VG-TVP: Multimodal Procedural Planning via Visually Grounded Text-Video Prompting

Muhammet Furkan Ilaslan[1, 2*], Ali Koksal[2], Kevin Qinhong Lin[1], Burak Satar[2], Mike Zheng Shou[1], Qianli Xu[2]

[1]Show Lab, National University of Singapore, Singapore
[2]Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore
{m.furkanilaslan, qinghonglin}@u.nus.edu, mike.zheng.shou@gmail.com
{stuimf, koksal_ali, burak_satar, qxu}@i2r.a-star.edu.sg

Abstract

Large Language Model (LLM)-based agents have shown promise in procedural tasks, but the potential of multimodal instructions augmented by texts and videos to assist users remains under-explored. To address this gap, we propose the Visually Grounded Text-Video Prompting (VG-TVP) method which is a novel LLM-empowered Multimodal Procedural Planning (MPP) framework. It generates cohesive text and video procedural plans given a specified high-level objective. The main challenges are achieving textual and visual informativeness, temporal coherence, and accuracy in procedural plans. VG-TVP leverages the zero-shot reasoning capability of LLMs, the video-to-text generation ability of the video captioning models, and the text-to-video generation ability of diffusion models. VG-TVP improves the interaction between modalities by proposing a novel Fusion of Captioning (FoC) method and using Text-to-Video Bridge (T2V-B) and Video-to-Text Bridge (V2T-B). They allow LLMs to guide the generation of visually-grounded text plans and textual-grounded video plans. To address the scarcity of datasets suitable for MPP, we have curated a new dataset called Daily-Life Task Procedural Plans (Daily-PP). We conduct comprehensive experiments and benchmarks to evaluate human preferences (regarding textual and visual informativeness, temporal coherence, and plan accuracy). Our VG-TVP method outperforms unimodal baselines on the Daily-PP dataset.

Dataset — https://github.com/mfurkanilaslan/VG-TVP
Extended version — https://arxiv.org/abs/2412.11621

Figure 1: VG-TVP generates MPP with multiple steps for a high-level goal, supplying textual and visual guidelines.

Introduction

To acquire procedural knowledge, such as operating a machine, a person can refer to procedure plans, which specify the steps to achieve a task. Procedure plans may take different formats such as text, image, video, and a combination of them. Procedure Planning (PP) is the process of generating procedure plans. Depending on the modality of procedure plans, PP can be implemented using different methods and various sources of information. For example, LLMs have been used to generate procedure plans either in a zero-shot manner (Huang et al. 2022) or by fusing information from various resources (e.g. WikiHow) (Lu et al. 2023b).

Instructional videos (IVs) are a useful source of information on procedural knowledge. It provides a rich context of the task steps and effectively incorporates the temporal information essential for procedural knowledge learning. However, the quality of IVs might be inconsistent and it usually involves considerable effort of a human learner to understand the information and acquire the knowledge. To address this issue, one may resort to computational methods to parse the IVs and use the extracted visual elements to generate informative procedure plans. This necessitate several capabilities, such as semantics understanding (Zhukov et al. 2019), action-step prediction (Niu et al. 2024), and scene understanding (Zhou et al. 2023). However, existing IVs may lack complete procedural information or contain visual content that does not align with the semantic plans.

* Corresponding author.

Hence, generating video content in association with the textual plans is desirable to form the final procedure plans.

In this paper, we aim to improve PP by generating multimodal content from different resources, including IVs. We are interested in enhancing human understanding by integrating visually grounded text and action-based video generations (Figure 1). Inspired by Text-to-Image Prompting (TIP) (Lu et al. 2023b), we cast the problem as MPP via Visually Grounded Text-Video Prompting (VG-TVP) (Figure 2). VG-TVP generates video-enhanced action and state procedures given text descriptions of a task and IVs, which is in contrast to generating image plans and using their descriptions as text plans (Lu et al. 2023b; Soucek et al. 2024).

We anticipate that video-augmented procedure plans are advantageous to image-augmented ones (Lu et al. 2023b; Ramesh et al. 2021) because images focus on the static "states" of the task, whereas videos display dynamic changes of "states" with human-centred "actions". However, the challenges faced by TIP, including textual-visual informativeness, coherent temporal alignment, and high-level accurate plan generations (Lu et al. 2023b), become even more prominent in MPP. In particular, the framework needs to ensure both temporal consistency (i.e. coherent temporal alignment between text and video plans) and spatial consistency (i.e. subsequent video steps must logically follow preceding ones in actions, objects, and contexts). Existing generative models, including LLMs and multimodal LLMs (MLLMs), cannot adequately address these challenges.

We propose a novel video-to-text−text-to-video (V2T-T2V) methodology that enhances the capability of LLMs for MPP tasks. Given the high-level goal-oriented task description, we first use LLMs to generate vanilla text plans. Meanwhile, we generate video captions from given IVs via V2T-Bridge (V2T-B). Subsequently, LLMs compile captions from various IVs to form a cohesive *"Fusion of Captioning (FoC)"* (Figure 3) that adheres to the required steps. We leverage the LLMs' capabilities by eliciting video captions and vanilla text plans to generate final revised text plans. Finally, video plans are generated by considering the generated text plans. Consequently, generated video plans are aligned with the generated text plans through VG-TVP.

Video-augmented PP is advantageous to IVs in several aspects. First, generated videos (a few seconds for each step) focus on the task steps, making them concise and relevant to the task. This may help alleviate the cognitive efforts of human learners than if they watch lengthy videos ($\approx 5-10$ minutes) with much auxiliary information and time management. Second, the structured PP format could also improve the clarity and reliability of the MPP content, which facilitates self-paced learning. Third, using well-crafted prompts, we guide the video generation module to generate human-centred content, i.e. showing human actions whenever possible (Figure 5). This potentially helps reduce users' cognitive load owing to a smaller cognitive gap than if only state changes of the scene (without humans) are visualized.

VG-TVP addresses 3 major technical challenges: *(1) Costly structure of foundational models (FM).* Although FMs are proficient in text generation, they must be trained with visual information. Training of data-hungry FMs from scratch demands high costs. In contrast, VG-TVP employs 3 distinct components that strategically leverage each other's inputs and outputs, thereby eliminating the need to train separate models for specific tasks. *(2) Alignment between procedures in IVs and generated text plans.* This challenge is addressed by FoC collaborating with VG-TVP alignment prompt. *(3) Inadequacy of video captioning models to capture detailed MPP.* Captioning algorithms aim to capture the descriptions of the scenes, instead of detailed MPP. For instance, VLog (Lin and Lei 2023) focuses on different aspects such as image, region, and the audio in the videos. Although it uses the latest LLMs, image, and audio models, it is not capable of combining the multimodal procedures for MPP tasks. A multimodal pre-trained video captioning model is proposed, focusing on dense events in the videos to capture descriptions in the same sequence (Yang et al. 2023). However, it only captures instructions, not coherent MPPs. In short, although existing methods are capable of captioning, additional information is required to generate MPPs.

The contributions of the study are four-fold. (1) We design a multi-modal framework, VG-TVP, that employs zero-shot prompting and avoids the costly training procedure. (2) We propose a novel method, FoC, which removes irrelevant contents in IVs, aligns mismatched steps between video captions and generates relevant text plans. They provide detailed MPP in the absence of IVs by exploiting video captions. (3) We propose Video-to-Text (V2T) and Text-to-Video (T2V) bridges to generate visually grounded text and video plans that are temporally coherent and accurate in planning. (4) We introduce a well-curated dataset called the Daily-Life Task Procedural Plans (Daily-PP) to mitigate the challenges in existing datasets for MPP content. VG-TVP's results show superior performance in the zero-shot setting compared to several baselines under the Daily-PP dataset.

Related Work

Video Understanding for Procedural Planning

There has been remarkable progress achieved in understanding IVs via various problem statements (Kuehne, Arslan, and Serre 2014; Miech et al. 2019; Ilaslan et al. 2023; Zhou, Xu, and Corso 2018; Chang et al. 2020). While prior studies handle the problems via fully (Elhamifar and Naing 2019; Zhu and Yang 2020) or weakly-supervised (Zhukov et al. 2019; Zhao et al. 2022) or unsupervised settings (Alayrac et al. 2016), others focus on step discovery (Dvornik et al. 2023), prompting (Lu et al. 2023a) or diffusion models (Fang et al. 2023; Wang et al. 2023a; Lian et al. 2023). Unlike existing studies, we focus on generating visually grounded text-video pairs to obtain coherent MPPs.

Multimodal Generation

Zero-shot planners (Huang et al. 2022; Song et al. 2023) or reasoners (Kojima et al. 2022) have leveraged the natural language generation capabilities of LLMs. While these efforts have yielded promising results, developing LLMs with multimodal content generation capabilities may lead to broader practical impacts. Although the performance of

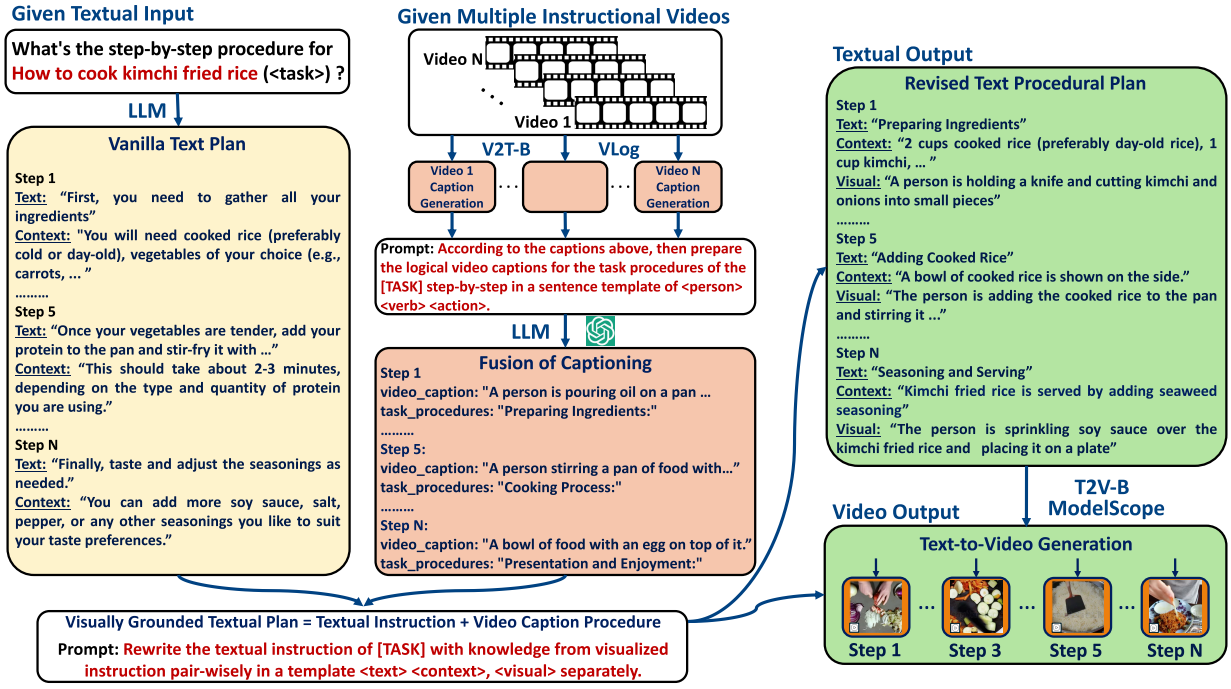

Figure 2: VG-TVP Model: Given the textual input and multiple instructional videos, VG-TVP generates visually grounded textual plans and video plans by using V2T-B and T2V-B. ChatGPT 3.5 is used to reorganize all captions to generate FoC.

T2V models lagged behind that of Text-to-Image (T2I) models (Ramesh et al. 2021), recent advances have enhanced T2V performances (Zhou et al. 2022; Singer et al. 2023; Khachatryan et al. 2023; Wu et al. 2023). Existing models lack robustness as exhibited by LLMs in text generation, highlighting the need for advanced LLM-based frameworks for MPP tasks. Unlike existing models, VG-TVP leverages LLM capabilities for MPP tasks, offering dynamic, and comprehensive guidance through video generation in a human-centered manner.

Integration of Visual Knowledge

IVs contain complex visual procedural knowledge that should be integrated into LLMs by reducing the complexity. Thus, it has been integrated with different methods such as usage of given images or generated texts as additional features (Yang et al. 2022; Lu et al. 2023b). However, image data results in inadequate descriptions of task actions since it provides only static information. To address that, VG-TVP harnesses IVs' captions to integrate visual knowledge into LLMs, leveraging their zero-shot reasoning capability. Moreover, in the PP tasks, integrated visual knowledge must be aligned for coherent outcomes (Shen, Wang, and Elhamifar 2021; Dvornik et al. 2023) which is a challenge for MPP tasks. To address that, we propose a method to combine video captions and vanilla text systematically.

VG-TVP Approach

We propose two key ideas in a methodology. The first idea involves generating MPPs for the tasks by exploiting their IVs, which is called "SEEN" (Idea 1). For the "SEEN" tasks, the model utilizes the relevant task IVs and captures their captions. Finally, video captions of these "SEEN" tasks and vanilla text plans are aligned to generate visually grounded text and video plans. The "UNSEEN" tasks (Idea 2) which the model has not previously encountered, involve generating MPPs for tasks that lack IVs. For example, "*Cooking Kimchi Fried Rice*" and "*Cooking Szechuan Chicken*" are two "SEEN" tasks under the Daily-PP dataset. We propose an exploration of "*Cooking Chicken Fried Rice*" as an "UNSEEN" task by utilizing the captions of "*Cooking Kimchi Fried Rice*" and *"Cooking Szechuan Chicken*".

Problem Statement

We formulate the problem as a visually grounded textual-video pairs alignment and generation task. G^V is given multiple IVs which represent high-level goal videos, and G^T is given textual input which denotes high-level goal texts, provided by the user in natural language. The model's output is a comprehensive multimodal procedural plan, represented as Goal Plan, G^P. It comprises the final sequence of visually-grounded text plan $TP = \{\{pt_1, pc_1\}, \{pt_2, pc_2\}, ..., \{pt_n, pc_n\}\}$ and video plan $VP = \{v_1, v_2, ..., v_n\}$ pairs. In this notation, pt_i represents the text of the final revised textual plan and pc_i denotes its corresponding context. Consequently, $G^P = \{TP, VP\}$ is generated by merging 3 multimodal features: vanilla text generation, video captioning via V2T-Bridge, and video generation via T2V-Bridge.

Method

We leverage the zero-shot reasoning ability of LLMs to generate vanilla text plans by proposing a step-by-step prompting template. To enhance the MPP, we propose the VG-TVP model (Figure 2), applying V2T-B and T2V-B for generating a comprehensive multimodal procedural plan.

Vanilla Text Plan Generation. The model generates vanilla text plans for the steps called Vanilla Textual Plan (VTP), VTP = $\{s_1, s_2, ..., s_n\}$ via vanilla prompt template function, $f_{prompt}(vanilla)$. This is a natural language format *"What's the step-by-step procedure for <[TASK]>?"* used to elicit information from LLMs shown in Figure 2. Inspired by the TIP approach (Lu et al. 2023b), we adopt the zero-shot chain-of-thought approach (Kojima et al. 2022) for VTP generation. For example, for *"Cooking Spaghetti"* task, the input text will be *"What's the step-by-step procedure for How to Cook Spaghetti?"*. Each step represents a VTP paired with text and context at timestamp i, $s_i = \{t_i, ct_i\}$. While $T = \{t_1, t_2, ..., t_n\}$ represents the VTP-text, $CT = \{ct_1, ct_2, ..., ct_n\}$ denotes VTP-context generation. The initial text plan VTP is derived using $f_{prompt}(vanilla)$ focused on the specified <[TASK]>.

Fusion of Captioning. We employ VLog model (Lin and Lei 2023) which integrates the capabilities of ChatGPT (OpenAI 2023), BLIP2 (Li et al. 2023), GRIT (Du, Rush, and Cardie 2021), Whisper (Radford et al. 2023) models. VLog generates video captions from 3 *modalities*: image, region, and audio. Our framework excludes audio captions as some videos lacked speech or significant auditory content. Thus, we follow the visual information. V2T-B is the method that textualizes the scenes of IVs by using a video captioning algorithm. For example, in the "apple juice" task, given that the $3rd$ caption of the $1st$ IV represents the step "wooden cutting board with sliced apples on it" between $17 - 26$ seconds (sec), similar steps appear in the $4th$ IV at $39 - 50$ sec, and the $7th$ IV at $9 - 19$ sec in the $2nd$ step (not in $3rd$). In other IVs' captions, steps similar to "pouring water into a blender" are included in different steps. Therefore, FoC reorders and aligns the steps for vanilla textual matching (Figure 15 in the Appendix). FoC is insufficient for completing the MPP alone. It provides additional visual information to VG-TVP for generating MPP tasks. Consequently, we formulate the V2T-B as FoC = $f_{captions} \oplus \{(G^V), f_{prompt}(description)\}$. The *Description* prompt is "*According to the captions above, then prepare the logical video captions for the task procedures of the <[TASK]> step-by-step in a sentence template of <person> <verb> <action>?"*. Detailed video captions and FoC can be found in the supplementary materials.

Visually-Grounded Textual and Video Plan Generation. There has been a shift toward leveraging LLMs for visual generations (Yu et al. 2023; Blattmann et al. 2023), moving beyond the conventional use of diffusion models. Specifically, (Lin et al. 2023) equipped with image annotations and a GPT4-based video planner, is prone to generating unnecessary instructions for PP tasks. For instance, *"A hungry cat is finding food."* prompt in (Lin et al. 2023) leads LLMs to segment the generated narrations into scenes. VG-TVP focuses on generating MPP to assist humans rather than storytelling or directing videos. VG-TVP's $GoalPlan$ is the final sequence of visually-grounded text plan (TP) and video plan (VP) pairs, $G^P = \{TP, VP\}$. However, multimodal generation brings an alignment challenge. Thus, we propose FoC collaborating with VG-TVP alignment prompt to revise VTPs and tackle this challenge.

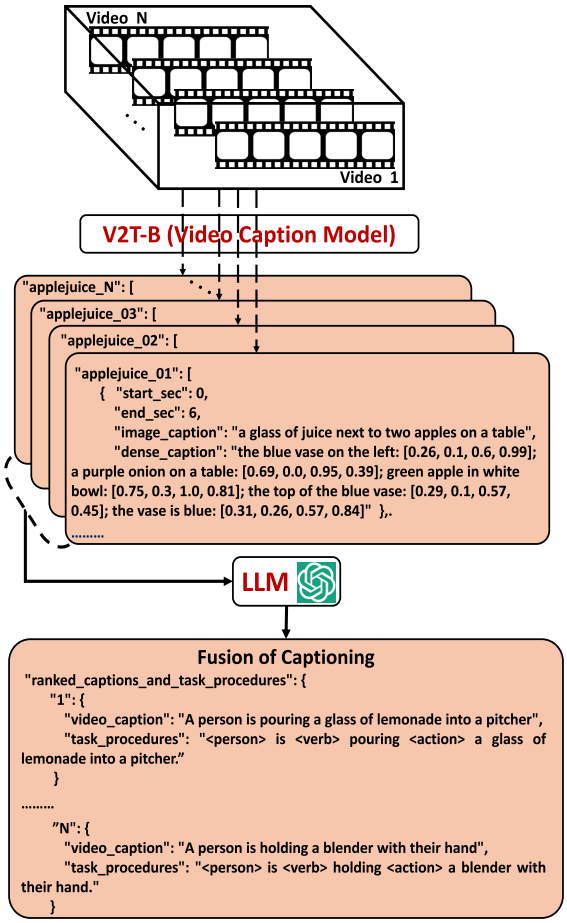

Figure 3: FoC captures and fuses IVs' captions. Then, it injects into the system by aligning them with vanilla textual.

Text-video alignment is essential for maintaining consistency and coherence between textual and visual plans. It synchronizes procedural steps across modalities, enabling users to correlate textual instructions with relevant generated videos and follow the instructions accurately. This aids comprehension and enhances the learning experience by offering unified and coherent instructions. We formulate the prompt alignment with the integration of VTPs and FoC via a prompt template: TP = $f_{prompt}(alignment) \oplus \{(FoC), (VTP)\}$. *Alignment* prompt represents "*Rewrite the step-by-step procedures of <[TASK]> by using video captions pair-wisely in a template <text>, <context> and <visual> separately.*". We follow this alignment to generate the relevant task videos which compose the Video Plan,

VG-TVP (Ours)	Textual Informative			Visual Informative			Temporal Coherence			Plan Accuracy		
versus	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Llama2-7B-q4	**44.00**	40.00	16.00	**74.00**	2.00	24.00	**68.00**	8.00	24.00	**74.00**	6.00	20.00
Llama2-7B-q8	**52.00**	22.00	26.00	**62.00**	18.00	20.00	**62.00**	24.00	14.00	**70.00**	12.00	18.00
Llama2-13B-q4	**40.00**	24.00	36.00	**46.00**	10.00	44.00	**42.00**	34.00	24.00	40.00	18.00	**42.00**
Llama2-13B-q8	20.00	**48.00**	32.00	**56.00**	18.00	26.00	36.00	**44.00**	20.00	**46.00**	22.00	32.00
Mistral-7B-q4	**58.00**	26.00	16.00	**56.00**	12.00	32.00	**52.00**	20.00	28.00	**48.00**	28.00	24.00
Mistral-7B-q8	**54.00**	22.00	24.00	**58.00**	14.00	28.00	**46.00**	34.00	20.00	**56.00**	16.00	28.00
GPT3.5	**40.00**	24.00	36.00	**56.00**	12.00	32.00	**54.00**	14.00	32.00	**56.00**	10.00	34.00
TIP Model	**71.43**	21.43	7.14	**57.14**	21.43	21.43	**42.86**	35.71	21.43	**42.86**	28.57	28.57

Table 1: Percentages of human evaluation comparisons between VG-TVP (Ours), baseline models by employing different LLMs (first 7 rows), and TIP (Lu et al. 2023b) for SEEN tasks. Win represents the preferred option for VG-TVP.

VP = (T2V-B) $\oplus$ (TP). VP is derived by employing T2V-B which exploits a sequence of visually-grounded text plan TP. Then, we generate relevant tasks' video plans by using ModelScope (Wang et al. 2023b). Consequently, a visually grounded approach leverages both text and visual information to tackle the alignment and coherency challenges.

Experiments

Baselines utilize text/context to generate text-aligned inclusive vanilla video plans. On the other hand, VG-TVP exploits vanilla text and FoC to generate visually grounded MPP content to assist individuals. We use a Win-Tie-Lose comparison on 50 seen and 15 unseen tasks, involving 28 human subjects for benchmarking. VG-TVP generated 2,504 videos for "seen" and 687 for "unseen" tasks, while baselines produced 2,701 and 681 vanilla textual videos, respectively. We included 2 tasks from WikiPlan&RecipePlan, TIP model (Lu et al. 2023b), to facilitate a fair benchmarking. Moreover, we conduct one more comparison with 2 different prompts (in the Appendix). This comparison aims to compare the effectiveness of injecting human orientation into text and video plans with VG-TVP against prompting. The qualitative results display that human orientation is more effectively integrated with VG-TVP. Additionally, we measure the textual relevance between generated baselines' and VG-TVP's text plans with reference text plans using BLEU (Papineni et al. 2002), and METEOR (Banerjee and Lavie 2005). Finally, we design an LLM (via ChatGPT4o) (with Socratic Method (Chang 2023)) evaluation protocol to evaluate baselines and VG-TVP on 4 aspects as in the human evaluation metric. Each scored out of 25 points, to assess the quality of task plans. The details are in the Appendix.

Daily-PP Dataset

Existing datasets are not suitable for the MPP content generation. Specifically, CrossTask (Zhukov et al. 2019), lacked specified task patterns, making it difficult to use for PP analysis. COIN (Tang et al. 2019), does not align with the structure needed for our research. Similarly, ProceL (Elhamifar and Naing 2019) required updates to meet the specific demands of IV analysis. WikiPlan&RecipePlan (Lu et al. 2023b) is not precisely tailored for following a structured task sequence. Thus, we curated a new dataset - called *Daily-Life Task Procedural Plans (Daily-PP)* - to better align with MPP content, drawing inspiration from the strengths and addressing the limitations of existing collections.

The Daily-PP consists of 5 domains (Breakfast, Dinner, Drink, Hobby&Crafts, and Home&Garage), 50 seen tasks, and 15 unseen tasks. Seen tasks include 7 or 10 IVs from YouTube, depending on the video density for each task. Moreover, 3 domains (Breakfast, Drink, and Dinner) have unseen tasks such as egg benedict, carrot mango lassi, and chicken fried rice. Unseen tasks are those without any IVs. The model generates their MPPs by using their vanilla text plan with 2 relevant seen tasks' video captions. For example, VG-TVP uses the video captions of *carrot juice* & *mango lassi* tasks to generate the MPP of *carrot mango lassi*. The Daily-PP dataset structure is shown in the Appendix.

Human Evaluation Metric

Existing metrics such as BLEU and METEOR evaluate text similarity by comparing generated texts with reference texts. However, they have limitations in MPP tasks. In cases where tasks do not have strict laws or steps, it cannot be assumed there is a single ground truth (GT). Daily life tasks such as "Hanging a Mirror", and "Cooking Pancakes" lack definitive instructions and can vary widely. They cannot be deemed as having only one correct GT. Thus, a human evaluation survey is the optimal metric for assessing tasks aimed at generating MPPs. An example of a survey is in the Appendix.

Results

Baselines. We employ 7 different LLMs to generate vanilla text plans (Touvron et al. 2023; Jiang et al. 2023; Brown et al. 2020) which are (1) LLama2-7B-q4, (2) LLama2-7B-q8, (3) LLama2-13B-q4, (4) LLama2-13B-q8, (5) Mistral-7B-q4, (6) Mistral-7B-q8 and (7) ChatGPT3.5.

Quantitative Analysis. The performance of VG-TVP in "SEEN" tasks is shown in Table 1. There are only two "TIE" results in the comparison between VG-TVP and Llama2-13B-q8, specifically for textual informativeness (48.00%) and temporal coherence (44.00%). The highest preference rate of VG-TVP is observed as 58.00% against Mistral-7B-q4. Benchmarking with the TIP model demonstrates that VG-TVP achieved better textual informativeness (71.43%). In terms of visual informativeness, VG-TVP's performance

VG-TVP (Ours) versus	Textual Informative			Visual Informative			Temporal Coherence			Plan Accuracy		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Llama2-7B-q4	**40.00**	26.67	33.33	**60.00**	26.67	13.33	**60.00**	26.67	13.33	**66.67**	20.00	13.33
Llama2-7B-q8	**66.67**	6.67	26.67	**53.33**	26.67	20.00	**53.33**	26.67	20.00	**60.00**	6.67	33.33
Llama2-13B-q4	26.67	**53.33**	20.00	**46.67**	13.33	40.00	**53.33**	26.67	20.00	**40.00**	33.33	26.67
Llama2-13B-q8	**53.33**	40.00	6.67	**60.00**	33.33	6.67	**80.00**	0.00	20.00	**66.67**	13.33	20.00
Mistral-7B-q4	**40.00**	**40.00**	20.00	**53.33**	26.67	20.00	**46.67**	46.67	6.67	**40.00**	53.33	6.67
Mistral-7B-q8	**73.33**	20.00	6.67	**86.67**	0.00	13.33	**66.67**	26.67	6.67	**73.33**	20.00	6.67
GPT3.5	13.33	**80.00**	6.67	**60.00**	6.67	33.33	20.00	**40.00**	**40.00**	26.67	33.33	**40.00**

Table 2: Percentages of human evaluation comparisons between VG-TVP (Ours) and baseline models by employing different LLMs (q: quantization) for UNSEEN tasks. WIN, and LOSE represent the better and worse results of VG-TVP, respectively.

Models	BLEU		METEOR	
	Base.	VG-TVP	Base.	VG-TVP
Llama2-7B-q4	0.013	0.013	**0.089**	0.082
Llama2-7B-q8	**0.014**	0.011	**0.082**	0.075
Llama2-13B-q4	**0.012**	0.011	0.067	**0.071**
Llama2-13B-q8	**0.014**	0.012	**0.082**	0.073
Mistral-7B-q4	0.012	**0.014**	0.091	**0.093**
Mistral-7B-q8	0.013	**0.015**	0.084	**0.094**
GPT3-5	0.017	**0.024**	0.072	**0.085**

Table 3: Automatic evaluations on 50 seen tasks from Daily-PP. Generated Baselines' (Base.) and VG-TVP's text plans are compared with the reference textual.

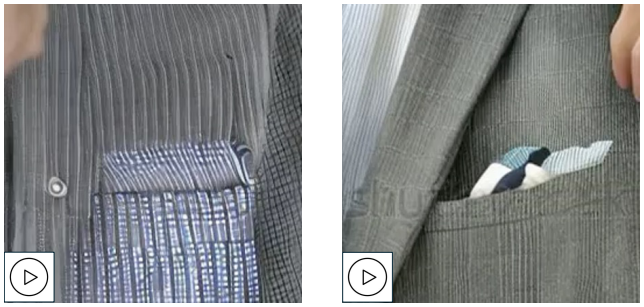

Figure 4: Impact of FoC on the task, "How to Fold the Presidential Pocket Square?". The right video shows well-displayed pocket square by employing the FoC.

ranges from 46.00% to 74.00%. VG-TVP, without losing any evaluation against the baseline models, achieves the highest average score in this challenge. VG-TVP generates better visual plans with a 57.14% preference score than the TIP model. Benchmarking with the TIP model, VG-TVP generates better visual plans with a 57.14% preference score. In temporal coherence, VG-TVP obtains a higher preference rate of 36.00% against Llama2-13B-q8, while achieving a preference parity of 44.00% among subjects. Benchmarking with the TIP model, VG-TVP reaches a better rate with 42.86%. In plan accuracy, VG-TVP's ratio is only 2.00% behind Llama2-13B-q4, yet it achieves successful ratios against other baselines. The benchmarking with the TIP, VG-TVP achieves more accurate plans with 42.86%.

For "UNSEEN" tasks, unlike the TIP model, VG-TVP can combine different task information to generate multimodal procedural plans for new, unseen content that lacks IVs or additional information. Thus, TIP model is not included in this benchmarking (Table 2). In terms of textual informativeness, VG-TVP is not preferred over Llama2-13B-q4 and GPT3.5. However, the majority of subjects rated the comparison as "TIE". In visual informativeness, VG-TVP achieves superior preference rates, ranging from 46.67% to 86.67%, against all baselines. Regarding temporal coherence and plan accuracy, VG-TVP obtains better ratios than all models except GPT3.5. Consequently, VG-TVP's superior performance can be seen from higher preferences over baselines across various tasks and models, particularly in visual informativeness, showing its effectiveness in MPP.

Table 3 shows the textual relevance evaluations of baseline and VG-TVP's text plans using BLEU and METEOR on 50 seen tasks from the Daily-PP dataset. Except for the LLama2-13B-q4 on the METEOR and the LLama2-7B-q4 on the BLEU, baseline-generated text plans outperform via LLama2-7B-q8, LLama2-13B-q4, and LLama2-13B-q8 models. For Mistral-7B-q4, Mistral-7B-q8, and GPT3.5, VG-TVP text plans perform better in both metrics. These results cannot show the semantic performance of the text plans because a single GT for daily life tasks is not adequate to cover the MPP content. Therefore, human evaluation is the optimum metric for MPP tasks to assist people.

Qualitative Analysis. Vanilla text plans successfully verbalize procedural information across many tasks. However, using vanilla text plans may not achieve the same efficiency for visuals while generating videos. We hypothesize that augmenting vanilla textual with IVs can enhance MPP, thereby more effectively assisting individuals. Our results and comparative analyses with VG-TVP and baselines confirm this hypothesis. We generate baseline models' videos using the text-context structure of vanilla textual. Figure 5 provides a qualitative analysis through a visual example using the Llama2-13B-q8+T2V. The unseen task, "How to bake kofta/meatballs and potatoes?" combines "Pan-Cooking Kofta/Meatballs" and "Cooking French Fries". Figure 5 above videos display vanilla text-video plan generations, where the baseline model outlines five steps to complete the task. In Figure 5 videos below, text-video plans are generated by VG-TVP, and video plans are highlighted

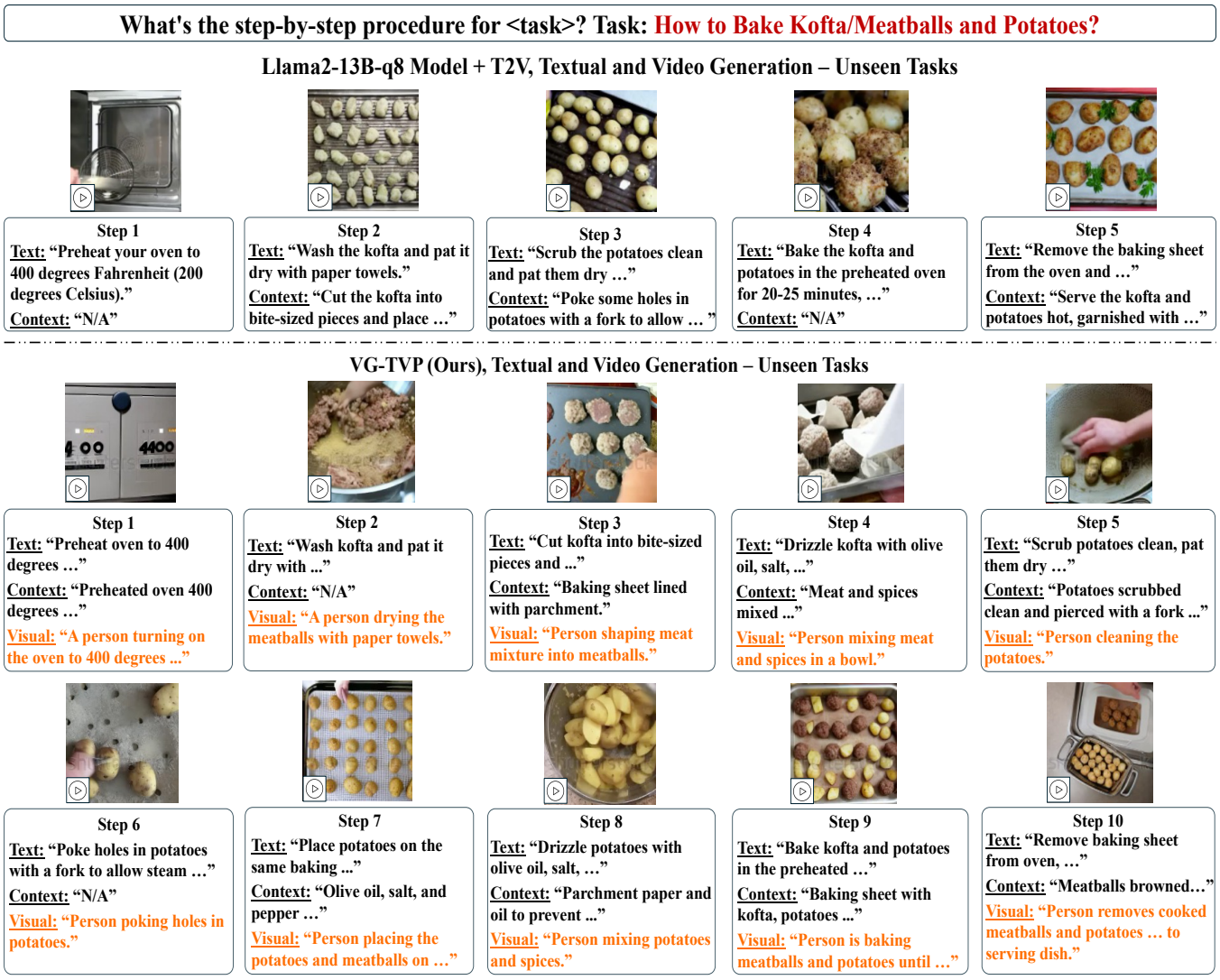

Figure 5: Qualitative comparison for an Unseen Task, between (above) Llama2-13B-q8 and (below) VG-TVP (Ours). Visuals are used to generate video plans. VG-TVP can increase the number of steps to generate the MPP more informative and accurate.

through visual prompts. VG-TVP enhances textual information, thereby improving the quality of generated video plans.

Ablations. FoC is the key technical component of VG-TVP which fuses IVs' captions to integrate into the model. For example, Figure 4 left video is generated based on *Text: "Finally, tuck the ends of the pocket square into your pocket to create a neat and tidy appearance"*, *Context: "Remember, the key to folding a pocket square is to be consistent and precise in your folds and to make sure the edges are aligned, and the corners are squared off."*. FOC leverages text to generate a "Visual" prompt as *"A person tucking the ends of a folded pocket square into their pocket, creating a neat and tidy appearance."* that is used to generate the video shown in Figure 4 (right). Unlike baseline's generation (Figure 4 (left)), VG-TVP generates a video with sharp lines of the suit jacket and folded square inside the pocket (Figure 4 (right). Consequently, FoC improves the VG-TVP's impact to generate more plan-accurate and informative visuals.

Conclusion and Future Work

This paper presents a novel approach to MPP through the development of LLM-powered frameworks, addressing the complexities of generating cohesive PPs for both seen and unseen tasks. By leveraging the capabilities of LLMs, VG-TVP enhances the coherence and consistency of PPs, demonstrating the efficacy of integrating visually grounded text and action-based video generations to enhance human assisting. Daily-PP dataset represents a significant stride towards overcoming the limitations of existing datasets, providing a more structured and comprehensive resource for evaluating MPP. VG-TVP may serve as an effective model for future research on leveraging multimodal information to enhance human learning experiences.

Acknowledgements

This research is supported by the Singapore International Graduate Award (SINGA) Scholarship.

References

Alayrac, J.; Bojanowski, P.; Agrawal, N.; Sivic, J.; Laptev, I.; and Lacoste-Julien, S. 2016. Unsupervised Learning from Narrated Instruction Videos. In *Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA*, 4575–4583. IEEE.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Association for Computational Linguistics.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada*, 22563–22575. IEEE.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; ...; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Annual Conference on Neural Information Processing Systems, NeurIPS*.

Chang, C.; Huang, D.; Xu, D.; Adeli, E.; Fei-Fei, L.; and Niebles, J. C. 2020. Procedure Planning in Instructional Videos. In *16th European Conference Computer Vision, ECCV, Glasgow,UK*, volume 12356 of *Lecture Notes in Computer Science*, 334–350. Springer.

Chang, E. Y. 2023. Prompting Large Language Models With the Socratic Method. arXiv:2303.08769.

Du, X.; Rush, A. M.; and Cardie, C. 2021. GRIT: Generative Role-filler Transformers for Document-level Event Entity Extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL*, 634–644. Association for Computational Linguistics.

Dvornik, N.; Hadji, I.; Zhang, R.; Derpanis, K. G.; Wildes, R. P.; and Jepson, A. D. 2023. StepFormer: Self-Supervised Step Discovery and Localization in Instructional Videos. In *Conference on Computer Vision and Pattern Recognition, CVPR, Vancouver, BC, Canada*, 18952–18961. IEEE.

Elhamifar, E.; and Naing, Z. 2019. Unsupervised Procedure Learning via Joint Dynamic Summarization. In *2019 International Conference on Computer Vision, ICCV, Seoul, Korea (South)*, 6340–6349. IEEE.

Fang, F.; Liu, Y.; Koksal, A.; Xu, Q.; and Lim, J. 2023. Masked Diffusion with Task-awareness for Procedure Planning in Instructional Videos. *CoRR*, abs/2309.07409.

Huang, W.; Abbeel, P.; Pathak, D.; and Mordatch, I. 2022. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In *International Conference on Machine Learning, ICML, Baltimore, Maryland, USA*, volume 162, 9118–9147. PMLR.

Ilaslan, M. F.; Song, C.; Chen, J.; Gao, D.; Lei, W.; Xu, Q.; Lim, J.; and Shou, M. 2023. GazeVQA: A Video Question Answering Dataset for Multiview Eye-Gaze Task-Oriented Collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 10462–10479. Singapore: Association for Computational Linguistics.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In *International Conference on Computer Vision, Paris, France*, 15908–15918. IEEE.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Annual Conference on Neural Information Processing Systems, NeurIPS, New Orleans, USA*. Curran Associates Inc.

Kuehne, H.; Arslan, A. B.; and Serre, T. 2014. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *Conference on Computer Vision and Pattern Recognition, CVPR, Columbus, OH, USA*, 780–787. IEEE.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Lian, L.; Shi, B.; Yala, A.; Darrell, T.; and Li, B. 2023. LLM-grounded Video Diffusion Models. *CoRR*, abs/2309.17444.

Lin, H.; Zala, A.; Cho, J.; and Bansal, M. 2023. VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning. *CoRR*, abs/2309.15091.

Lin, K. Q.; and Lei, S. W. 2023. VLog: Video as a Long Document. https://github.com/showlab/VLog. Accessed:2024-12-15.

Lu, Y.; Feng, W.; Zhu, W.; Xu, W.; Wang, X. E.; Eckstein, M. P.; and Wang, W. Y. 2023a. Neuro-Symbolic Procedural Planning with Commonsense Prompting. In *The Eleventh International Conference on Learning Representations, ICLR, Kigali, Rwanda*.

Lu, Y.; Lu, P.; Chen, Z.; Zhu, W.; Wang, X. E.; and Wang, W. Y. 2023b. Multimodal Procedural Planning via Dual Text-Image Prompting. *CoRR*, abs/2305.01795.

Miech, A.; Zhukov, D.; Alayrac, J.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision, ICCV, Seoul, Korea (South)*, 2630–2640. IEEE.

Niu, Y.; Guo, W.; Chen, L.; Lin, X.; and Chang, S.-F. 2024. SCHEMA: State Changes Matter for Procedure Planning in Instructional Videos. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning, ICML, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 28492–28518. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139, 8821–8831. PMLR.

Shen, Y.; Wang, L.; and Elhamifar, E. 2021. Learning To Segment Actions From Visual and Language Instructions via Differentiable Weak Sequence Alignment. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 10156–10165. IEEE.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; and Taigman, Y. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations, ICLR, Kigali, Rwanda*.

Song, C. H.; Sadler, B. M.; Wu, J.; Chao, W.; Washington, C.; and Su, Y. 2023. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In *International Conference on Computer Vision, ICCV, Paris, France*, 2986–2997. IEEE.

Soucek, T.; Damen, D.; Wray, M.; Laptev, I.; and Sivic, J. 2024. GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. In *Conference on Computer Vision and Pattern Recognition,CVPR, Seattle, WA, USA*, 6561–6571. IEEE.

Tang, Y.; Ding, D.; Rao, Y.; Zheng, Y.; Zhang, D.; Zhao, L.; Lu, J.; and Zhou, J. 2019. COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis. In *Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA*, 1207–1216. IEEE.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; ...; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Wang, H.; Wu, Y.; Guo, S.; and Wang, L. 2023a. PDPP: Projected Diffusion for Procedure Planning in Instructional Videos. In *Conference on Computer Vision and Pattern Recognition, CVPR, Vancouver, BC, Canada*, 14836–14845. IEEE.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023b. ModelScope Text-to-Video Technical Report. *CoRR*, abs/2308.06571.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *International Conference on Computer Vision, ICCV, Paris, France*, 7589–7599. IEEE.

Yang, A.; Nagrani, A.; Seo, P. H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; and Schmid, C. 2023. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In *Conference on Computer Vision and Pattern Recognition, CVPR, Vancouver, BC, Canada*, 10714–10726. IEEE.

Yang, Y.; Yao, W.; Zhang, H.; Wang, X.; Yu, D.; and Chen, J. 2022. Z-LaVI: Zero-Shot Language Solver Fueled by Visual Imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, Abu Dhabi, UAE*, 1186–1203. Association for Computational Linguistics.

Yu, L.; Lezama, J.; Gundavarapu, N. B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Gupta, A.; Gu, X.; Hauptmann, A. G.; Gong, B.; Yang, M.; Essa, I.; Ross, D. A.; and Jiang, L. 2023. Language Model Beats Diffusion - Tokenizer is Key to Visual Generation. *CoRR*, abs/2310.05737.

Zhao, H.; Hadji, I.; Dvornik, N.; Derpanis, K. G.; Wildes, R. P.; and Jepson, A. D. 2022. P^3IV: Probabilistic Procedure Planning from Instructional Videos with Weak Supervision. In *Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, USA*, 2928–2938. IEEE.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. MagicVideo: Efficient Video Generation With Latent Diffusion Models. *CoRR*, abs/2211.11018.

Zhou, H.; Martín-Martín, R.; Kapadia, M.; Savarese, S.; and Niebles, J. C. 2023. Procedure-Aware Pretraining for Instructional Video Understanding. In *Conference on Computer Vision and Pattern Recognition, CVPR, Vancouver, BC, Canada*, 10727–10738. IEEE.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA*, 7590–7598. AAAI Press.

Zhu, L.; and Yang, Y. 2020. ActBERT: Learning Global-Local Video-Text Representations. In *Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA*, 8743–8752. IEEE.

Zhukov, D.; Alayrac, J.; Cinbis, R. G.; Fouhey, D. F.; Laptev, I.; and Sivic, J. 2019. Cross-Task Weakly Supervised Learning From Instructional Videos. In *Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA*, 3537–3545. IEEE.