. Clair: Evaluating image captions with large language models. In *EMNLP*. Association for Computational Linguistics, 2023. 5, 7

[9] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023. 3

[10] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 2, 7, 4

[11] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023. 7, 8, 1, 2

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 6, 7, 8, 3

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33 – 55, 2020. 1

[14] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[16] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1, 8

[17] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[19] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. 1

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 4, 6, 7, 8, 3

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

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

[23] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 2, 6, 7, 8, 3

[24] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[26] Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv*, abs/2306.05424, 2023. 2, 6, 7, 8, 3

[27] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 2

[28] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022. 1

[29] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 2

[30] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 2, 3

[31] 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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 4, 6

[32] H. Rasheed, M. Khattak, M. Maaz, S. Khan, and F. Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, jun 2023. 2

[33] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 5

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[35] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021. 1

[36] A. Sabir, F. Moreno-Noguer, and L. Padro. Visual semantic relatedness dataset for image captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5598–5606. IEEE Computer Society, jun 2023. 5

[37] Roger C Schank and Robert P Abelson. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157, 1975. 4

[38] Soomro, Khurram, Zamir, Amir Roshan, Shah, and Mubarak. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[39] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 2

[40] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, 2019. 5

[41] 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. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 1

[42] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 1, 2, 3

[43] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[44] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1, 2

[45] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022. 2

[46] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044, 2023. 2

[47] Jinmeng Wu, Chong Cao, Hao Zhang, Basura Fernando, Yanbin Hao, and Hanyu Hong. Pointtfa: Training-free clustering adaption for large 3d point cloud models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, 2024. 1

[48] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 1, 2, 3

[49] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. 2024. 2

[50] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. 2

[51] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023. 1, 2, 6, 7, 8, 3

[52] Andy Zeng, Adrian S. Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Peter R. Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, 2022. 2, 3

[53] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 917–925, 2021. 1

[54] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 2

[55] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 5

[56] Qipeng Zhao, Ce Zhang, Shijie Wang, Changcheng Fu, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *ICLR*, 2024. 2

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2, 4, 1

[58] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 2

[59] Zhukov, Dimitri, Alayrac, Jean-Baptiste, Cinbis, Ramazan Gokberk, Fouhey, David, Laptev, Ivan, Sivic, and Josef. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 5

The supplementary material is structured as follows: Section 1 provides an extensive set of ablation studies and analysis. Section 2 details the implementation of baseline methods and ViDSE, including the prompts used for large language model in our experiments. Section 3 showcases qualitative results with additional examples illustrating the inference process.

1. Supplemental Ablations and Analysis

1.1. Ablation Evidence Generator on ActivityNet dataset

We extended the ablation of evidence generator on ActivityNet-v1.3 dataset since we have shows the generalization ability of proposed ViDSE on action recognition task. We report the results in Table 9. The performance of ViDSE with evidence generator is outperform the counterpart that without evidence generator. The untrimmed videos in the ActivityNet dataset contain many frames unrelated to the target actions. Therefore, we shows the effectiveness of evidence generator in deducing and selecting the more relevant frames in order to perform action recognition.

ActivityNet	ρ = 100%				
	S	C	M	B	SB
w/o ES	21.2	79.6	12.7	22.3	57.4
with ES	**24.0**	**94.0**	**14.7**	**28.8**	**61.0**

Table 9. Ablation study of the evidence generator component on ActivityNet dataset.

1.2. Result Table of Analysis on Impact of Evidence Generator

In addition to the plots of visual-textual similarities with and without evidence generator on goal inference task, we report the full results number in Table 10. We have also included experiments on ActivityNet dataset to shows that evidential frames selected by ViDSE with evidence generator have better alignment with actual labels compared to uniformly or randomly frame sampling.

1.3. Prompt for LLM-as-Judge

The Figure 5 shows the complete prompt for Llama3-8B to act as judge and evaluate open-ended inferences.

1.4. Ablation Number of Iteration for Deduction and Selection Process

We compare ViDSE (1 iteration) with the counterparts that perform 2 and 3 iterations of frame deduction and selection process. Table 11 shows that more iterations of frame selection does not yield improvements. This reflects that one evidence generator is sufficient to select relevant frames for make inference and balance computations and performance well.

1.5. Ablation Number of Frames.

We also study the influence of the number of sampled frames, L, and selected frames, M together, by varying the frame number limit so that $L, M \leq \{4, 8, 16, 32\}$. Table 12 shows that performance is optimal when limited to 16 frames, as it also indicates that including more frames does not improve performance.

1.6. Ablation on Large Language Model.

We conduct ablation on using different LLM (e.g. Vicuna [57], GPT-3.5-Turbo [6], Llama-3-8B-Instruct) in the $\mathcal{F}_{\text{LLM}}$ and compare their inference performance. As shown in Table 13, the Vicuna-13B model performs better than Vicuna-7B while achieving comparable performance with GPT-3.5. In addition, we also compared with the quantized Vicuna-13B-8bit model and Vicuna-13B model from [11] which compresses the LLM and speeds up the inferences. This ablation study suggests that using more robust LLMs could enhance inference performance.

1.7. In-Context Learning Prompt.

We ablate the effect of In-Context Learning [6, 28, 35] (ICL) within the LLM prompt for open-vocabulary inference in the LLM prompt. Table 14 results suggest that using ICL helps improve open-vocabulary inference performance.

1.8. Hypothesis from CLIP.

We also study the impact of the hypothesis h_c from CLIP for video inference. The Table 15 shows using ($\mathcal{H} \oplus \ddot{\mathcal{H}} \oplus h_c$) as an option list for the final stage inference brings a slight improvements.

1.9. Operators to Combine Hypotheses List.

We test two types of operators $\oplus$ to combine $\mathcal{H}$, $\ddot{\mathcal{H}}$ and h_c. One is list concatenation: [$\mathcal{H}$] + [$\ddot{\mathcal{H}}$] + [h_c] and another is union of set $\{ \mathcal{H} \} \cup \{ \ddot{\mathcal{H}} \} \cup \{h_c\}$. Their main difference is list concatenation allows redundant options, but the union operator does not; this would affect the frequency of individual hypotheses inputted to LLM. As in Table 16, the concatenation operator performs better than the union operator.

2. Implementation Details

In this section, we provide the implementation details of both baselines and the proposed ViDSE framework, including the prompts used to query the vision-language models (VLM) and large language model (LLM).

Method	CrossTask			COIN			ActivityNet
	10%	30%	50%	10%	30%	50%	100%
Uniformly sampled	0.764	0.780	0.788	0.768	0.793	0.800	0.815
Randomly sampled	0.759	0.777	0.783	0.763	0.789	0.796	0.813
ViDSE dynamic sampled	**0.784**	**0.802**	**0.806**	**0.781**	**0.802**	**0.818**	**0.831**

Table 10. Similarity score between visual and text features by CLIP after frame selection process.

```
Let A = <Ground Truth Label>, Let B = <Inferences>.
Determine if A and B have similar meanings, then provide a binary output of 'Yes'
or 'No' only.
```

Figure 5. Prompt for Llama3 to judge correctness between the generated inferences and ground truth.

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
1 iteration	23.0	80.1	15.4	32.3	47.6	23.1	91.7	16.9	35.0	50.9	24.4	80.8	16.3	34.5	50.2
2 iterations	23.1	73.6	15.0	33.3	47.5	21.8	76.2	15.8	33.4	49.2	23.4	83.2	16.1	32.5	49.4
3 iterations	23.5	74.6	15.4	32.8	47.6	20.7	72.4	15.2	32.7	48.6	22.9	80.3	16.2	33.5	49.7

Table 11. Ablation study on iteration of frame selection.

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
4 frames	19.1	59.5	12.9	29.4	43.3	16.8	68.6	13.2	30.2	44.0	16.5	69.6	13.1	31.6	45.5
8 frames	20.4	70.8	13.7	30.7	46.2	21.1	82.8	15.6	33.6	49.6	22.7	**84.7**	16.2	**35.7**	50.8
16 frames	**23.0**	**80.1**	**15.4**	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	80.8	16.3	34.5	50.2
32 frames	19.3	64.0	14.8	31.1	46.4	21.0	79.9	15.5	30.7	47.3	23.5	83.8	**17.1**	34.5	**51.5**

Table 12. Ablation of number of sampled frames (L) and relevant frames selected (M).

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
Vicuna (7B)	20.1	77.2	13.4	30.5	45.4	21.5	88.6	14.3	32.0	47.4	21.2	86.5	14.8	32.6	48.6
Vicuna (13B)	23.0	**80.1**	15.4	32.3	47.6	**23.1**	91.7	16.9	35.0	50.9	**24.4**	80.8	16.3	34.5	50.2
Vicuna (13B) by [11]	**23.8**	78.6	15.6	33.5	48.3	21.3	82.9	15.7	33.3	49.4	22.7	76.1	16.0	33.0	49.6
Vicuna (13B) 8bit	21.0	74.9	**16.8**	**34.2**	**48.9**	20.7	80.6	17.1	35.2	50.7	23.9	82.5	17.0	36.5	51.5
GPT-3.5-Turbo	18.7	75.4	15.5	31.3	47.0	19.6	92.3	16.7	35.5	**51.3**	20.9	88.6	17.5	37.8	**52.5**
Llama3 (8B)	18.8	75.4	15.4	29.8	44.6	21.9	**109.3**	**18.0**	**37.6**	**51.3**	23.3	**116.9**	**17.9**	**40.4**	51.7

Table 13. Ablation study of the LLMs.

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
without ICL	19.7	46.4	12.1	19.0	42.4	18.9	38.2	11.9	16.7	42.3	18.5	36.3	11.2	16.1	41.8
with ICL	**23.0**	**80.1**	**15.4**	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	**80.8**	**16.3**	**34.5**	**50.2**

Table 14. Ablation study of the In-Context Learning (ICL) prompt.

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
w/o h_c	22.7	80.1	15.2	32.3	47.2	22.4	91.7	16.5	34.5	50.3	23.7	76.2	15.9	33.8	49.2
With h_c	**23.0**	80.1	**15.4**	32.3	**47.6**	**23.1**	91.7	**16.9**	**35.0**	**50.9**	**24.4**	**80.8**	**16.3**	**34.5**	**50.2**

Table 15. Ablation study of hypothesis from CLIP (h_c).

Method	ρ = 10%					ρ = 30%					ρ = 50%				
	S	C	M	B	SB	S	C	M	B	SB	S	C	M	B	SB
Set Union Operator	22.8	77.1	15.4	31.8	47.2	21.8	83.0	15.8	33.2	49.5	23.4	78.2	15.9	33.8	49.8
List concatenation	**23.0**	**80.1**	15.4	**32.3**	**47.6**	**23.1**	**91.7**	**16.9**	**35.0**	**50.9**	**24.4**	**80.8**	**16.3**	**34.5**	**50.2**

Table 16. Ablation study on concatenation of hypotheses.

2.1. Open-vocabulary Inference Baselines

2.1.1 BLIP-2

BLIP-2 [20] has proficient zero-shot image question-answering ability; we use it for frame-level inference (16 frames) as it is designed for image-to-text tasks. We use BLIP-2 with FLanT5-XXL model with the prompts: ``Question: What is the intention or goal of the person in the photo? Short answer: '' for goal inference task, while ``Question: What is the ongoing action of the person in the photo? Short answer: '' for the action recognition task. We then computed the evaluation metrics of each frame-level caption against the ground truth label and took the mean values as the final measurement of each video-level inference.

2.1.2 InstructBLIP

InstructBLIP [12] with FLanT5-XXL model is instruction-tuned based on pre-trained BLIP-2 [20]. Instead of a question-answer format, we use an instruction format prompts: ``Please provide the intention or goal of the person in the photo.'' for goal inference task, whereas ``Please provide a short answer of the ongoing action of the person in the photo.'' for the action recognition task. We use the same evaluation method as the BLIP-2 baseline since both are applied for frame-level inference (16 frames).

2.1.3 Video-ChatGPT

Video-ChatGPT [26] is pre-trained on 100K video-caption pairs and works well in various open-vocabulary video question-answering tasks. We provide the direct and clear question prompt, ``What is the intention or goal of the person in the video?'' and ``What is the ongoing action of the person in the video?'' to the model for zero-shot video goal inference and action recognition, respectively. We set the frame number parameter to 16.

2.1.4 mPLUG-Owl

mPLUG-Owl [51] is another large MLM demonstrating remarkable zero-shot abilities on various open-vocabulary visual inference tasks. We follow the suggested prompt template, ```The following is a conversation between a curious human and an AI assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. Human: <|video|> Human: {Question} AI: '''. The Question is filled with ``What is the intention or goal of the person in the video?'' for the goal inference task, whereas ``What is the ongoing action of the person in the video?'' for the action recognition task. The number of sampled frames per video is 16.

2.1.5 Video-LLaVA

Video-LLaVA [23] proposed as MLM that uses a unified visual representation before projection to enhance downstream visual-language understanding. We use it as a baseline to perform open-vocabulary video inference with the following prompts: ``Write a short answer of the intention or goal of the person in the video. The person in the video is: '' for goal inference, whereas ``Write a short answer of the ongoing action of the person in the video. The person in the video is: '' for action recognition. It is only supporting to take a maximum of 8 frames for each video inference at the moment we implemented it.

2.1.6 Combination of mPLUG-Owl & Vicuna-13B

mPLUG-Owl + Vicuna-13B is another baseline method that use the mPLUG-Owl as a visual descriptor and Vicuna-13B as LLM agent to make inference without any frame selection process. We input the prompt to mPLUG-Owl as ```` ```The following is a conversation between a curious human and AI assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. Human: <|video|> Human: What is the content of the video? AI: ''' ````, and then we use the LLM to infer directly on top of the video description generated by mPLUG-Owl. The prompt for LLM is similar to the prompt template used by ViDSE as shown in Table 18. Instead of list the top-k hypotheses, we ask the LLM to provide only one answer.

2.2. ViDSE Framework

2.2.1 Seeing through Visual Descriptor.

We use BLIP-2 with FLanT5-XXL [20] to generate a caption for every sampled frame by using a general prompt (φ_d): ```` ``Question: What is the content of the image? Answer: '' ```` for all inference tasks. After L number of captions are generated, we preprocess the captions by deduplicate the identical captions if there is any and concatenate the rest by using the word "*then*" to create a high-level description so that $\mathcal{D}$ follows the form of "`<caption 1>, then, <catpion 2>, then, ... <caption L>`". In a later process, we also do the same for the M selected frames to generate a new description $\ddot{\mathcal{D}}$.

2.2.2 Deducing and Selecting by Evidence Generator.

The evidence generator module is pivotal in aligning visual features with text features to identify the evidential frames. We employ the frozen visual and text towers from the CLIP [31] model by using the ViT-B/16 backbone to effectively integrate visual and textual information for optimal evidence frame selection. Specifically, we use CLIP vision encoder to encode N visual frames and generate the frame features, then we use CLIP text encoder to generate text features by encoding the hypothesized steps S generated by the LLM. Subsequently, we compute similarity between visual features and text features. We select the top similarity score of M frames and resulting in a new set of evidence frames.

2.2.3 Guessing Hypotheses and Final Inference by LLM.

We use the readily available LLMs, specifically Vicuna-13B [10], in the goal inference and action recognition experiments. For Vicuna, we set the temperature to 0.001 and the repetition penalty to 1.0. The full prompt template ($\varphi_v, \varphi_l, \varphi_f$) that are used to generate hypotheses ($\mathcal{H}$ or $\ddot{\mathcal{H}}$), hypothesized step sequence ($\mathcal{S}$), and final inference (h) are shown in Table 18. The prompt template is applied to both goal inference and action recognition tasks without requiring crafting the prompt again from task to task.

3. Qualitative Results

We present a few more detailed qualitative examples as in Figure 6, 7, and 8 that included detail intermediate outputs along the inference process in the ViDSE framework. We also show a failure example in Figure 9. Best viewed on computer full screen.

Inference Task	ICL Examples
Goal Inference	Based on the description: The person is standing on a stepladder, holding a light bulb in one hand and reaching towards the ceiling fixture with the other. There is a toolbox on the floor, and another light bulb is in his hand. Answer: 1: Replace Ceiling Light Bulb 2: Replace Ceiling Fan Blades 3: Install a Ceiling Medallion 4: Adjust Smoke Detector 5: Paint Ceiling Based on the description: The person is seated at a table covered with a large sheet of white paper. They are holding a heat gun and aiming it at a colorful arrangement of crayon pieces placed along the top edge of the paper. Then, crayon wax is melting and dripping down the paper onto a canvas below. Answer: 1: Make Melted Crayon Art 2: Make Crayon Candles 3: Prepare Crayon Canvas 4: Make a Fresco Painting 5: Paint Bookshelves
Action Recognition	Based on the description: The human is holding a paintbrush or other painting tool, with their arm extended towards a canvas or surface, possibly leaning or sitting in front of it. Answer: 1: Painting 2: Drawing 3: Sketching 4: Coloring 5: Crafting Based on the description: The human is sitting on a bicycle, hands on the handlebars, feet on the pedals, and body leaning forward. Answer: 1: Cycling 2: Biking 3: Wheeling 4: Pedaling 5: Riding

Table 17. ICL examples used in open-vocabulary inference tasks

Inference Task	Prompt
φ_v or φ_f to infer top-K hypotheses, $\mathcal{H}$ / $\ddot{\mathcal{H}}$ or final answer h	I want to perform ⟨TASK NAME⟩ after observing some visual descriptions. ⟨ICL EXAMPLE⟩ Based on the description: $\langle\ \mathcal{D}$ or $\ddot{\mathcal{D}}\ \rangle$ {Based on these options: $\langle\ \mathcal{H} \oplus \ddot{\mathcal{H}} \oplus h_c\ \rangle$} List the most likely ⟨K NUMBER⟩ correct ⟨TARGET⟩ without any explanation. Answer:
φ_l to generate hypothesized steps, $\mathcal{S}$	"Briefly list down the steps to perform $\langle\ \mathcal{H}\ \rangle$. List down in point format without require any specific quantity or unit."

Table 18. Prompt template for LLM used in both goal and action inference tasks. The placeholder ⟨TASK NAME⟩ also denote as ϕ which is replaceable with the specific inference task name (*e.g.* goal inference, action recognition), whereas ⟨ICL EXAMPLE⟩ is for insert the In-Context Learning (ICL) example when infer the hypotheses only, otherwise, it will be empty when not required. The $\langle\mathcal{D}$ or $\ddot{\mathcal{D}}\rangle$ indicate the input of visual descriptions. For {Based on these options: $\langle\mathcal{H} \oplus \ddot{\mathcal{H}} \oplus h_c\rangle$}, it is only applied when there is an option list provided to prompt LLM select the final inference from the hypotheses. The ⟨K NUMBER⟩ is an integer value to control how many hypotheses suppose be inferred. Lastly, the ⟨TARGET⟩ is the term of desired outcome (*e.g.* "action goal" or "ongoing action") to help LLM understand the specific output for the inference task.

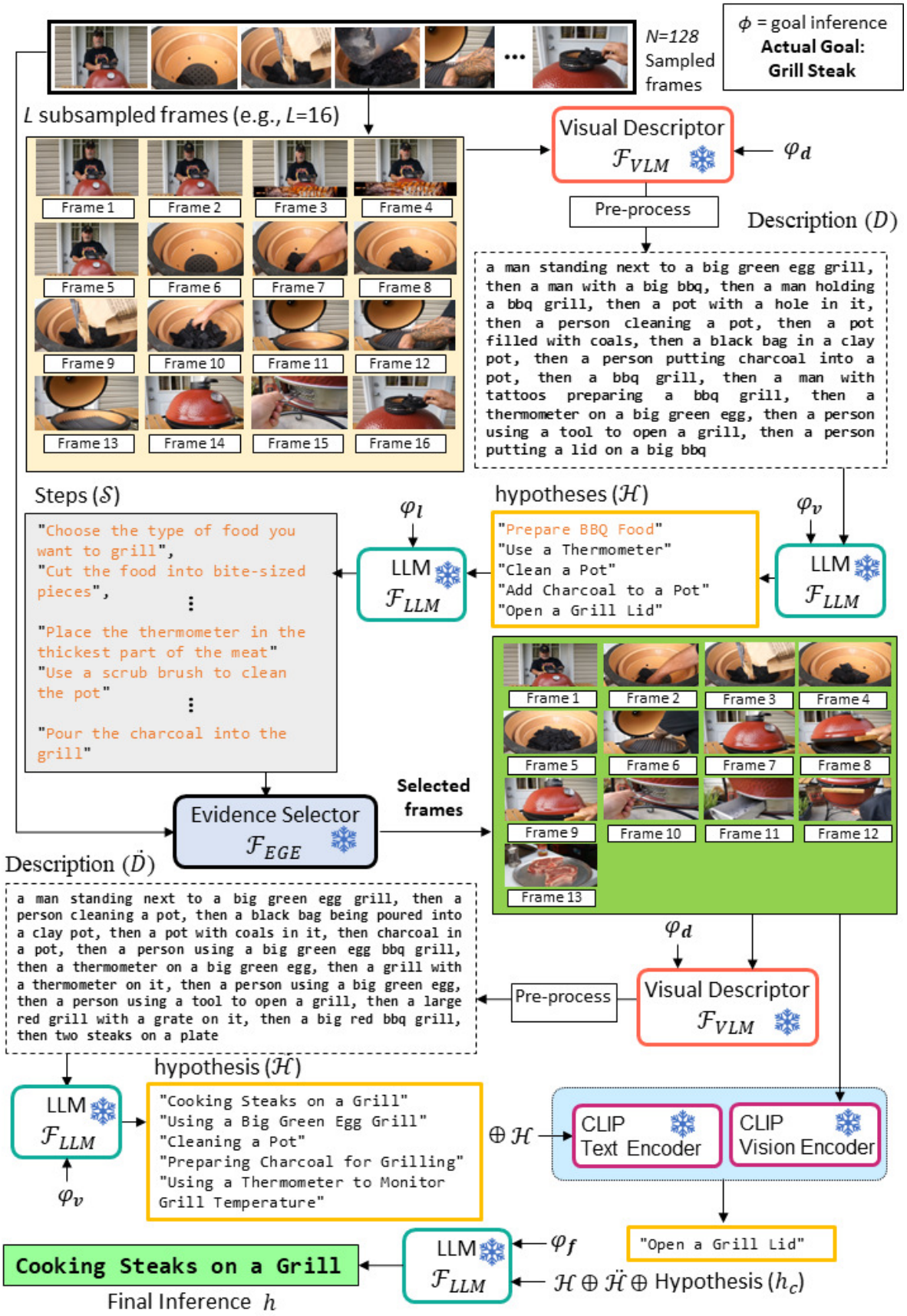


Figure 6. Qualitative example of goal inference by ViDSE (V13B) framework on CrossTask video ($\rho = 50\%$). We demonstrate the frames selection process of the evidence generator which leads to better hypotheses and final inference: "**Cooking Steaks on a Grill**" *vs* ground truth: "**Grill Steak**" (obtain 86.3 SBERT score). We can see the selected frames are more relevant to the grill with charcoal and steak after frame selection process.

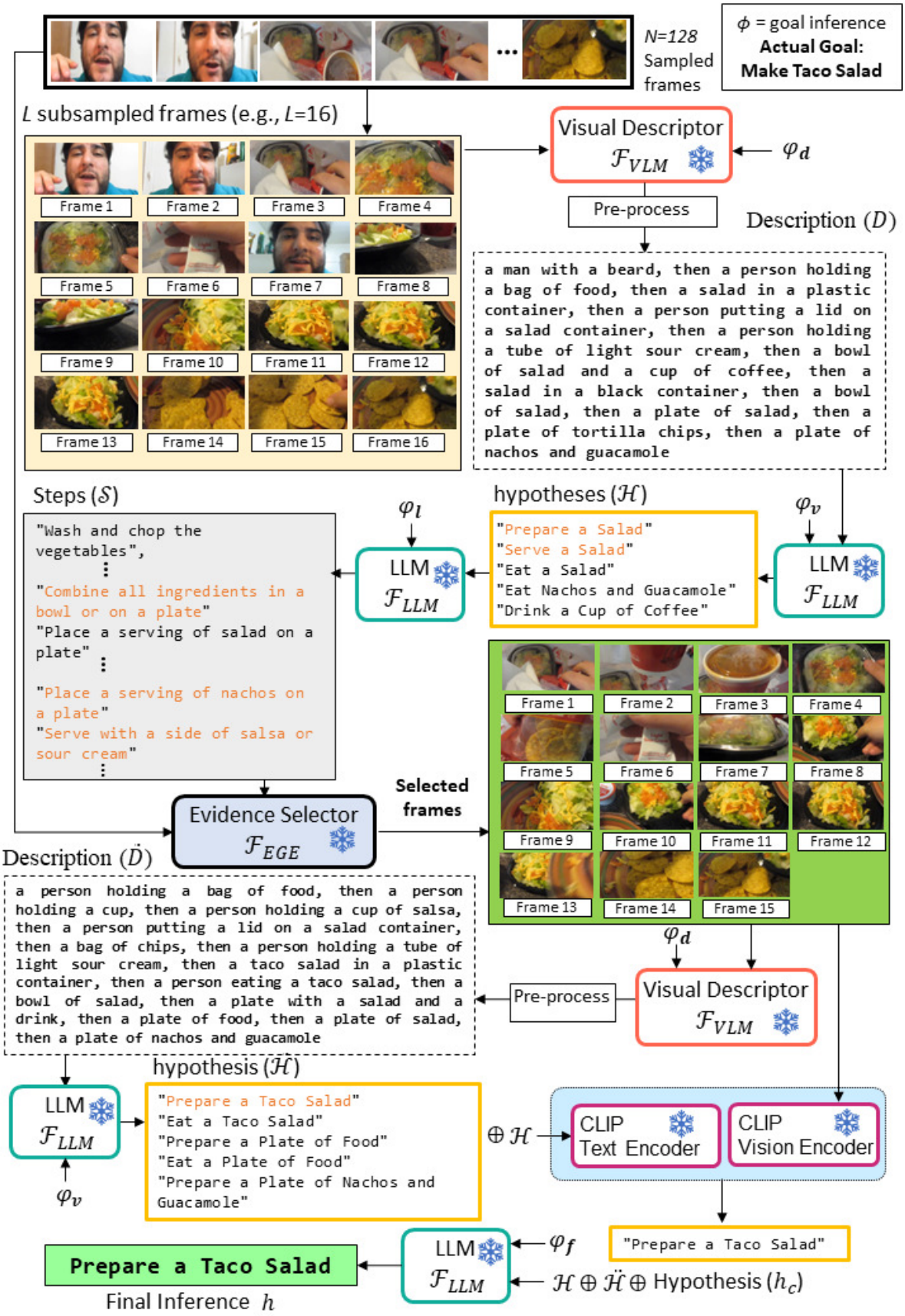


Figure 7. Qualitative example of goal inference by ViDSE (V13B) framework on CrossTask video (ρ = 50%). We can noticed the initial sampled frames that related to a man with beard are filtered out after frame selection process as it is not relevant to the goal. We also can find the inference direction shift from salad only to taco salad related after matching the frames with the hypothesized steps that contained of taco or nachos related steps.

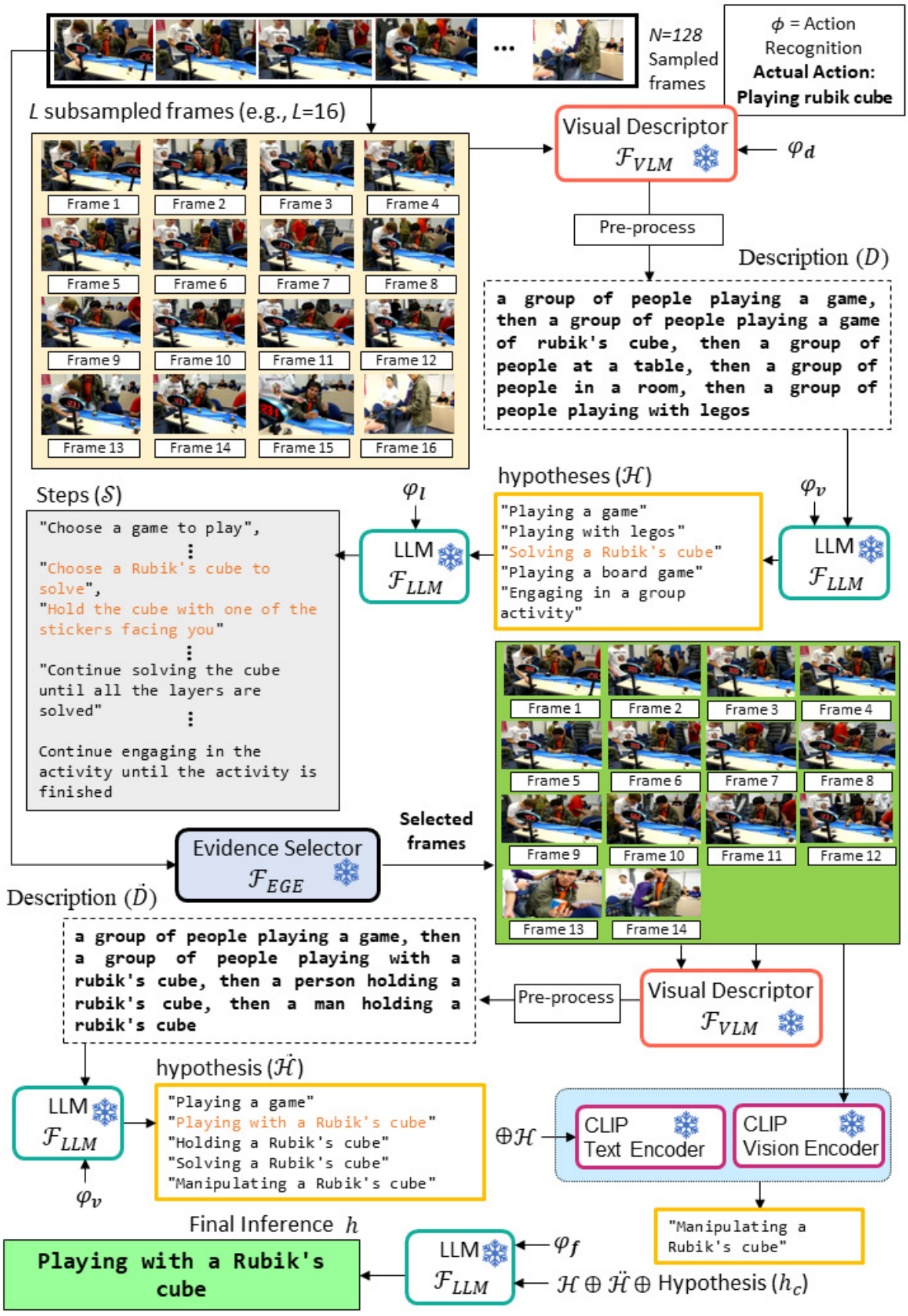


Figure 8. Qualitative example of action recognition by ViDSE (V13B) framework on a video (ρ = 100%) from ActivityNet. Although video action recognition task is more straightforward, it is still challenging when infer on longer untrimmed video that contained many ongoing actions. We can see that initial hypotheses $\mathcal{H}$ is uncertain about the action, whereas $\ddot{\mathcal{H}}$ inference after frame selection process is more certain that the action is related to the Rubik's Cube.

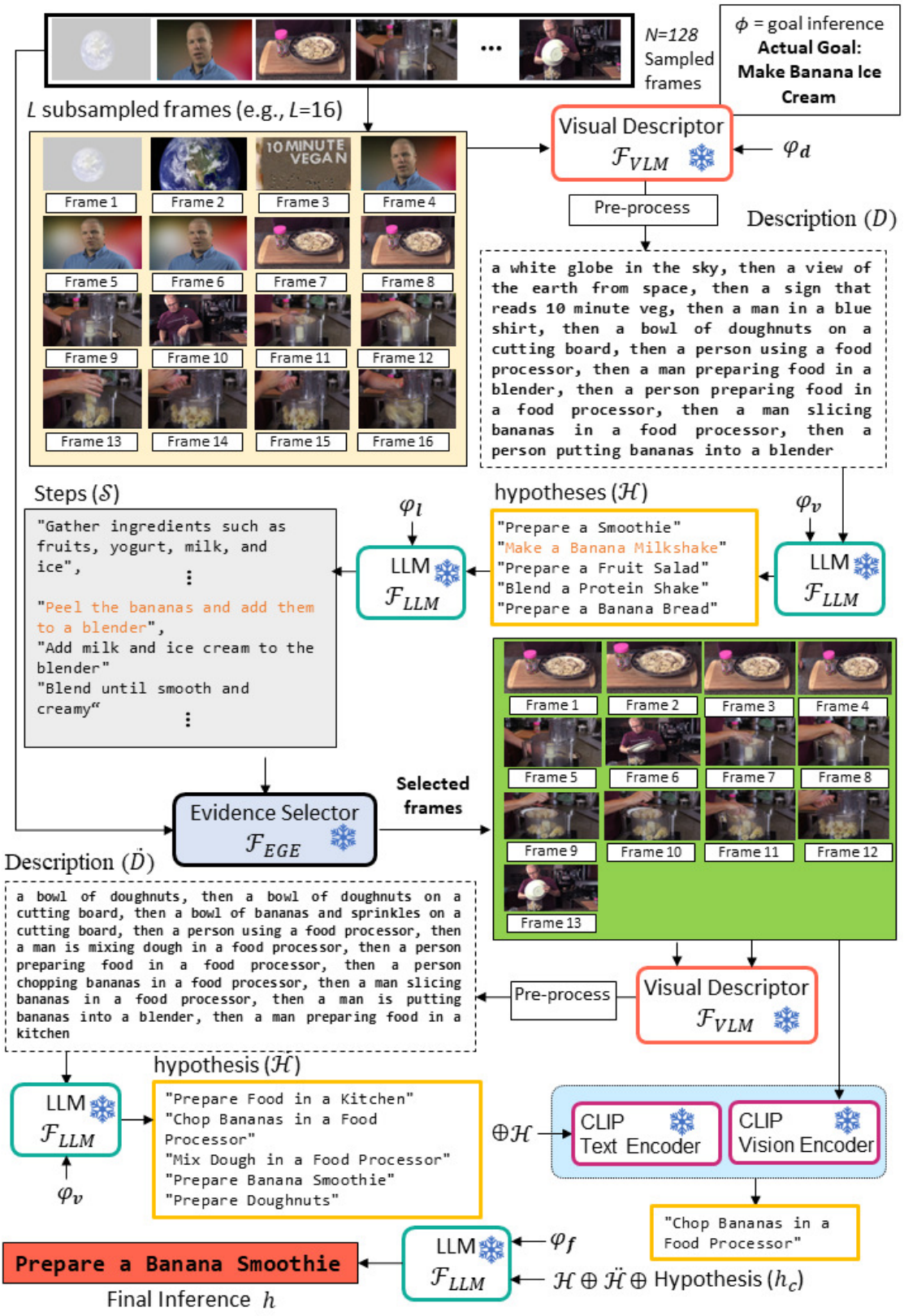


Figure 9. Example of incorrect goal inference by ViDSE (V13B) framework on CrossTask video (ρ = 30%). We can notice that the banana slices in the bowl is wrongly recognized as "doughnuts" in a bowl. This suggests that a visual descriptor with better object-recognizing ability could mitigate this misidentified problem. Moreover, the ice cream related frames are not seen, the LLM is missing this important clue and hence it cannot relate to banana ice cream related goals. We also notice that the frames of *"view of the earth from space"* and *"a man in blue shirt"* are filtered out after frame selection process. This shows that the evidence generator is able to select the frames that are more relevant to the hypotheses.