Supplementary Material: Propose, Assess, Search: Harnessing LLMs for Goal-Oriented Planning in Instructional Videos

Md Mohaiminul Islam[1,2⋆] Tushar Nagarajan[1] Huiyu Wang[1]
Fu-Jen Chu[1] Kris Kitani[1] Gedas Bertasius[2] Xitong Yang[1]

[1]FAIR, Meta [2]UNC Chapel Hill

Our supplementary material includes additional related works, implementation details, more experimental results, and qualitative results. We will release the code upon acceptance of the paper.

1 Additional Related Works

Several recent studies have introduced various approaches for future action anticipation from videos [2,11]. Additionally, several intriguing methodologies have emerged for procedural planning in instructional videos. Notable among these are prompt-based techniques, including commonsense prompting [6] and multimodal image-text prompting [7]. Another work [4] proposed a condensed action space learning method for procedural planning. In contrast, our work presents a unified framework for different goal-oriented planning tasks (i.e., visual planning for assistance and procedural planning), employing a deliberate propose, assess, and search technique.

Concurrently, there is a growing body of literature exploring reasoning and planning with large language models (LLMs). Various studies leverage LLMs with different tools to tackle complex tasks [2,10,12]. Moreover, there has been significant research on planning in multimodal domains using LLMs [1,2,9]. On the contrary, our approach harnesses LLMs as both a knowledge base and an assessment tool for goal-oriented planning in instructional videos.

2 Implementation Details

Video History Understanding Model for VPA task. We use VideoCLIP [13] as our video history understanding model following the prior work VLaMP [8] to ensure fair comparison. To obtain action steps prediction from the video history, we divide the video into fixed-length windows of 1-second clips. Then, we classify each clip using the pretrained VideoCLIP [13] model. Afterward, we marge the consecutive clips with the same action category as the one step prediction, and thus, we get a sequence of action history from the video.

⋆ Work done during an internship at Meta.

Step Classification Model for PP task. Following the prior work [5], we use a retrieval model to predict the initial and goal action steps from the visual observations. In particular, we utilize a BLIP [3] base model to retrieve the initial and goal steps simultaneously from the provided images. We follow the double retrieval method proposed by [5] and finetune a BLIP [3]-base model following their approach. Please refer to [5] for more details on the retrieval-based step classification for this task.

Combining Value Functions. As described in Section 3.3 of the main paper, we utilize four value functions (text generation score, text mapping score, task graph score, and partial plan evaluation) to guide the breadth-first search of VidAssist to find the optimal action plan. In particular, we utilize a weighted combination of all four functions as the assessment criteria for generated partial plans at each step. We find the optimal weights of each value function from the held-out validation set and use them for the test set. Specifically, we use the following function for the visual planning for assistance task.

$$V^k = 0.2 * V_G^k + 0.1 * V_M^k + 0.1 * V_{TG}^k + 0.7 * V_P^k \tag{1}$$

Here, V_G^k is the text generation score of a particular sample, V_M^k is the text mapping score, V_{TG}^k is the task graph score, V_P^k is the partial plan evaluation score, and V^k is the combined value score. On the other hand, we use the following function for the procedural planning task.

$$V^k = 0.1 * V_G^k + 0.1 * V_M^k + 0.3 * V_{TG}^k + 0.5 * V_P^k \tag{2}$$

3 Additional Quantitative Results

Different Combination of Value Functions. In section 5.3 and Table 6 of the main draft, we ablate the importance of the proposed four value functions-text generation score, text mapping score, task graph score, and partial plan evaluation. In Table 1, we show the performance of using more combinations of value functions. In particular, we compare the performance of combining two and three value functions in both visual planning for assistance and procedural planning tasks. We observe that the proposed LLM-based partial plan evaluation is really important, and the combination of generation score, mapping score and partial plan evaluation leads to the second-highest performance in all metrics. Finally, we notice that the combination of all four value functions leads to the highest performance, demonstrating the importance of the proposed value functions for goal-oriented planning tasks.

4 Qualitative Results

We present qualitative results on both visual planning for assistance (VPA) and procedural planning (PP) tasks.

Table 1: Different Combination of Value Functions. The combination of four value functions leads to the highest performance. We show the success rate for different planning horizons (T) for both tasks.

Generation Score	Mapping Score	Task Graph	Partial Plan	VPA			PP	
				T=1	T=3	T=4	T=3	T=4
✓	✓			41.17	12.73.	6.11	19.21	12.36
✓		✓		42.21	13.69	6.89	18.13	12.41
✓			✓	44.26	15.16	8.59	22.45	14.89
✓	✓	✓		45.66	16.96	9.51	24.38	16.89
✓	✓		✓	50.67	19.91	12.33	27.81	19.03
✓		✓	✓	49.69	19.71	12.01	26.76	18.67
✓	✓	✓	✓	**52.20**	**21.08**	**13.80**	**29.69**	**20.78**

4.1 Qualitative Results on VPA task

We show one example from the COIN dataset in Figure 1 and one example from the CrossTask dataset in Figure 2. In both examples, we observe that while our model successfully predicts the future action plan, the LLM baseline fails to predict the correct action steps. Moreover, we show the propose, assess and search technique of the VidAssist model in Figure 1 (b) and Figure 2 (b). We notice that our model is able to search the optimal action plan from the generated trees utilizing the proposed value scores. For instance, in Figure 1 (b), we observe that the model assigns the highest score to the 'Unscrew the wheel' action in step 1; however, the correct action for step 1 in 'remove old tire'. Nevertheless, the VidAssist rectifies its error in the final step by choosing the highest value path at step 3. Thus, it finds the optimal plan for the particular task. This shows the effectiveness of our search-based technique and the proposed value functions for this particular task.

We also show two failure cases of our model in Figure 3. In Figure 3 (a), we observe that the video history prediction model fails to identify the correct steps. From the provided video history, only one step, 'take out some rice,' was identified, while the user also performed the step 'soak and wash the rice' in the video. Thus, our model predicts 'soak and wash the rice' and 'put rice into the cooker' as future steps. This shows though the error is coming from the video history understanding model, VidAssist still makes valid predictions based on the observed history. This also indicates that the performance of our model can be further enhanced by improving the video history understanding model. On the other hand, in Figure 3 (b), we observe that although the predicted future action steps do not match perfectly with the ground-truth actions, our model still makes reasonable predictions for the particular task.

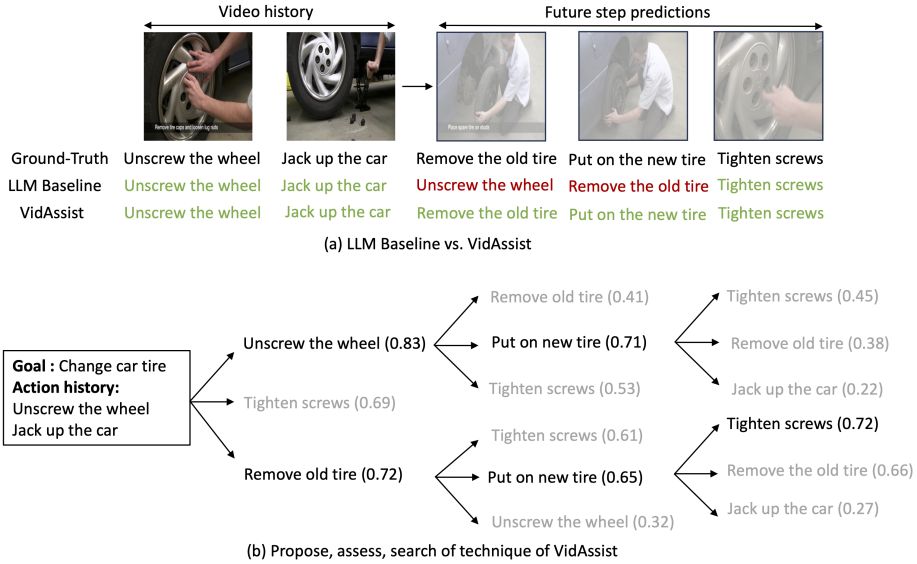
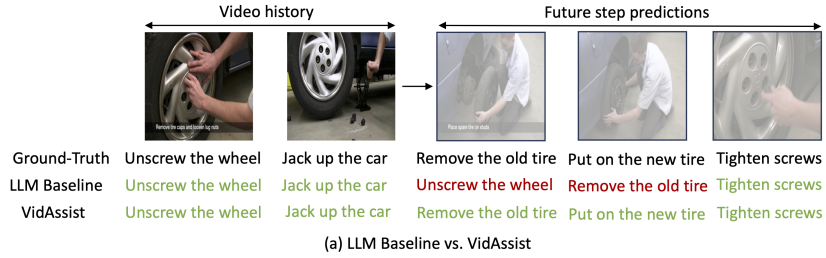


Fig. 1: Example of visual planning for assistance in COIN dataset. (a) VidAssist successfully predicts the future action steps while the LLM baseline fails. (b) Visualization of the proposed search technique with intermediate steps and value scores. We only show three generated actions at each step for brevity and clarity.

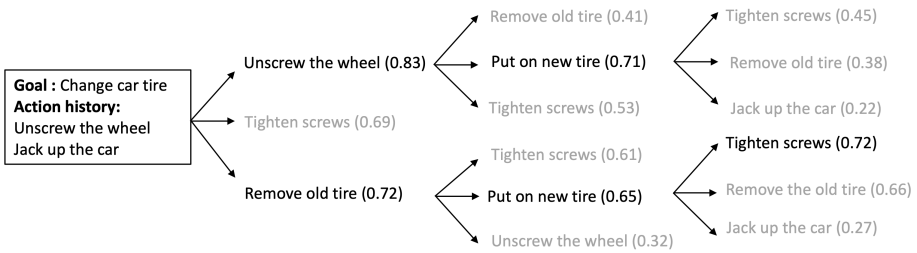
4.2 Qualitative Results on PP task

For the procedural planning task, We show one example from the COIN dataset in Figure 4 and one example from the CrossTask dataset in Figure 5. VidAssist successfully predicts the correct future action plan in both cases while the LLM baseline fails. We also visualize the propose, assess and search technique of the VidAssist model in Figure 4 (b) and Figure 5 (b). We observe that our model is able to search the optimal action plan from the generated trees utilizing the proposed value scores. This demonstrates the effectiveness of our search-based technique and the proposed value functions for the procedural planning task.

Moreover, we present two failure cases of our model in Figure 6. In Figure 3 (a), we observe that the step classification model fails to predict the correct initial step. However, the intermediate steps generated by our model are reasonable considering the predicted initial and goal steps. Therefore, the error stems from the step classification model in this case rather than our LLM-based search technique. This also indicates that the performance of our model can be further enhanced by improving the video step classification model. On the other hand, in Figure 3 (b), we observe that although the predicted intermediate step does not match perfectly with the ground truth, our model still makes a valid prediction for the particular task.

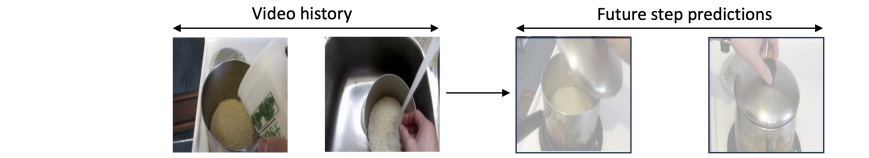


(a) LLM Baseline vs. VidAssist



(b) Propose, assess, search of technique of VidAssist

Fig. 2: Example of visual planning for assistance in CrossTask dataset. (a) VidAssist successfully predicts the future action steps while the LLM baseline fails. (b) Visualization of the proposed search technique with intermediate steps and value scores. We only show three generated actions at each step for brevity and clarity.



(a) Goal: Use rice cooker to cook rice

(a) Goal: Make lemonade

Fig. 3: Examples failure cases of VidAssist in visual planning for assistance. (a) The video history understanding model fails to identify the correct history steps, which leads to the wrong future step predictions. However, our model makes valid predictions based on the predicted action history. (b) VidAssist makes reasonable predictions, though they do not match with the ground truth.

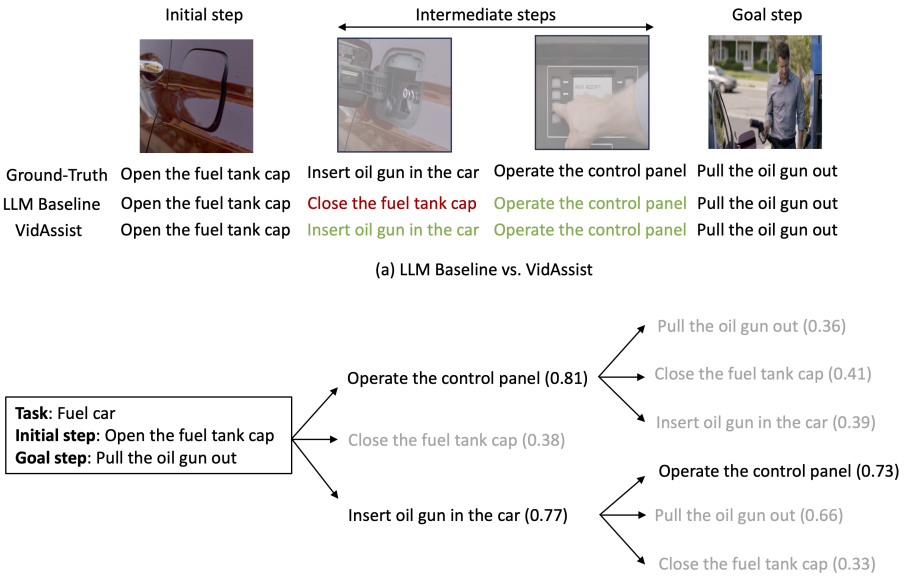


Fig. 4: Example of procedural planning for assistance in COIN dataset. (a) VidAssist successfully predicts the future action steps while the LLM baseline fails. (b) Visualization of the proposed search technique with intermediate steps and value scores. We only show three generated actions at each step for brevity and clarity.

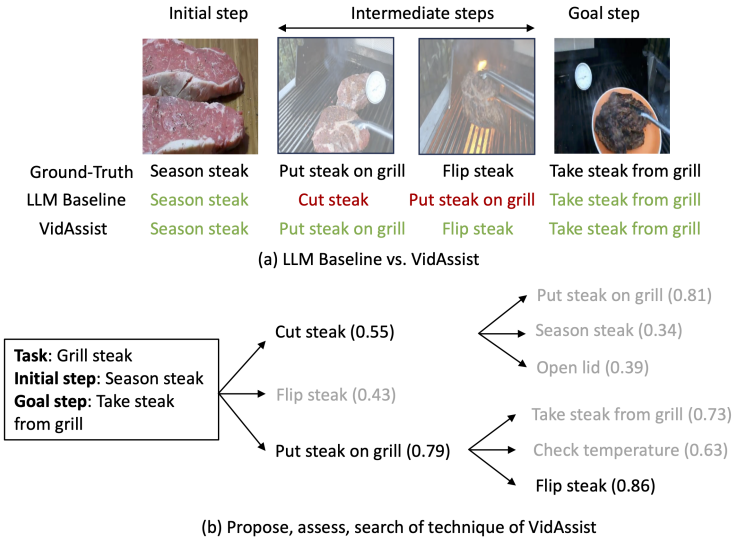


Fig. 5: Example of procedural planning for assistance in CrossTask dataset. (a) VidAssist successfully predicts the future action steps while the LLM baseline fails. (b) Visualization of the proposed search technique with intermediate steps and value scores. We only show three generated actions at each step.

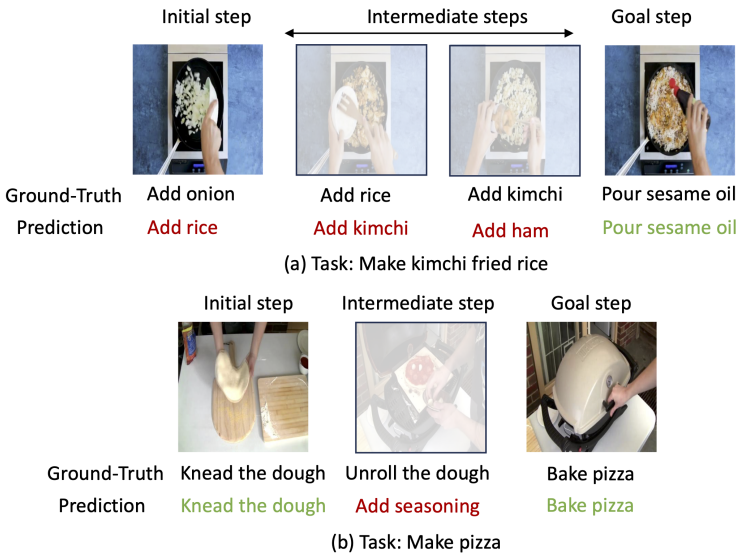


Fig. 6: Examples failure cases of VidAssist in procdural planning. (a) The step classification model fails to identify the correct initial steps, which leads to the wrong intermediate step predictions. However, our model makes valid predictions based on the predicted initial and goal steps. (b) VidAssist makes a reasonable prediction for the intermediate step, though it does not match with the ground truth.

References

1. Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J.B., et al.: Video language planning. arXiv preprint arXiv:2310.10625 (2023) 1
2. Gao, D., Ji, L., Zhou, L., Lin, K.Q., Chen, J., Fan, Z., Shou, M.Z.: Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv preprint arXiv:2306.08640 (2023) 1
3. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022) 2
4. Li, Z., Geng, W., Li, M., Chen, L., Tang, Y., Lu, J., Zhou, J.: Skip-plan: Procedure planning in instructional videos via condensed action space learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10297–10306 (2023) 1
5. Liu, J., Li, S., Wang, Z., Li, M., Ji, H.: A language-first approach for procedure planning. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 1941–1954 (2023) 2
6. Lu, Y., Feng, W., Zhu, W., Xu, W., Wang, X.E., Eckstein, M., Wang, W.Y.: Neuro-symbolic procedural planning with commonsense prompting. arXiv preprint arXiv:2206.02928 (2022) 1
7. Lu, Y., Lu, P., Chen, Z., Zhu, W., Wang, X.E., Wang, W.Y.: Multimodal procedural planning via dual text-image prompting. arXiv preprint arXiv:2305.01795 (2023) 1
8. Patel, D., Eghbalzadeh, H., Kamra, N., Iuzzolino, M.L., Jain, U., Desai, R.: Pretrained language models as visual planners for human assistance. arXiv preprint arXiv:2304.09179 (2023) 1
9. Rose, D., Himakunthala, V., Ouyang, A., He, R., Mei, A., Lu, Y., Saxon, M., Sonar, C., Mirza, D., Wang, W.Y.: Visual chain of thought: Bridging logical gaps with multimodal infillings. arXiv preprint arXiv:2305.02317 (2023) 1
10. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems **36** (2024) 1
11. Sener, F., Yao, A.: Zero-shot anticipation for instructional activities. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 862–871 (2019) 1
12. Sun, S., Liu, Y., Wang, S., Zhu, C., Iyyer, M.: Pearl: Prompting large language models to plan and execute actions over long documents. arXiv preprint arXiv:2305.14564 (2023) 1
13. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021) 1