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 [58], 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.

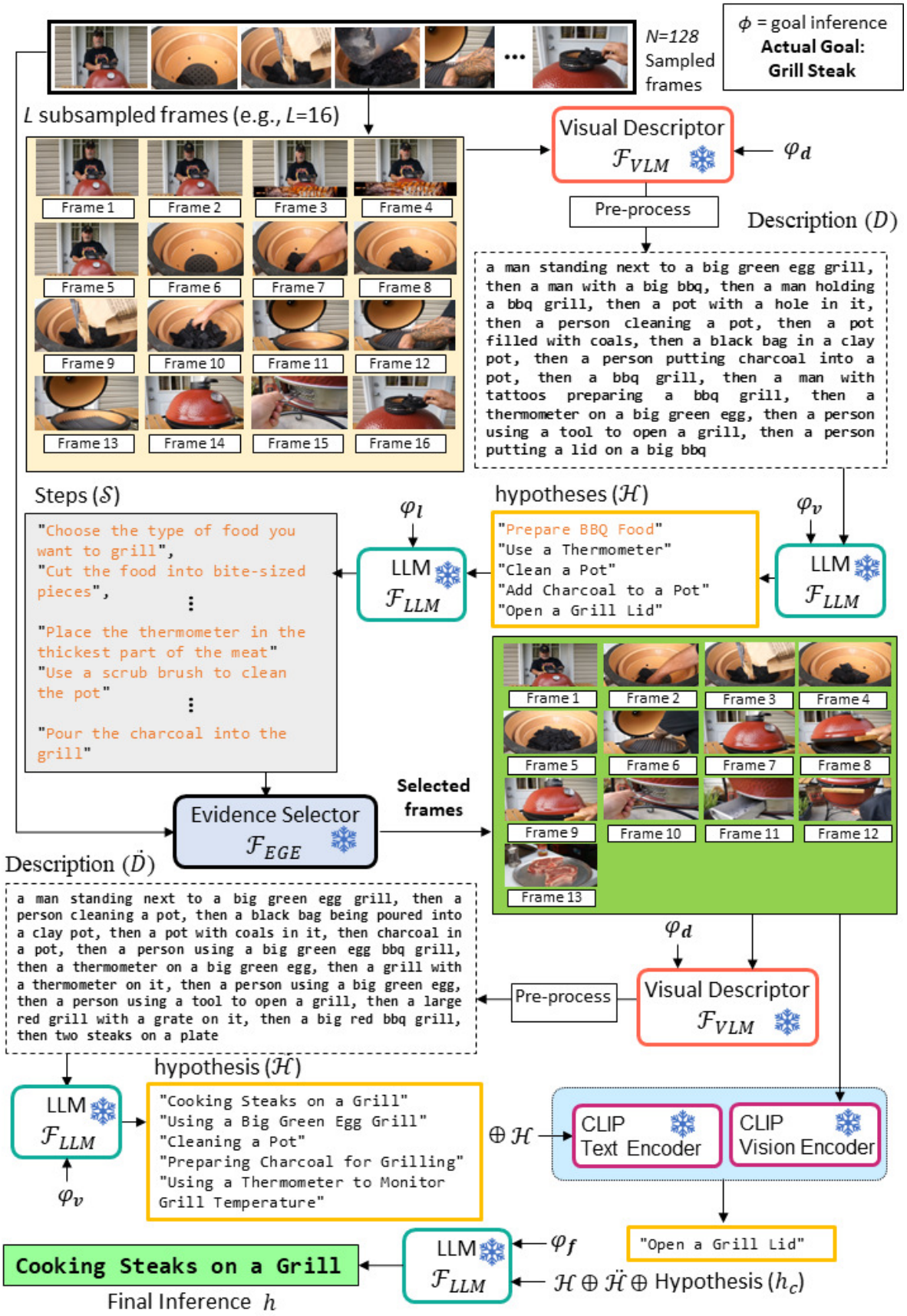

Figure 6. Qualitative example of goal inference by ViDSE (V13B) framework on CrossTask video ($\rho = 50\%$). We demonstrate the frames selection process of the evidence generator which leads to better hypotheses and final inference: "**Cooking Steaks on a Grill**" *vs* ground truth: "**Grill Steak**" (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.

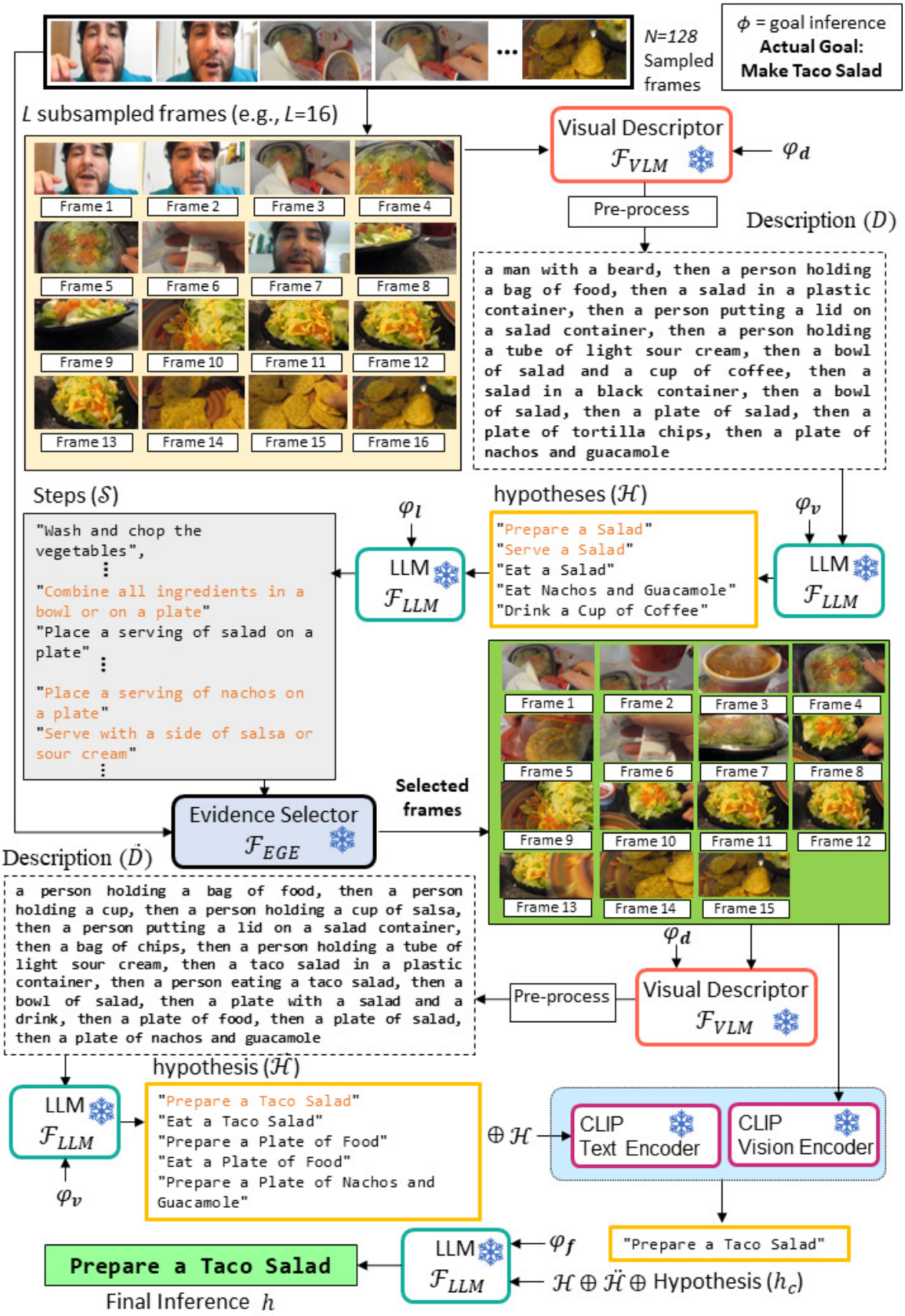

Figure 7. Qualitative example of goal inference by ViDSE (V13B) framework on CrossTask video (ρ = 50%). We can noticed the initial sampled frames that related to a man with beard are filtered out after frame selection process as it is not relevant to the goal. We also can find the inference direction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

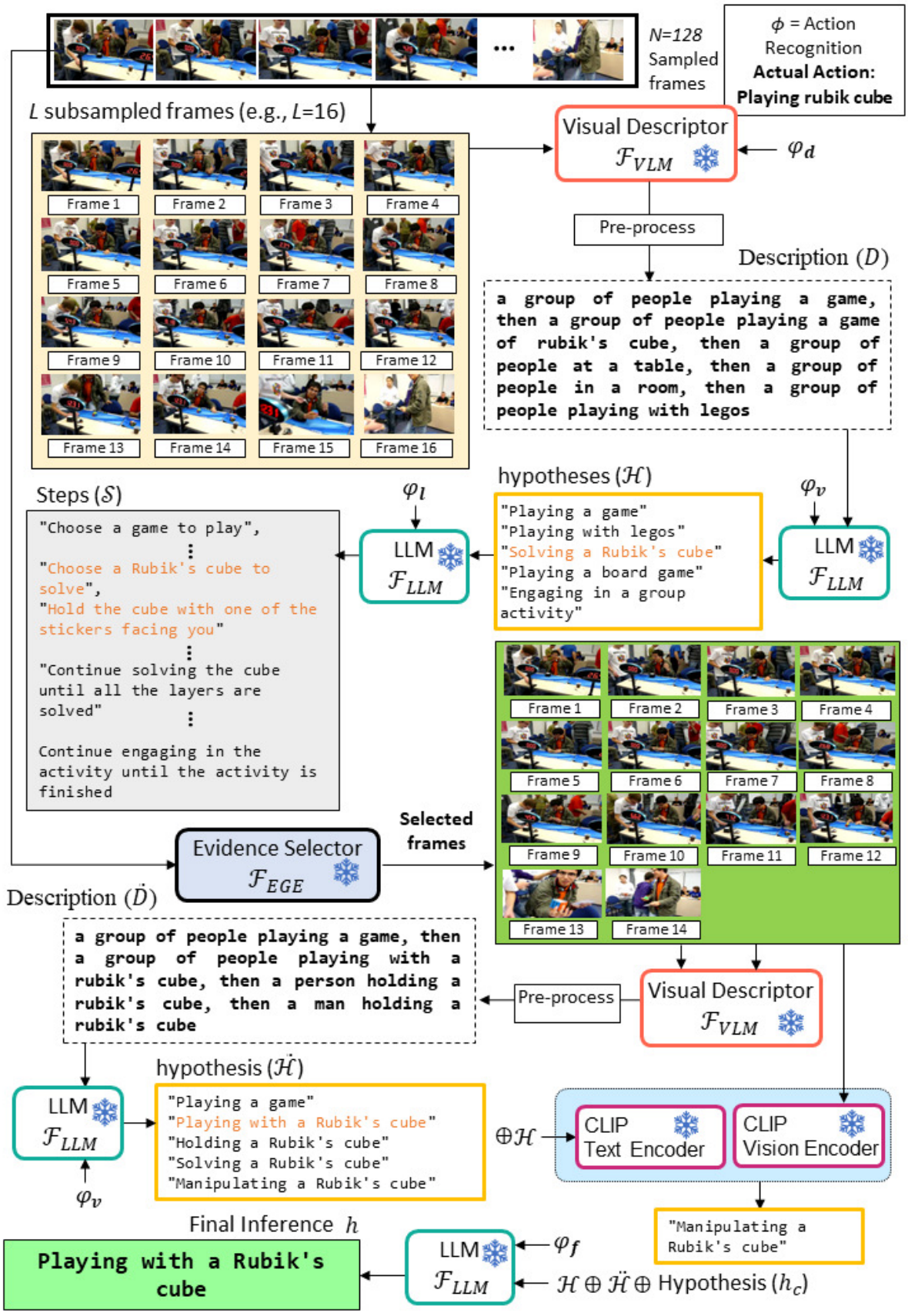

Figure 8. Qualitative example of action recognition by ViDSE (V13B) framework on a video (ρ = 100%) from ActivityNet. Although video action recognition task is more straightforward, it is still challenging when infer on longer untrimmed video that contained many ongoing actions. We can see that initial 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.

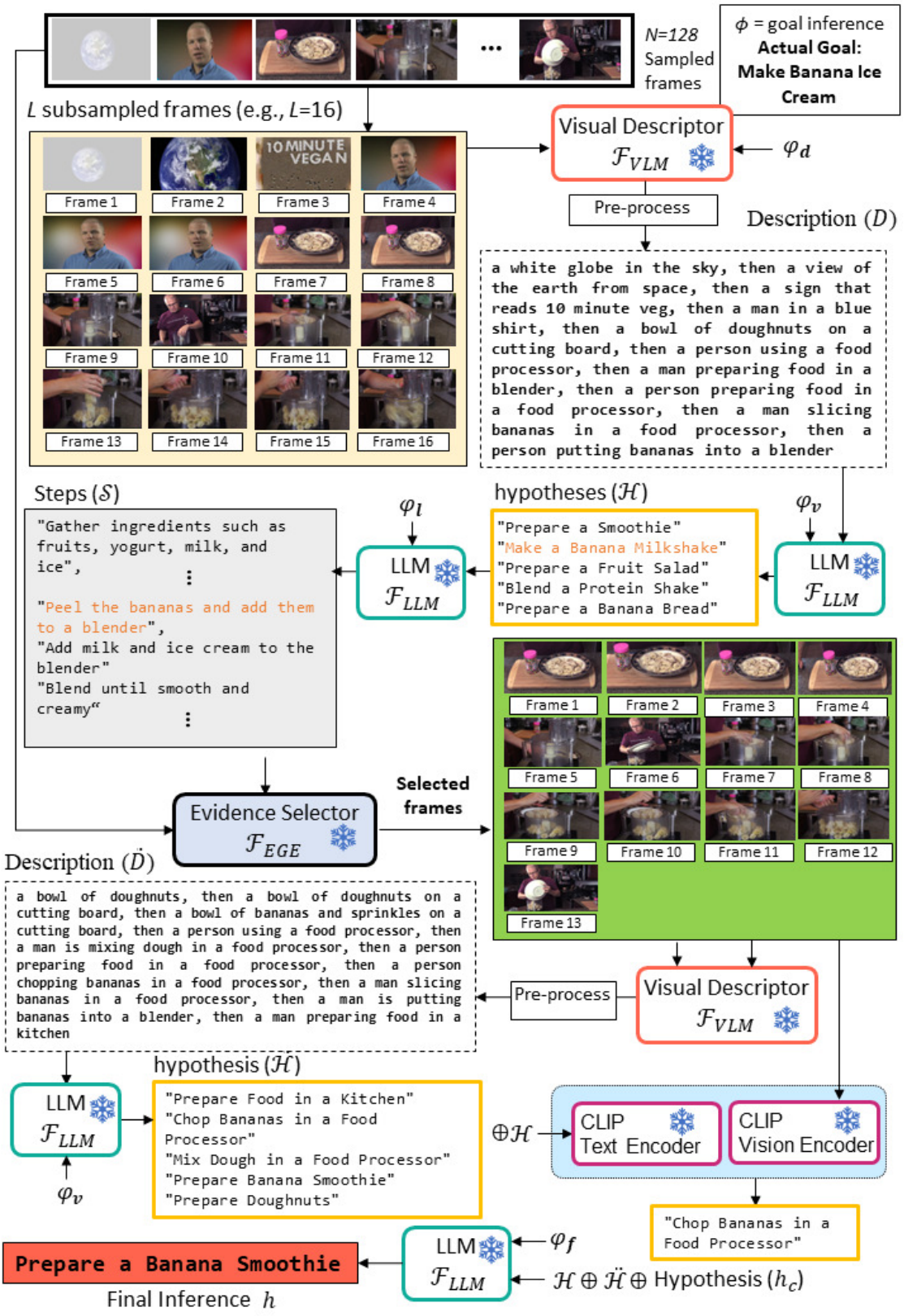

Figure 9. Example of incorrect goal inference by ViDSE (V13B) framework on CrossTask video (ρ = 30%). We can notice that the banana slices in the bowl is wrongly recognized as "doughnuts" in a bowl. This suggests that a visual descriptor with better object-recognizing ability could mitigate this misidentified 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.