Using Left and Right Brains Together: Towards Vision and Language Planning

Jun Cen [1 2 3 *] **Chenfei Wu** [2 *] **Xiao Liu** [2] **Shengming Yin** [2] **Yixuan Pei** [4] **Jinglong Yang** [1 5] **Qifeng Chen** [3 †] **Nan Duan** [2 †] **Jianguo Zhang** [1 6 †]

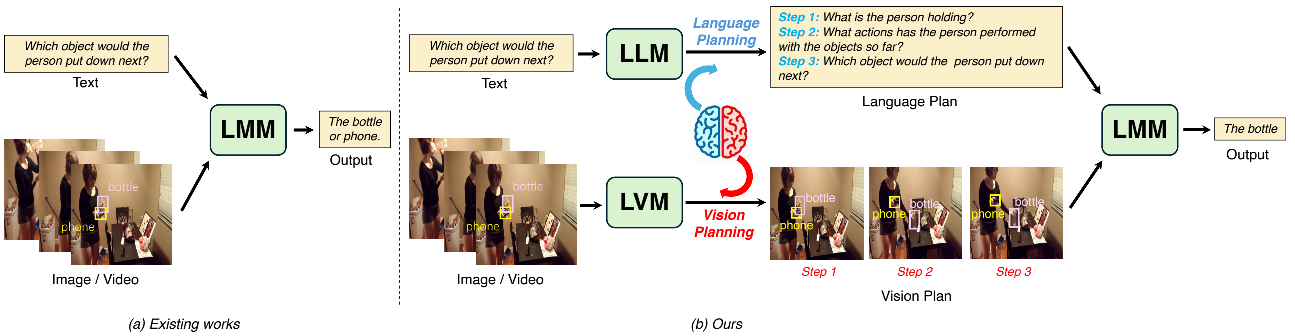


Figure 1. Existing Large Multi-modality Models (LMMs) directly make the decision based on text and image inputs. Our Vision-Language Planning (VLP) framework conducts both language planning and vision planning first, which serves as the left hemisphere and the right hemisphere of a human brain, and then uses an LMM for the final decision making.

Abstract

Large Language Models (LLMs) and Large Multi-modality Models (LMMs) have demonstrated remarkable decision masking capabilities on a variety of tasks. However, they inherently operate planning within the language space, lacking the vision and spatial imagination ability. In contrast, humans utilize both left and right hemispheres of the brain for language and visual planning during the thinking process. Therefore, we introduce a novel vision-language planning framework in this work to perform concurrent visual and language planning for tasks with inputs of any form. Our framework incorporates visual planning to capture intricate environmental details, while language planning enhances the logical coherence of the overall system. We evaluate the effectiveness of our framework across vision-language tasks, vision-only tasks, and language-only tasks. The results demonstrate the superior performance of our approach, indicating that the integration of visual and language planning yields better contextually aware task execution.

[*]Equal contribution [1]Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology [2]Microsoft Research Asia [3]The Hong Kong University of Science and Technology [4]Xi'an Jiaotong University [5]City University of Hong Kong [6]Peng Cheng Lab, Shenzhen, China. †Correspondence to: Qifeng Chen <cqf@ust.hk>, Nan Duan <nanduan@microsoft.com>, Jianguo Zhang <zhangjg@sustech.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

The advent of large-scale auto-regressive text pre-training equips Large Language Models (LLMs) with a powerful ability to conduct sophisticated dialogue and advanced cognitive functions (Brown et al., 2020). Building upon the strong LLMs, plenty of Large Multi-modality Models (LMMs) (Achiam et al., 2023) and agents (Wu et al., 2023) have been developed to address the multi-modality user demands. These LMMs have shown remarkable achievements across various domains, such as robotics (Du et al., 2023a), medical diagnosis (Singhal et al., 2023), and games (Wang et al., 2023a).

Most LMMs incorporate a trainable bridge network designed to align visual features with linguistic representations (Liu et al., 2023), thereby facilitating the processing of both visual and language tokens by a LLM. Recently, language planning such as Chain-of-Thought (CoT) (Wei et al., 2022; Zhang et al., 2023d) has been integrated into LMMs, offering a structured methodology to decompose intricate questions into more tractable components and enabling a sequenced and step-wise reasoning approach. This kind of CoT language planning has been demonstrated to be effective in both few-shot and zero-shot contexts (Wei et al., 2022; Kojima et al., 2022).

Despite the pivotal role of language planning in LMMs, there is a notable shortfall in their capability for vision-based associative reasoning, a process we call vision planning. Language planning alone might result in the generation of responses that are not aligned satisfactorily with the dynamic nature of real-world events, since it is hard to describe the real world with the same granularity and exhaustiveness as visual images by pure language descriptions. In contrast, vision planning could facilitate more realistic reasoning in the form of generating a video that predicts subsequent events using vision inputs. This vision planning is different from the visual branch of current LMMs, which typically maps visual perceptual information into the textual space and still depends on LLMs for linguistic reasoning.

From a cognitive perspective, human cognition relies on a symbiotic operation of the brain's hemispheres, with the left primarily governing language and logical reasoning, and the right hemisphere managing spatial awareness and holistic visual intuition (Gazzaniga, 2005; Corballis, 2014; Zhang et al., 2023b). For instance, when tackling algebraic mathematical challenges, humans often draw upon geometric interpretations to facilitate the reasoning. Current LLMs exhibit functionalities that are akin to the human left hemisphere, specializing in linguistic processing. Yet, they lack the capacity for visual cognition that is intrinsic to the right hemisphere.

Based on the above observations, we propose a Visual-Language Planning (VLP) framework for multi-modality tasks. With respect to language planning, our approach leverages an LLM such as ChatGPT (Brown et al., 2020) to decompose the input text into several steps which are helpful for responding to the overarching inquiry. With respect to vision planning, we employ a Large Vision Model (LVM) such as Stable Video Diffusion (Blattmann et al., 2023a) to generate future video sequences from current images or videos, maximizing the use of visual information for reasoning that aligns with real-world scenarios. For instance, in Fig. 1, by observing the state of a woman drinking water and holding a cellphone, we generate the subsequent videos where the woman is putting down the bottle. Ultimately, our methodology integrates the outcomes of language and vision planning through an LMM and makes the final decision. Our experiments show the effectiveness of our VLP framework across vision-language tasks, vision-only tasks, and language-only tasks.

In summary, our contributions include the following:

- We propose Visual-Language Planning (VLP), a general multi-modal reasoning architecture, which involves not only language planning (serves as left brain) but also vision planning (serves as right brain).
- We implement Visual-Language Processing (VLP) by integrating advanced language generative models, such as ChatGPT, with vision generative models like Stable Video Diffusion, thereby enabling them to collaborate in solving complex problems.
- We show that our VLP not only significantly enhances performance in vision-language tasks but also demonstrates great potential in pure vision and language tasks.

2. Related Work

2.1. Large Multi-modality Models

Large Language Models (LLMs) have exhibited impressive capabilities in conversation and reasoning, owing to extensive auto-regressive pre-training methodologies (Brown et al., 2020; Touvron et al., 2023). Building on the foundation of LLMs, a series of Large Multi-modal Models (LMMs) have been developed, which can process both visual and linguistic inputs (Achiam et al., 2023; Team et al., 2023). The majority of open-source LMMs employ a strategy that aligns visual features with linguistic representations, and conduct visual instruction tuning to improve performance (Liu et al., 2023; Zhu et al., 2023; Team, 2023). These LMMs make decisions based solely on text and image inputs, which constrains their reasoning abilities. In contrast, our VLP framework initially engages in both language and vision planning, analogous to the left and right hemispheres of the human brain, respectively. A LMM is used finally for the final decision-making process.

2.2. Planning with Large Language Models

Most LLMs and LMMs perform planning in the linguistic aspect. The Chain-of-Thought (CoT) approach has been established as an effective technique for prompting LLMs to engage in sequential reasoning (Wei et al., 2022). Zero-shot CoT (Kojima et al., 2022) demonstrates that the prompt "let's think step by step" can enhance the model's output without additional effort. In contrast, few-shot CoT (Wei et al., 2022; Zhang et al., 2023c) employs reasoning templates that guide the LLM through to think in a sequential reasoning format. The recent advent of multi-modal CoT (Zhang et al., 2023d) introduces a two-stage framework that separates rationale generation from answer inference, allowing the latter to fully leverage multi-modal rational information. However, the above works only consider planning in linguistic modality, limiting their capability in visual imagination during planning. Recent studies have employed LMMs in conjunction with video generation models to facilitate task planning in robotics (Du et al., 2023a;b; Ajay et al., 2023), where the video generation model functions as a format of visual planning. However, these works only focus on the robotic domain, limiting the exploration in

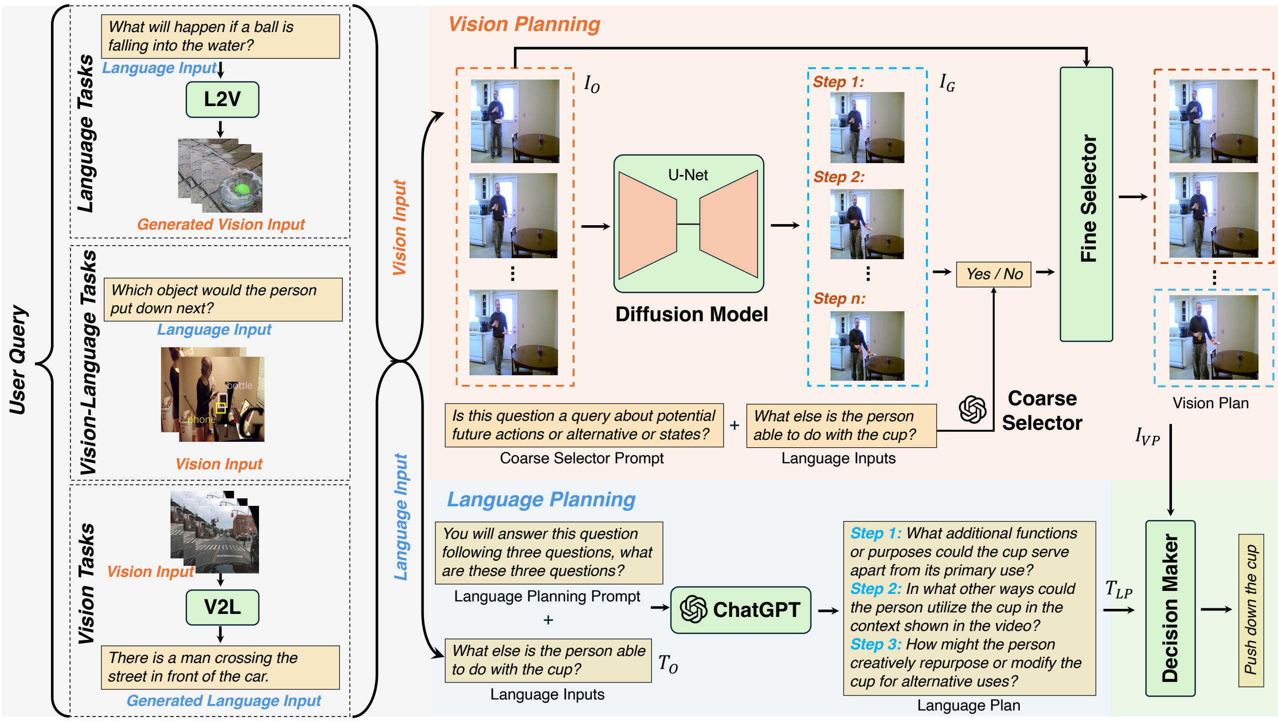


Figure 2. Vision-Language Planning (VLP) Framework. We begin by transforming the user queries into the vision input I_O and language input T_O for tasks of different modalities. Subsequently, the vision planning and language planning are conducted in parallel to obtain the vision plan I_{VP} and language plan T_{LP}. A decision maker then synthesizes these plans to generate the final output.

open-domain scenarios. To address this issue, we design a general-purpose VLP that includes both language planning and vision planning and conduct detailed experiments on a variety of downstream tasks, including vision-language tasks, vision-only tasks, and language-only tasks.

2.3. Video Generation

Initial video generation methodologies (Tulyakov et al., 2018; Skorokhodov et al., 2022; Lu et al., 2022; Wang et al., 2023d) utilized generative adversarial networks (GANs) (Goodfellow et al., 2020), yet they were limited in producing high-quality videos (Blattmann et al., 2023b). The advent of diffusion models (Rombach et al., 2022), characterized by their stable training process and superior generative capabilities, has led to their adoption in contemporary video generation techniques (Ho et al., 2022; Blattmann et al., 2023a; Luo et al., 2023; Yin et al., 2023; Zhang et al., 2023a). Among these, Stable Video Diffusion (Blattmann et al., 2023a) has gained recognition for its robust text-to-video and image-to-video generation capabilities across various domains. DMVFN (Hu et al., 2023) tailors video generation to specific applications, such as autonomous driving, by operating on video inputs. Meanwhile, MCVD (Voleti et al., 2022) innovatively masks and reconstructs video frames, facilitating video prediction and interpolation. In our Visual Language Processing (VLP) framework, we integrate a video generation model to augment the visual aspect of the reasoning process.

3. Vision-Language Planning

3.1. Framework Overview

As shown in Fig. 2, Our VLP system handles user queries of different modalities, including pure language tasks, pure vision tasks, and vision-language tasks. For pure language tasks, a Language-to-Vision (L2V) model is used to convert language queries to corresponding visual content, such as images or videos. Conversely, for pure vision tasks, relevant language descriptions are produced using a Vision-to-Language (V2L) model. Therefore, whatever modalities the user queries are, our approach enables the acquisition of both vision input I_O and language input T_O.

The vision input I_O undergoes processing by the vision planning branch to yield the vision planning outcomes I_{VP}. A Vision Planning Generator (VPG) is employed to synthesize future frames that constitute the vision plan based on the current frames, followed by the use of a Vision Planning Selector (VPS) which contains a coarse selector and a fine selector to choose frames that are potentially beneficial for the current task. The language input T_O is processed by an

LLM to produce the language plan T_{LP}. Finally, a Decision Maker which is a LMM takes both the vision plan I_{VP} and language plan T_{LP} into consideration and makes the final decision.

3.2. Vision Planning

Vision Planning Generator (VPG). The vision input is denoted as $I_O = \{I_O^1, I_O^2, ..., I_O^N\}$, where N represents the number of input images. $N = 1$ means we input an image and $N > 1$ means the vision input is a video. Then a Vision Planning Generator (VPG) G is applied to generate the future frames I_G:

$$I_G = G(I_O), \tag{1}$$

where $I_G = \{I_G^1, I_G^2, ..., I_G^n\}$ and n denotes the number of generated images or vision planning steps. The video diffusion model G is an image-to-video model if the input is an image ($N = 1$), and G is a video prediction model if the input is a video ($N > 1$).

Vision Planning Selector (VPS). Although VPG generates potentially useful future frames, directly using them may cause the following issues: 1) We notice that not all problems are related to the future states, in which case the inclusion of generated frames could introduce irrelevant noise. 2) Besides, the video generation model's limitations might result in artifacts and superfluous frames within the generated content. To address above issues, we employ a Vision Planning Selector (VPS) comprising two modules: 1) Coarse Selector (CS) to determine whether the current task needs the generated video frames or not. 2) Fine Selector (FS) to determine which frames should be selected to help solve problems if current task requires generated frames.

For the Coarse Selector (CS), we simply add the prompt *[Is this question a query about potential future actions or alternative or states?]* to ChatGPT, so that it will output Yes or No to judge if the language query T_O should use the generated frames or not.

For the Fine Selector (FS), it selects the useful frames for the query T_O among original inputs I_O and generated frames I_G as the ultimate vision plan I_{VP}. FS takes a video as the input, and assigns selection scores for each frame, so that we can select the frames with top-K highest scores to form the final vision plan I_{VP}. Specifically, for each frame, FS first extracts visual features by a CLIP vision encoder (Radford et al., 2021). Then visual query features are generated by a Q-former (Li et al., 2023) and concatenated with the text prompt like *[Does the information within the frame provide the necessary details to accurately answer the given question]* (Yu et al., 2023). Finally, a LLM takes the visual and text tokens as inputs and we use the output probability of the token "Yes" as the selection score for the frame.

In summary, I_{VP} can be formed as:

$$I_{VP} = \begin{cases} \mathit{FS}(\mathit{Concate}(I_O, I_G)),\ if\ \mathit{CS}(T_O) =\ \mathit{Yes}, \\ \mathit{FS}(I_O),\ if\ \mathit{CS}(T_O) =\ \mathit{No}. \end{cases} \tag{2}$$

The vision plan I_{VP} may contain the generated future frames which could provide additional useful information for the user query.

3.3. Language Planning

In our language planning branch, we implement the zero-shot chain-of-thought (Kojima et al., 2022) technique to decompose the language input into a series of sub-steps, forming the language plan T_{LP}. We use the prompt like *[Imagine that you are trying to answer a Video Q&A Multi-choice Question. You will firstly watch a video and then answer this question. Question here. You will answer this question following three questions, what are these three questions?]*. ChatGPT will answer 3 steps which could help the decision maker to think step by step and make the decision according to these sub-questions.

For instance, the user asks *[What else is the person able to do with the cup?]* in Fig 2. To answer this question, ChatGPT generates the three-step language plan including *[what additional functions could the cup serve, in what other ways could the person utilize the cup, and how might the person repurpose the cup]*. These language plan steps provide complementary information for the initial query and guide the following decision maker to give the final answer from different perspectives.

3.4. Decision Maker

The decision maker is a LMM and is responsible for making the final output according to the vision plan I_{VP} and language plan T_{LP}. We design a multi-round conversation strategy to guide the LMM to think sequentially. 1) Vanilla Answering. We directly give the original vision inputs I_O and language inputs T_O to LMM, and prompt LMM to give the vanilla answer. 2) Language Answering. For language plan T_{LP}, we first let the LMM answer three language steps one by one, and then give the answer for the original query T_O based on the answers of all steps. 3) Vision Answering. We prompt LMM to give the answer using generated vision plan I_{VP} and the original query T_O. 4) Voting. We propose a voting mechanism to strengthen the vanilla answering by the language answering and vision answering, since they provide the alternatives from different modality reasoning perspectives. LMM will evaluate the validity again between the vanilla answer and language answer or vision answer to obtain the voted language answer and vision answer, and finally make the ultimate decision between these two voted answers. See Fig. 5 for an example.

Table 1. Results on video question answering.

Model (# Frames)	STAR					NExT-QA			
	Int.	Seq.	Pre.	Fea.	Avg.	Tem.	Cau.	Des.	Avg.
ViperGPT (dense/1fps) (Surís et al., 2023)	-	-	-	-	-	-	-	-	60.0
Flamingo-80B (30) (Alayrac et al., 2022)	-	-	-	-	39.7	-	-	-	-
VFC (32) (Momeni et al., 2023)	-	-	-	-	-	45.4	51.6	64.1	51.5
InternVideo* (8) (Wang et al., 2022)	43.8	43.2	42.3	37.4	41.6	43.4	48.0	65.1	49.1
BLIP-2^{voting} (4) (Li et al., 2023)	41.8	39.7	40.2	39.5	40.3	59.1	61.3	74.9	62.7
BLIP-2^{concat} (4) (Li et al., 2023)	45.5	41.8	41.8	40.0	42.2	59.7	60.8	73.8	62.4
SEVILA (4) (Yu et al., 2023)	48.3	45.0	44.4	40.8	44.6	**61.3**	<u>61.5</u>	<u>75.6</u>	<u>63.6</u>
LLAVA (4) (Liu et al., 2023)	<u>49.0</u>	<u>47.3</u>	<u>45.5</u>	<u>47.8</u>	<u>47.4</u>	55.7	60.6	74.3	61.1
VLP (4 + 1 (Generated Frame))	**52.0**	**50.1**	**50.8**	**49.0**	**50.5**	<u>60.5</u>	**63.7**	**76.7**	**64.7**

Table 2. Results of Video Captioning on the BDD-X dataset. 'B', 'C', and 'M' refer to BLEU-4, CIDEr, and METEOR, respectively.

Method	B	C	M
S2VT (Venugopalan et al., 2015)	30.2	179.8	27.5
S2VT++ (Venugopalan et al., 2015)	27.1	157.0	26.4
SAA (Kim et al., 2018)	31.8	214.8	29.1
WAA (Kim et al., 2018)	32.3	215.8	29.2
ADAPT (Jin et al., 2023)	34.6	247.5	30.6
VLP (Ours)	**35.7**	**256.7**	**31.1**

Table 3. Ablation study of VP and LP on STAR dataset.

Model	Int.	Seq.	Pre.	Fea.	Avg.
LLAVA (Liu et al., 2023)	49.0	47.3	45.5	47.8	47.4
LLAVA+VP (Ours)	51.5	<u>49.9</u>	50.0	47.1	49.6
LLAVA+LP (Ours)	**52.3**	**50.1**	**51.1**	<u>48.2</u>	<u>50.4</u>
LLAVA+LP+VP (Ours)	<u>52.0</u>	**50.1**	<u>50.8</u>	**49.0**	**50.5**

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our VLP on various scenarios, covering the open-domain scenario (STAR (Wu et al., 2021) and NExT-QA (Xiao et al., 2021)), autonomous driving scenario (BDD-X (Kim et al., 2018)), and robotics operation scenario (BAIR (Ebert et al., 2017)). The dataset details are as following:

- *STAR.* Situated Reasoning in Real-World Videos (STAR) dataset (Wu et al., 2021) comprises 60k situated reasoning questions accompanied by programs and answers, 24k candidate choices, and 22k trimmed situation video clips. It covers four types of questions: interaction, sequence, prediction, and feasibility, in which prediction and feasibility questions are strongly related to what will happen next. We evaluate the accuracy of the multiple-choice questions.
- *NExT-QA.* NExT-QA (Xiao et al., 2021) comprises 5440 videos, each with an average duration of 44 seconds. It includes approximately 52k manually annotated question-answer pairs, categorized into causal (48%), temporal (29%), and descriptive (23%) questions. We evaluate the accuracy of the multiple-choice questions.
- *BDD-X.* BDD-X (Kim et al., 2018) is a textual autonomous driving dataset. It annotates the descriptions and actions of 77 hours within 6,970 videos from BDD dataset (Xu et al., 2017). The video captioning performance is evaluated by the BLEU-4 score, CIDEr score, and METEOR score. The actions include the course and speed, and we use root mean squared error (RMSE) and a tolerant accuracy (A_σ) (Jin et al., 2023) to measure the acition prediction performance.
- *BAIR.* BAIR dataset (Ebert et al., 2017) records 30k videos of a robot pushing multiple objects. The X, Y, and Z location of the robot gripper is provided for each frame, and we use root mean squared error (RMSE) for evaluation. We conduct prediction and planning tasks on the BAIR dataset. In the prediction task, we aim to predict the location of the robot gripper of the subsequent frames given the preceding ones. In the planning task, we provide the initial frames and goal (end) frames and predict the locations of the robot gripper between them.

Implementation Details. We build our VLP framework based on the existing baselines, which refer to using a single Decision Maker without VP and LP. Existing baseline Decision Makers are distinct for different tasks and datasets, and we implement our VLP on top of them.

For STAR and NExt-QA dataset, we use Stable Video Dif-

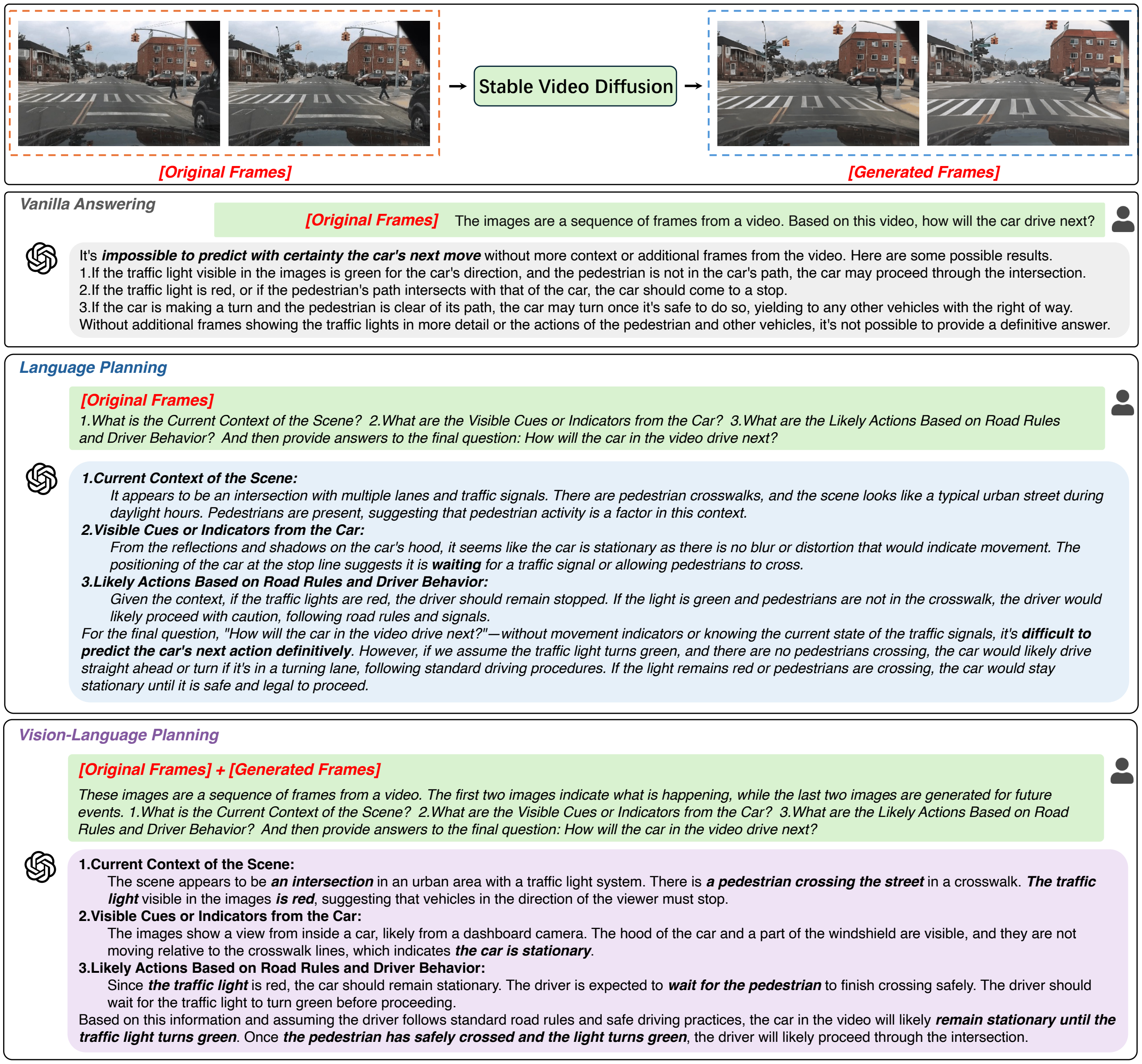


Figure 3. VLP with GPT4-V on vision-related tasks. Languege planning helps the GPT4-V to consider the information from different perspectives and reason sequentially, and vision planning provides the possible future states which are essential for the ultimate decisions.

fusion (Blattmann et al., 2023a) model for future frames generation. The Visual Planning Selector (VPS) contains a ChatGPT (Coarse Selector) and a BLIP-2 (Fine Selector) (Li et al., 2023; Yu et al., 2023), and the detail is explained in Sec. 3.2. The decision maker is based on a LLAVA model (Liu et al., 2023), as shown in Fig. 5. We conduct zero-shot experiments without finetuning LLAVA.

For BDD-X and BAIR datasets, we use DMVFN (Hu et al., 2023) and MCVD (Voleti et al., 2022) for video generation. The Fine Selector of VPS is a temporal transformer following (Wang et al., 2023b), and the decision maker is a BERT model (Devlin et al., 2018) based on ADAPT (Jin et al., 2023). We train the decision maker and Fine Selector end-to-end on BDD-X and BAIR datasets, and we follow the training details of ADAPT (Jin et al., 2023).

4.2. Results

Results on Video Multiple Choice Questions Datasets STAR and NexT-QA. Video multiple choice questions are typical vision-language tasks, which require the model to have a deep understanding of both the videos and questions. The corresponding results are shown in Table 1. It shows that our VLP achieves the best performance among all LMM-based baselines, including the state-of-the-art method

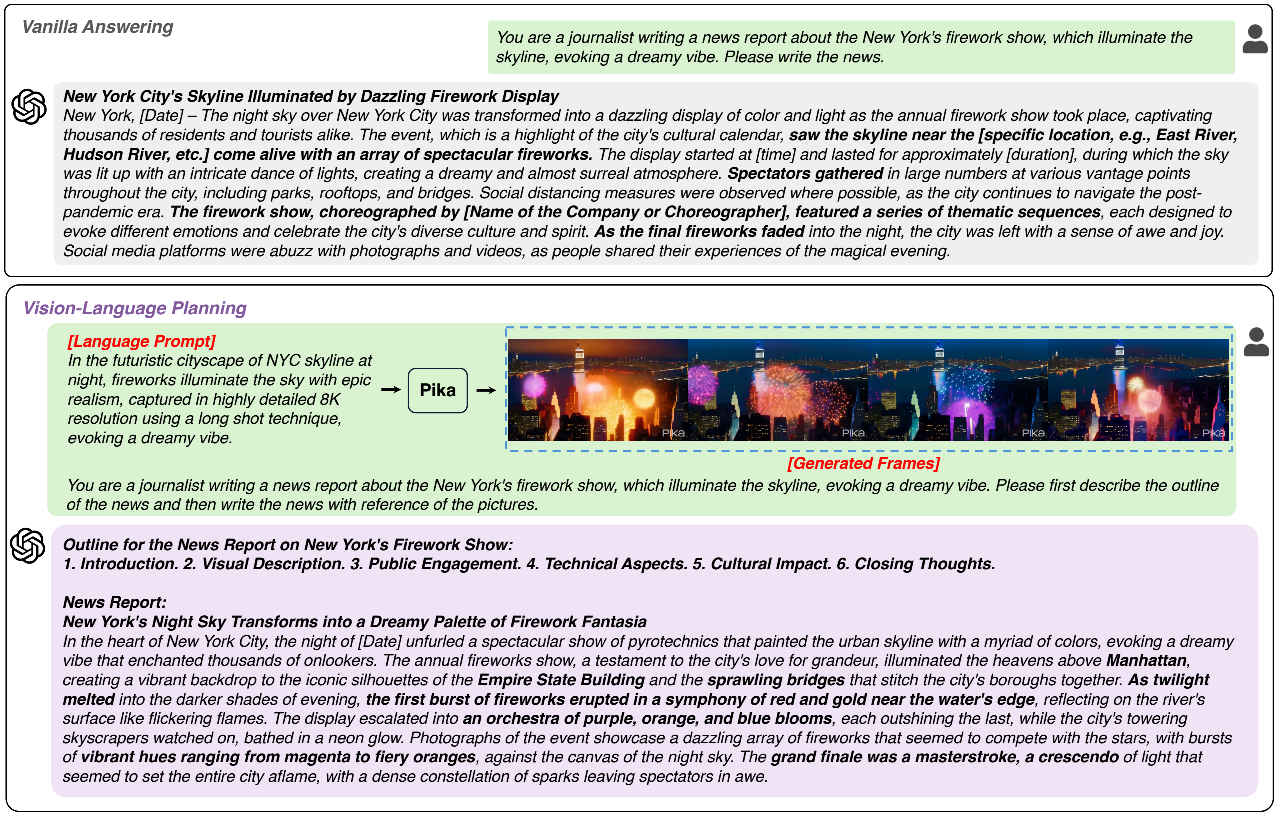


Figure 4. VLP with GPT4-V on language-only tasks. GPT4-V gives more detailed and vivid descriptions using the generated videos.

Table 4. Ablation study of VP and LP on BDD-X dataset.

Method	B	C	M
ADAPT (Jin et al., 2023)	34.6	247.5	30.6
ADAPT+LP (Ours)	35.2	242.6	30.8
ADAPT+VP (Ours)	35.7	**256.7**	**31.1**
ADAPT+VP+LP (Ours)	**36.2**	251.7	30.6

Table 5. Ablation study of VPS (including CS and FS) on STAR.

Model	Int.	Seq.	Pre.	Fea.	Avg.
VP w/o CS	49.6	47.7	48.6	45.9	48.0
VP w/o FS	51.4	**50.3**	38.1	42.4	45.6
VP	**51.5**	49.9	**50.0**	**47.1**	**49.6**

SEVILA and our implemented baseline LLAVA. Table 1 illustrates the effectiveness of our VLP in the open-domain scenario.

Results on Video Captioning Datasets BDD-X. Video captioning is a vision-only input task. Table 2 shows that our VLP surpasses the state-of-the-art method ADAPT with a clear margin.

Case Study with GPT4-V. We cannot conduct quantitative

Table 6. Ablation study of voting in decision maker on STAR.

Model	Int.	Seq.	Pre.	Fea.	Avg.
VP w/o voting	51.4	**50.4**	48.1	43.3	48.3
VP	**51.5**	49.9	**50.0**	**47.1**	**49.6**
LP w/o voting	48.3	49.8	44.2	42.9	46.3
LP	**52.3**	**50.1**	**51.1**	**48.2**	**50.4**

experiments using GPT-4V due to the usage limit restrictions per day. Instead, we provide two case studies to demonstrate the effectiveness of VLP with GPT4-V. Fig. 3 shows that vanilla answering cannot give the results with current videos (*[It's impossible to predict with certainty the car's next move]*). Language planning provides sequential reasoning steps but still *[difficult to predict the next action]*. With the generated future frames from vision planning, which shows the pedestrian is crossing the road, GPT4-V gives the correct answer that the car should *[remain stationary]* and move *[once the pedestrian has safely crossed and the light turns green]*. So vanilla answering and language planning give vague response and VLP gives the certain and correct response. Fig. 4 shows that our VLP generates more detailed and vivid descriptions based on the generated future frames in language-only tasks. For example, VLP generates

Table 7. Control Signals Prediction Accuracy on BDD-X dataset.

Method	Course						Speed					
	RMSE(degree)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑	RMSE(m/s)↓	$A_{0.1}$ ↑	$A_{0.5}$ ↑	$A_{1.0}$ ↑	$A_{5.0}$ ↑	$A_{10.0}$ ↑
Single	6.3	8.3	84.7	**90.5**	97.2	98.7	3.4	5.0	25.5	37.8	86.8	98.7
ADAPT	6.4	62.2	85.5	89.9	97.2	**98.8**	2.5	11.1	28.1	45.3	94.3	99.5
ADAPT + VP	**6.2**	**65.5**	**86.2**	90.3	**97.3**	**98.8**	**2.3**	**16.1**	**35.3**	**51.8**	**95.2**	**99.6**

Table 8. Action Prediction (2+0, 4+0) and Planning (1+1, 1+2) RMSE(cm) on BAIR. i and e refer to initial and end (goal) frames.

# Inputs	Method	X	Y	Z	Sum
2 (i) + 0 (e)	Baseline	8.75	7.24	3.86	19.85
	Baseline + VP	**8.68**	**6.83**	**3.84**	**19.36**
4 (i) + 0 (e)	Baseline	8.06	6.70	**3.63**	18.39
	Baseline + VP	**7.72**	**6.47**	3.68	**17.86**
1 (i) + 1 (e)	Baseline	5.74	5.67	3.42	14.83
	Baseline + VP	**5.48**	**5.46**	**3.40**	**14.34**
1 (i) + 2 (e)	Baseline	5.54	**5.45**	3.41	14.39
	Baseline + VP	**5.05**	5.46	**3.35**	**13.85**

Table 9. Results of different numbers of generated frames on STAR. † means using ground truth future frames.

# Generated Frames	Int.	Seq.	Pre.	Fea.	Avg.
1	**51.5**	**49.9**	**50.0**	47.1	49.6
2	**51.5**	**49.9**	49.8	**47.6**	**49.7**
3	51.4	**49.9**	**50.0**	47.3	**49.7**
$1^{\dagger}$	**51.3**	**50.6**	**57.5**	51.4	52.7
$2^{\dagger}$	**51.3**	50.5	55.5	**54.9**	**53.0**
$3^{\dagger}$	51.2	50.5	48.2	50.4	50.1

the phases like *[fireworks erupted in a symphony of red and gold near the water's edge]* and *[the display escalated into an orchestra of purple, orange, and blue blooms, each outshining the last]* while vanilla answering does not.

4.3. Ablation Study

Effects of VP and LP. We conduct an ablation study of VP and LP on Video Q&A dataset STAR and Video Captioning dataset BDD-X. Table 3 and Table 4 show that both VP and LP could clearly boost the performance of the baseline. For example, VP and LP improved 2.2% and 3.0% Accuracy on STAR and 1.1 and 0.6 BLEU-4 score on BDD-X. LP brings more benefit than VP on vision-language task STAR while this circumstance is contrary on vision task BDD-X. This is because understanding the language question is also significant for the Q&A task, while the captioning task has a consistent output demand based on only vision input.

Effects of VPS in VP. Coarse Selector (CS) is to determine whether the generated video is needed for the current task.

Table 10. Ablation Study of video generation model on BDD-X.

Video Generation Method	B	C	M
MCVD - Cityscapes	31.2	195.3	26.8
DMVFN - Cityscapes	35.0	230.1	**29.4**
DMVFN - Kitti	**35.2**	**234.2**	**29.4**
Stable Video Diffusion	33.9	229.6	28.8
Ground Truth Frames	34.6	247.5	30.6

Table 11. Results of different numbers of generated frames on BDD-X.

# Generated Frames	2	4	8	16	30
BLEU-4	32.0	33.5	**35.2**	34.4	33.3
CIDEr	212.6	216.8	**234.2**	228.0	223.3
METEOR	29.0	29.3	**29.4**	29.2	28.6

The *Interaction* and *Sequence* questions in STAR are not supposed to be related to the future frames, and Table 5 shows the performance of them drops about 2% without CS, which means introducing generated frames might bring noisy information for questions independent of the future. Most of the *Prediction* and *Feasibility* questions are related to the future states so they will be chosen by CS to use generated future frames. Without FS, the performance of *Prediction* and *Feasibility* questions drop dramatically, which illustrates the significance of using FS for picking up useful and high-quality generated frames.

Effects of Voting in Decision Maker. The generated language plan and vision plan may not always be reliable due to the limited ability of the language and video generation model. Table 6 shows that letting the model vote again between the vanilla answer and the answer with language or vision plan could effectively enhance the performance.

VP for Action Prediction and Planning. In addition to the language output tasks including video Q&A and captioning, we also implement VP on the action model. Table 7 shows that with the help of generated future frames, the model could predict the course and speed more accurately in the driving scenario. We also conduct robotics gripper trajectory prediction (predict the future actions given initial states) and planning (generate the future actions given initial and goal states). Table 8 shows that VP also helps in this application.

Video Generation Quality Matters. The video generation quality plays a significant role in our visual planning.

Table 9 shows that using real future frames has significantly better performance than generated frames using Stable Video Diffusion, *e.g.*, 57.5 and 54.9 compared to 50.0 and 47.6 on *Prediction* and *Feasibility* questions. Fig. 8 and Fig. 10 give cases where the generated contents are not reasonable enough to provide positive information while ground truth future frames are helpful. Due to the limited quality of the generated future frames in the open domain, selecting more frames does not have clear performance improvements according to Table 9.

On the BDD-X dataset, we select the first 2 frames out of all 30 frames as the input to conduct the ablation study. Table 10 shows using ground truth future frames achieves better overall performance than generated frames. MCVD performs worst since it generates low-resolution images. DMVFN trained on the driving datasets including Cityscapes (Cordts et al., 2016) and Kitti (Geiger et al., 2013) show better performance because of higher resolution. Stable Video Diffusion does not perform better as it is not specifically trained for the driving scenario. Table 11 shows that a proper number of generated frames is helpful when using domain-specific generative models, but long sequence generated videos are not reliable enough.

5. Conclusion

In conclusion, we propose a Visual-Language Planning (VLP) framework in this work. By incorporating both vision-based associative reasoning and language planning, our VLP framework has demonstrated enhanced capabilities in handling multi-modality tasks, which aligns with the cognitive processing strategies of humans involving both hemispheres of the brain. We hope our work could inspire the community to develop more advanced and human-like artificial intelligence systems.

Acknowledgements

This work is supported by National Key Research and Development Program of China (2021YFF1200800), National Natural Science Foundation of China (Grant No. 62276121), Shenzhen International Research Cooperation Project (Grant No. JHZ20220913142611021), Guangdong Science and Technology Department (Grant No. 2023A0505010004), and the Research Grant Council of the Hong Kong Special Administrative Region (Grant No. 16212623).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ajay, A., Han, S., Du, Y., Li, S., Gupta, A., Jaakkola, T., Tenenbaum, J., Kaelbling, L., Srivastava, A., and Agrawal, P. Compositional foundation models for hierarchical planning. *arXiv preprint arXiv:2309.08587*, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., and Dao, T. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Corballis, M. C. Left brain, right brain: facts and fantasies. *PLoS biology*, 2014.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Du, Y., Yang, M., Dai, B., Dai, H., Nachum, O., Tenenbaum, J. B., Schuurmans, D., and Abbeel, P. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2023a.

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*, 2023b.

Ebert, F., Finn, C., Lee, A. X., and Levine, S. Self-supervised visual planning with temporal skip connections. *CoRL*, 2017.

Gazzaniga, M. S. Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience*, 2005.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hu, X., Huang, Z., Huang, A., Xu, J., and Zhou, S. A dynamic multi-scale voxel flow network for video prediction. In *CVPR*, 2023.

Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., and Liu, J. Adapt: Action-aware driving caption transformer. *arXiv preprint arXiv:2302.00673*, 2023.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. Textual explanations for self-driving vehicles. In *ECCV*, 2018.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Lu, Q., Zhang, Y., Lu, M., and Roychowdhury, V. Action-conditioned on-demand motion generation. In *ACM International Conference on Multimedia*, 2022.

Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., and Tan, T. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023.

Momeni, L., Caron, M., Nagrani, A., Zisserman, A., and Schmid, C. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 2023.

Skorokhodov, I., Tulyakov, S., and Elhoseiny, M. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022.

Surís, D., Menon, S., and Vondrick, C. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team, I. Internlm: A multilingual language model with progressively enhanced capabilities, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. Sequence to sequence-video to text. In *ICCV*, 2015.

Voleti, V., Jolicoeur-Martineau, A., and Pal, C. MCVD - masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Wang, X., Zhang, S., Cen, J., Gao, C., Zhang, Y., Zhao, D., and Sang, N. Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, 2023b.

Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., and Sang, N. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023c.

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Wang, Y., Jiang, L., and Loy, C. C. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, 2023d.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Wu, B., Yu, S., Chen, Z., Tenenbaum, J. B., and Gan, C. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., and Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Xiao, J., Shang, X., Yao, A., and Chua, T.-S. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

Xu, H., Gao, Y., Yu, F., and Darrell, T. End-to-end learning of driving models from large-scale video datasets. In *CVPR*, 2017.

Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.

Yu, S., Cho, J., Yadav, P., and Bansal, M. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023.

Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., and Zhou, J. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023a.

Zhang, Y., Aghajan, Z. M., Ison, M., Lu, Q., Tang, H., Kalender, G., Monsoor, T., Zheng, J., Kreiman, G., Roychowdhury, V., et al. Decoding of human identity by computer vision and neuronal vision. *Scientific reports*, 2023b.

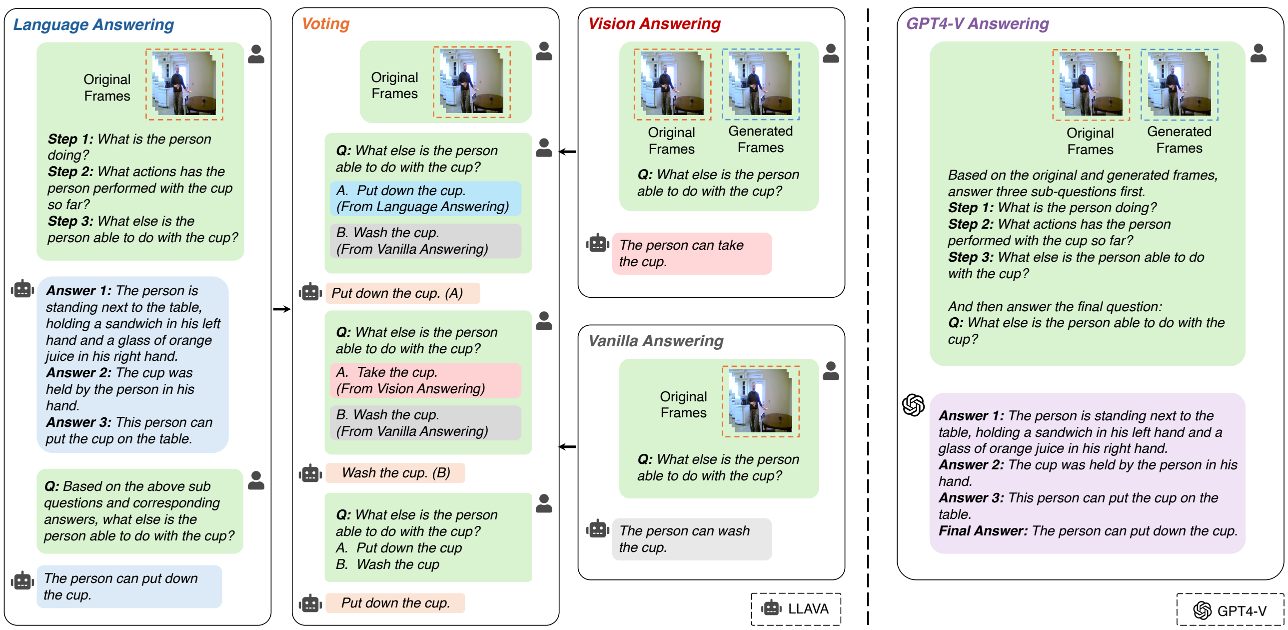
Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *ICLR*, 2023c.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023d.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Decision Maker

We give a detailed example of our proposed LMM-based decision maker in Fig. 5. For LMM that has very strong visual instruction following ability like GPT4-V, it could directly answer the overall question following the language plan and vision plan, as shown in Fig. 5 (b). However, we find that the open-source LMM such as LLAVA can only follow simple visual instructions and cannot handle flexible and complicated visual instructions. For example, LLAVA cannot answer the questions sequentially in one reply. So we design a multi-round conversation strategy, as shown in Fig. 5 (a), which is explained in Sec. 3.4. Either vision planning and language planning could introduce the noise and our voting mechanism could effectively strengthen the robustness of the final answer, which is testified by Table 6.



(a) Decision Maker using LLAVA

(b) Decision Maker using GPT4-V

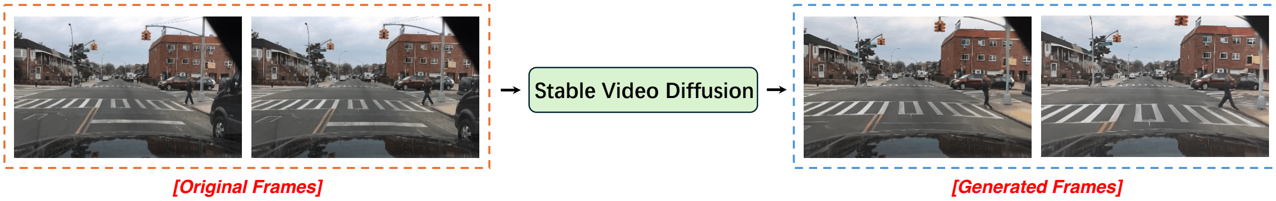
Figure 5. Decision maker using open-sourced LMM like LLAVA and GPT4-V.

B. VLP Case Study with GPT4-V

We provide the case study of vision-related task in Fig. 6 and language-only task in Fig. 7.

Fig. 6 is the detailed version of Fig. 3. The vanilla answering gives a general and ambiguous answer and requires more information for the decision making. It is *[impossible to predict]* the next move based on the current condition. Language planning decomposes the question into three sub-questions, and let the model answer these sub-questions one by one. Although more information is obtained through language planning, it is still *[difficult to predict]* the next move. Then we use the Stable Video Diffusion for vision planning to generate the future frames, which show that the pedestrian is crossing the road. Using this vision plan, GPT4-V gives the ideal answer that the car should proceed *[once the pedestrian has safely crossed and the light turns green]*.

Fig. 7 is the detailed version of Fig. 4. The language planning provides a specific outline for the answer, resulting in a better-structured long article. For example, GPT4-V shows that the result should follow *[Introduction, Visual Description, Public Engagement, etc.]*. Besides, it shows that the answer of GPT4-V is more vivid and realistic because of the introducing of generated videos. For example, GPT4-V gives specific location such as *[Empire State Building and the sprawling bridges]* and detailed description of fireworks like *[the first burst of fireworks erupted in a symphony of red and gold near the water's edge]* and *[an orchestra of purple, orange, and blue blooms]*.



Vanilla Answering

[Original Frames] The images are a sequence of frames from a video. Based on this video, how will the car drive next?

The images you've provided are two separate frames from a video, showing a car's perspective at an intersection with a pedestrian crossing the street. It's ***impossible to predict with certainty the car's next move*** without more context or additional frames from the video. However, we can make some educated guesses based on traffic rules and typical driving behavior:
1.If the traffic light visible in the images is green for the car's direction, and the pedestrian is not in the car's path, the car may proceed through the intersection.
2.If the traffic light is red, or if the pedestrian's path intersects with that of the car, the car should come to a stop.
3.If the car is making a turn and the pedestrian is clear of its path, the car may turn once it's safe to do so, yielding to any other vehicles with the right of way.
It is also possible that the car could be preparing to park or pull over, depending on the driver's intent and the presence of parking spaces.
Without additional frames showing the traffic lights in more detail or the actions of the pedestrian and other vehicles, it's not possible to provide a definitive answer.

Language Planning

[Original Frames]

These images are a sequence of frames from a video. You will firstly watch this video and answer three questions.
1.What is the Current Context of the Scene? 2.What are the Visible Cues or Indicators from the Car? 3.What are the Likely Actions Based on Road Rules and Driver Behavior?
And then provide answers to the final question: How will the car in the video drive next?

1.Current Context of the Scene:
It appears to be an intersection with multiple lanes and traffic signals. There are pedestrian crosswalks, and the scene looks like a typical urban street during daylight hours. Pedestrians are present, suggesting that pedestrian activity is a factor in this context.
2.Visible Cues or Indicators from the Car:
From the reflections and shadows on the car's hood, it seems like the car is stationary as there is no blur or distortion that would indicate movement. The positioning of the car at the stop line suggests it is waiting for a traffic signal or allowing pedestrians to cross.
3.Likely Actions Based on Road Rules and Driver Behavior:
Given the context, if the traffic lights are red, the driver should remain stopped. If the light is green and pedestrians are not in the crosswalk, the driver would likely proceed with caution, following road rules and signals.

For the final question, "How will the car in the video drive next?"—without movement indicators or knowing the current state of the traffic signals, it's ***difficult to predict the car's next action definitively****. However, if we assume the traffic light turns green, and there are no pedestrians crossing, the car would likely drive straight ahead or turn if it's in a turning lane, following standard driving procedures. If the light remains red or pedestrians are crossing, the car would stay stationary until it is safe and legal to proceed.*

Vision-Language Planning

[Original Frames] + [Generated Frames]

These images are a sequence of frames from a video. The first two images indicate what is happening, while the last two images are generated for future events. You will firstly watch this video and answer three questions.
1.What is the Current Context of the Scene? 2.What are the Visible Cues or Indicators from the Car? 3.What are the Likely Actions Based on Road Rules and Driver Behavior?
And then provide answers to the final question: How will the car in the video drive next?

1.Current Context of the Scene:
The scene appears to be ***an intersection*** in an urban area with a traffic light system. There is ***a pedestrian crossing the street*** in a crosswalk. ***The traffic light*** visible in the images ***is red***, suggesting that vehicles in the direction of the viewer must stop.
2.Visible Cues or Indicators from the Car:
The images show a view from inside a car, likely from a dashboard camera. The hood of the car and a part of the windshield are visible, and they are not moving relative to the crosswalk lines, which indicates ***the car is stationary***. The reflection on the car's hood shows that the traffic lights for the lane of the car are red.
3.Likely Actions Based on Road Rules and Driver Behavior:
Since ***the traffic light*** is red, the car should remain stationary. The driver is expected to ***wait for the pedestrian*** to finish crossing safely. The driver should wait for the traffic light to turn green before proceeding.

Based on this information and assuming the driver follows standard road rules and safe driving practices, the car in the video will likely ***remain stationary until the traffic light turns green***. Once ***the pedestrian has safely crossed and the light turns green***, the driver will likely proceed through the intersection.

Figure 6. VLP with GPT4-V on vision-related tasks.

Vanilla Answering

You are a journalist writing a news report about the New York's firework show, which illuminate the skyline, evoking a dreamy vibe. Please write the news.

New York City's Skyline Illuminated by Dazzling Firework Display
New York, [Date] – The night sky over New York City was transformed into a dazzling display of color and light as the annual firework show took place, captivating thousands of residents and tourists alike.
*The event, which is a highlight of the city's cultural calendar, **saw the skyline near the [specific location, e.g., East River, Hudson River, etc.] come alive with an array of spectacular fireworks.** The display started at [time] and lasted for approximately [duration], during which the sky was lit up with an intricate dance of lights, creating a dreamy and almost surreal atmosphere.*
***Spectators gathered** in large numbers at various vantage points throughout the city, including parks, rooftops, and bridges. Social distancing measures were observed where possible, as the city continues to navigate the post-pandemic era.*
***The firework show, choreographed by [Name of the Company or Choreographer], featured a series of thematic sequences**, each designed to evoke different emotions and celebrate the city's diverse culture and spirit. The display was accompanied by a synchronized soundtrack that added to the immersive experience.*
***City officials noted** that this year's event was especially significant as it symbolized not only the city's resilience in the face of recent challenges but also its hope and optimism for the future. Mayor [Name] stated, "Tonight's firework display is a testament to New York City's enduring vibrancy and unity. It's a reminder of the strength and spirit of our people."*
*The firework display also had an **economic impact,** with local businesses near viewing areas reporting increased activity. Restaurants, bars, and street vendors welcomed the boost in customers, seeing it as a positive sign of the city's recovery.*
***Environmental and safety measures** were also a priority, with the fire department and police ensuring that the event was conducted safely and with minimal environmental impact.*
***As the final fireworks faded** into the night, the city was left with a sense of awe and joy. Social media platforms were abuzz with photographs and videos, as people shared their experiences of the magical evening.*
The firework show in New York City, apart from being a visual spectacle, continues to be an important event that brings together people from all walks of life, reflecting the city's spirit of unity and celebration.

Vision-Language Planning

[Language Prompt]
In the futuristic cityscape of NYC skyline at night, fireworks illuminate the sky with epic realism, captured in highly detailed 8K resolution using a long shot technique, evoking a dreamy vibe. → **Pika** →

[Generated Frames]

You are a journalist writing a news report about the New York's firework show, which illuminate the skyline, evoking a dreamy vibe. Please first describe the outline of the news and then write the news with reference of the pictures.

Outline for the News Report on New York's Firework Show:
*1. **Introduction:***
Brief overview of the event's significance.
Mention of the date and location.
*2. **Visual Description:***
Descriptions of the fireworks illuminating the skyline. Mention of the colors and patterns observed.
Reference to the mood and atmosphere created by the fireworks.
*3. **Public Engagement:***
Quotes from spectators and officials.
The number of people in attendance and the general public response.
*4. **Technical Aspects:***
Details about the fireworks used (types, quantity, etc.).
Information on the planning and execution of the firework display.
*5. **Cultural Impact:***
Discussion on how the fireworks display adds to New York's cultural tapestry.
Reference to past events and the tradition of fireworks in the city.
*6. **Closing Thoughts:***
Reflection on the success of the event.
Final remarks on the night and what it symbolizes for the city.

News Report:
New York's Night Sky Transforms into a Dreamy Palette of Firework Fantasia
*In the heart of New York City, the night of [Date] unfurled a spectacular show of pyrotechnics that painted the urban skyline with a myriad of colors, evoking a dreamy vibe that enchanted thousands of onlookers. The annual fireworks show, a testament to the city's love for grandeur, illuminated the heavens above **Manhattan**, creating a vibrant backdrop to the iconic silhouettes of the **Empire State Building** and the **sprawling bridges** that stitch the city's boroughs together.*
***As twilight melted** into the darker shades of evening, **the first burst of fireworks erupted in a symphony of red and gold near the water's edge**, reflecting on the river's surface like flickering flames. The display escalated into **an orchestra of purple, orange, and blue blooms**, each outshining the last, while the city's towering skyscrapers watched on, bathed in a neon glow.*
*Photographs of the event showcase a dazzling array of fireworks that seemed to compete with the stars, with bursts of **vibrant hues ranging from magenta to fiery oranges**, against the canvas of the night sky. In one particularly striking image, a solitary firework radiates a celestial white, like a star descending to join the city's skyline, offering a stark contrast to the warm glow emanating from the buildings below.*
*The **grand finale was a masterstroke, a crescendo** of light that seemed to set the entire city aflame, with a dense constellation of sparks leaving spectators in awe. The resulting imagery captured by attendees and shared across social media channels has become an instant sensation, with many declaring the event an ethereal experience that uplifted spirits and showcased the city's vibrant essence.*
In an age where digital spectacles are the norm, New York's firework show stands out as a beacon of collective joy and shared experience. It is a reminder of the city's capacity to inspire and amaze, to hold onto moments of beauty that transcend the ordinary. As the last sparkle faded into the night, it was clear that the event was much more than a display of light and sound—it was a celebration of New York City's enduring spirit, a dazzling dance of light that will be remembered for years to come.

Figure 7. VLP with GPT4-V on language-only tasks.

C. VLP Case Study with LLAVA

We provide several case studies with LLAVA from STAR dataset. Fig. 8 and Fig. 9 show the successful and unsuccessful cases of vision planning and language planning. Fig. 10 shows the vision planning using ground truth future frames.

C.1. Vision Planning Case Study

Fig. 8 (a) shows that LLAVA thinks the man is going to open the cabinet without vision planning, which is reasonable according to the background and the action of the man in the video. Stable Video Diffusion generates the future frames which show that the man is reaching out his hand to the paper, so LLAVA gives the correct answer. In Fig. 8 (b), LLAVA gives the correct answer with original frames since there is a white box in the man's hand (please zoom in the figure for better visualization). However, the generated future frames show the man continue turning around and does not put down the box, so LLAVA gives the wrong answer.

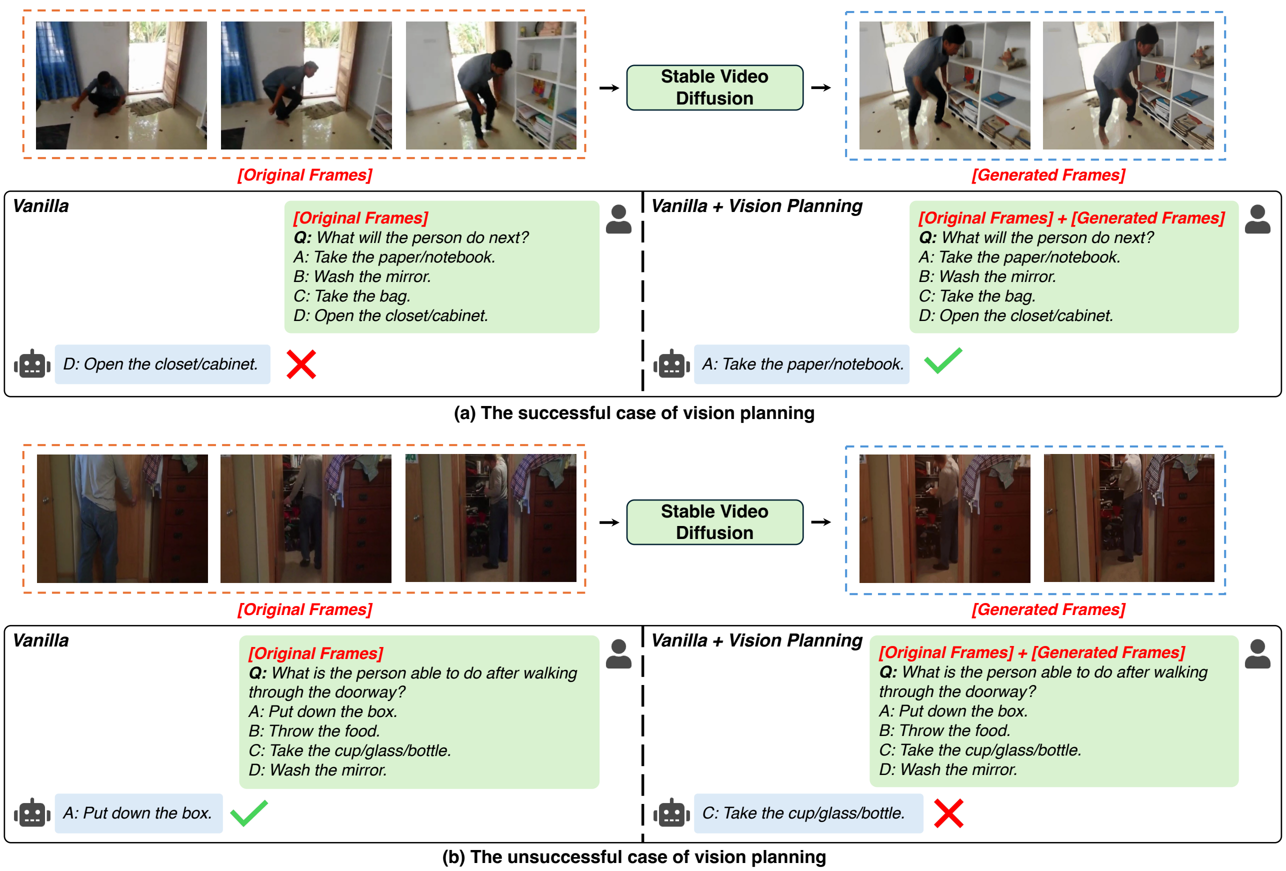


Figure 8. The successful and unsuccessful cases of vision planning.

C.2. Language Planning Case Study

Fig. 9 (a) shows that LLAVA believes that the person puts down the blanket before he took the book, and we believe that the reason is the blanket is prominent in the images. Using Language planning, LLAVA first answers three sub-questions and mentions the man *[puts the plate down before taking the book]*. The sandwich is the only choice that is related to the plate, so LLAVA chooses sandwich and gives the correct answer.

In Fig. 9 (b), LLAVA gives the correct answer for the question without language planning. However, LLAVA gives the wrong responses when answering the sub-questions of language planning, *e.g.*, the person is *[not expecting it]* and *[thrown away]* the sandwich. Therefore, the final answer picked up by LLAVA is putting down the sandwich which is directly related to *[thrown away]*. This case shows that the quality of generated language plans and corresponding answers for the language plans are significant.

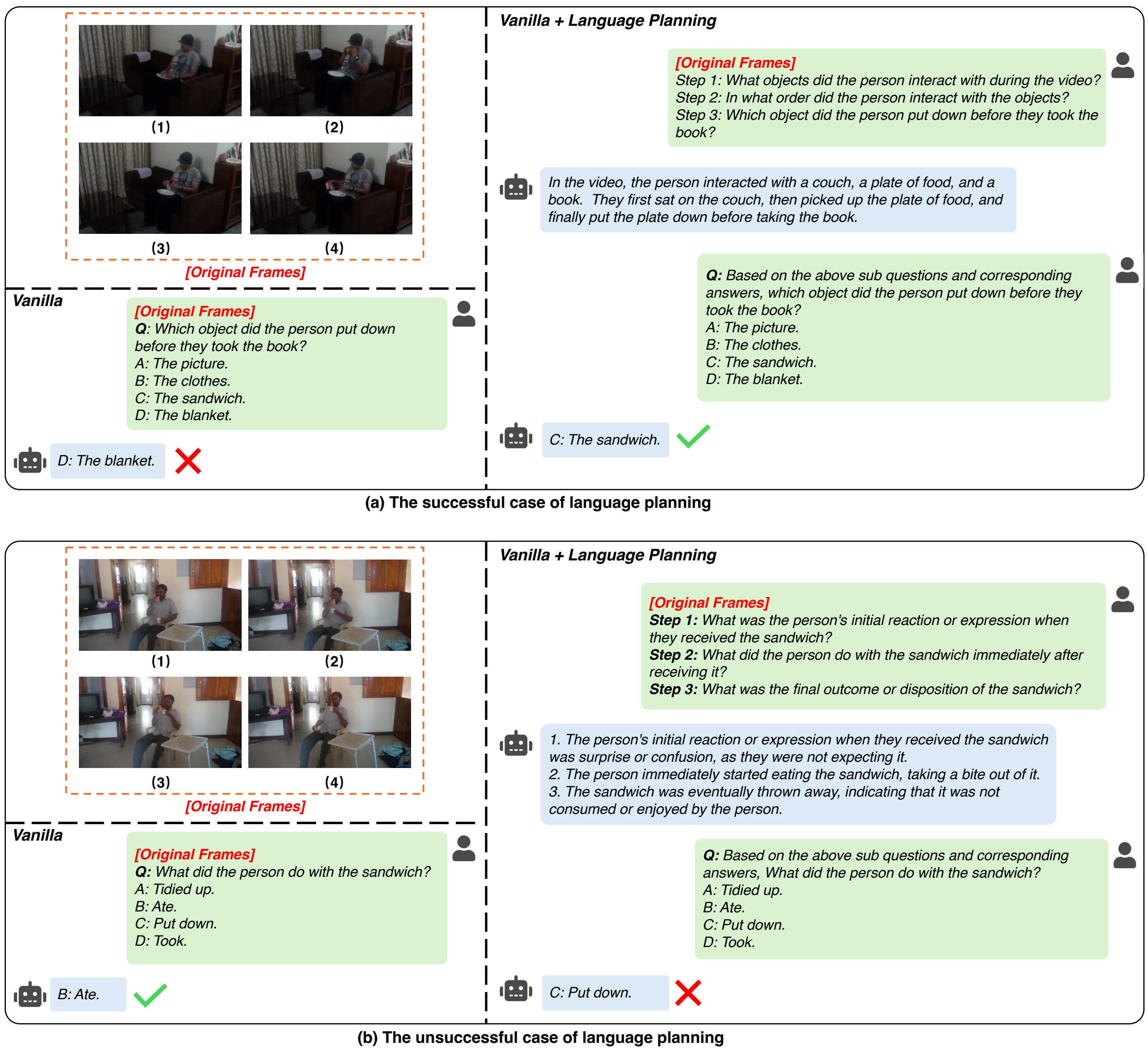


Figure 9. The successful and unsuccessful cases of vision planning.

C.3. Vision Planning using Ground Truth Future Frames

Current LLMs and LMMs have strong text generation ability across open domains, but the video generation models are still far behind in terms of generation ability. In Fig. 10 (a), the generated arm action is not reasonable enough. In Fig. 10 (b), the moving part of the generated video is blurry. In Fig. 10 (c), the person disappears without opening the door. For these cases, using generated future frames cannot correct mistakes for LLAVA but using ground truth future frames can. These cases show that the ability of video generation model is a bottleneck for vision planning, which has been discussed in the last part of Sec. 4.3.

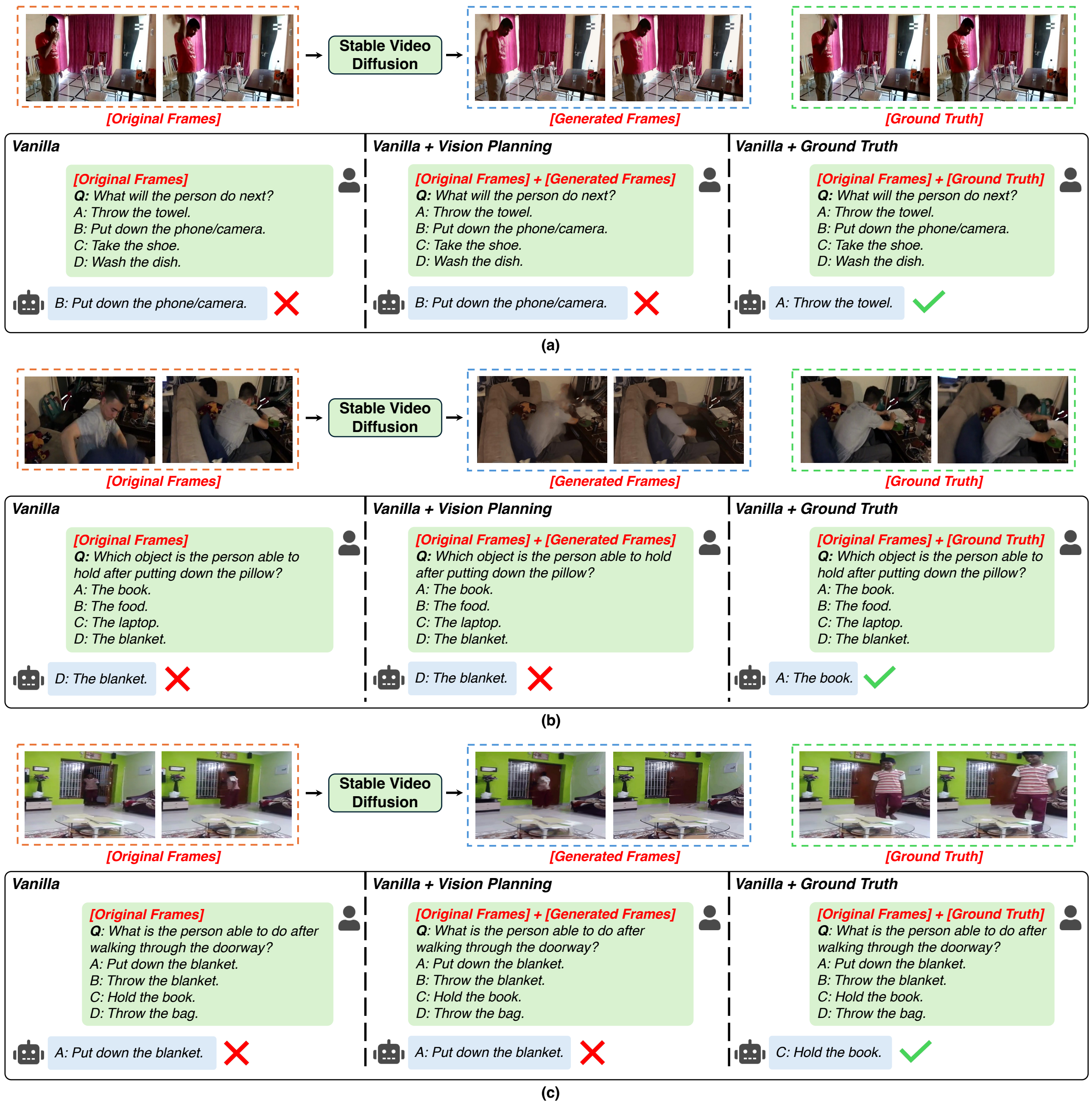


Figure 10. The successful cases of vision planning using ground truth future frames.

D. VLP Case Study with BERT

Our VLP framework can be utilized not only with recent LLMs and LMMs, but it can also be applied to the traditional BERT for captioning task (Fig. 11 from BDD-X dataset) and action generation task (Fig. 12 and Fig. 13 from BAIR dataset).

D.1. Video Captioning Case Study

Fig. 11 (a) shows that BERT model predicts the car merges left which is contradictory to the truth that the car is merging right and driving down the highway. With generated future frames, BERT model gives the correct answer. In Fig. 11 (b), both vanilla and vision planning do not give the correct answer, while language planning provides the optimal response with the hint from generated language descriptions.

Figure 11. Vision Planning and language planning for the video captioning task.

D.2. Action Generation Case Study

We provide several cases for action prediction (predict the next actions based on first two frames) and action planning (predict the actions between the initial frame and end frame) in Fig. 12 and Fig. 13, respectively.

Fig. 12 (a) and (b) show that the generated vision plans successfully predict the gripper to grab the green ball and leave the yellow ball, but Fig. 12 (c) shows that the gripper in the generated future frames circles around the green ball without makes contact, which is not the original intention in the ground truth.

Fig. 13 shows that the inference results of the planning task are closer to the ground truth than that of the prediction task, since the goal state is given and the video generation process is guided by the goal. Fig. 13 (a) shows that the generated video successfully predicts the gripper to approach the white ball. Fig. 13 (b) is a more sophisticated task, where the gripper first approaches the black object, and then moves the black object up, and finally leaves the black object. The generated video successfully reproduces the whole process according to the final position of the black object, which shows the great potential of vision planning. In Fig. 13 (c), the video generation model fails to generate the correct process that the gripper is moving the green ball.

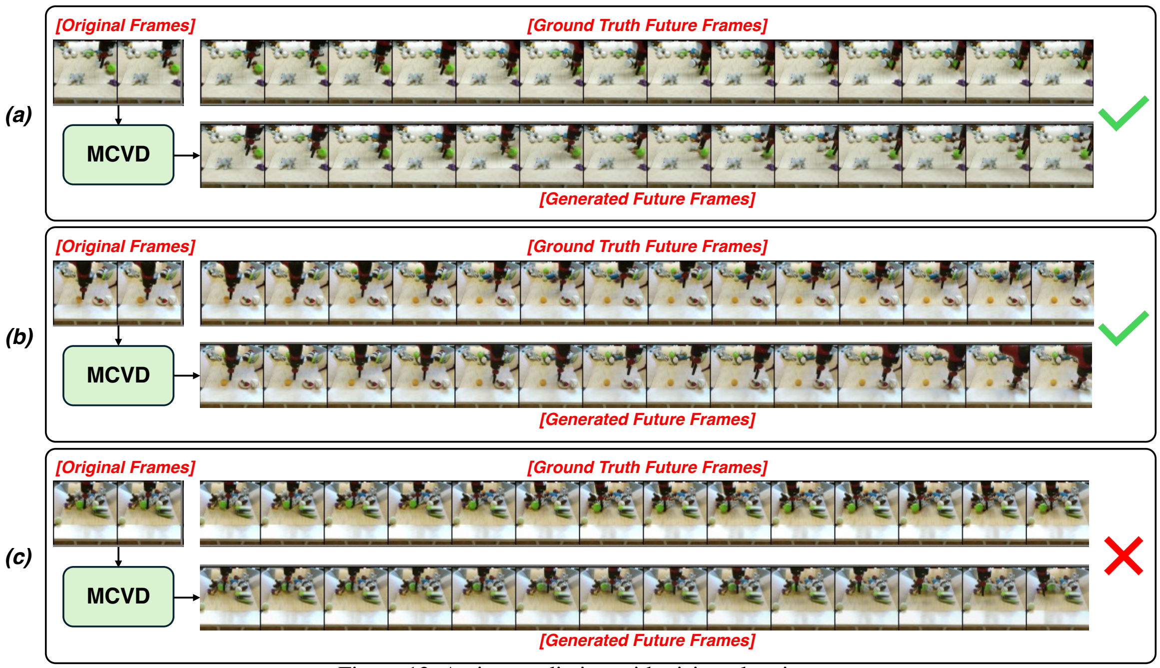


Figure 12. Action prediction with vision planning.

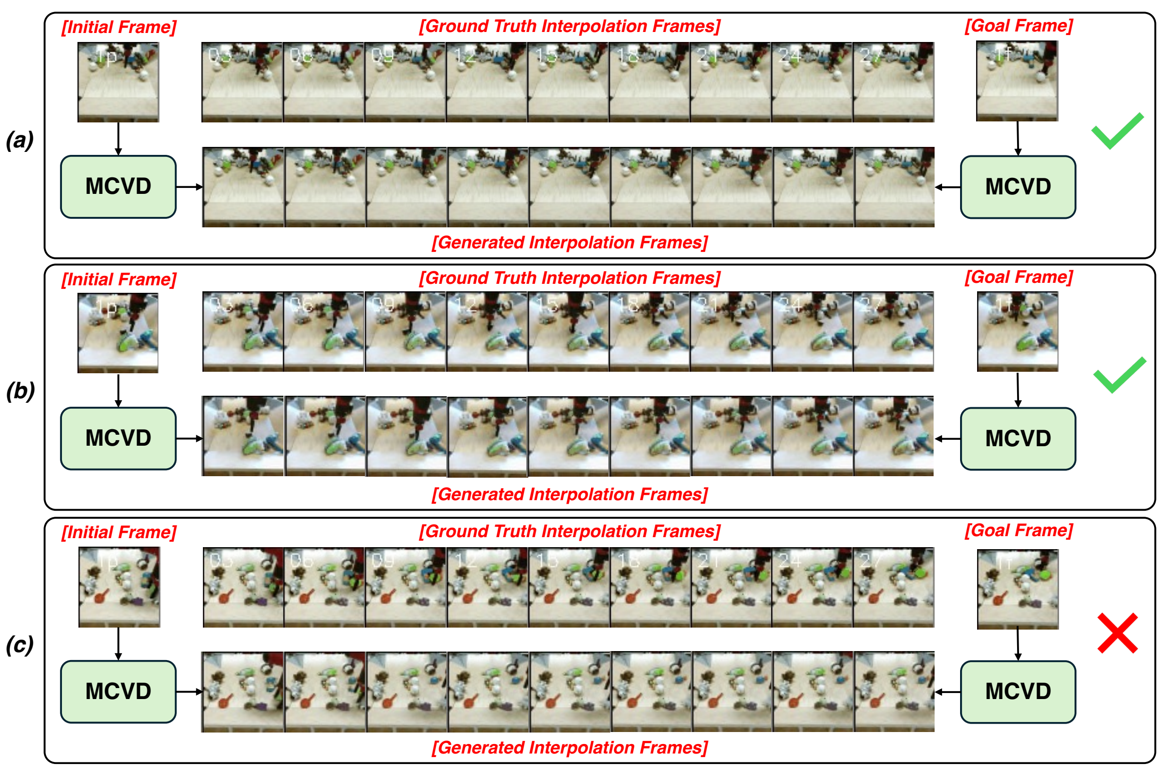


Figure 13. Action planning with vision planning.

E. Inference Time Bottleneck

The inference time of GPT models through OpenAI's api is short, which depends on the input and output text length and is usually less than 1s. This might benefit from work on LLM inference acceleration such as (Cai et al., 2024). The inference time bottleneck of VLP lies in the video generation model, which consumes much longer time than LLM. For example, Stable Video Diffusion takes more than 60s to generate a short video. There are some works that could reduce the diffusion steps from 50 steps to 4 steps and minimize the inference time to 10s (Wang et al., 2023c). Since the video generation model is not mature as LLM and still under the fast development, we believe that the inference time of the video generation would be much faster if the algorithm is mature enough and more acceleration techniques are designed.