

# Multimodal Procedural Planning via Dual Text-Image Prompting

Yujie Lu<sup>1</sup>, Pan Lu<sup>2</sup>, Zhiyu Chen<sup>1</sup>, Wanrong Zhu<sup>1</sup>, Xin Eric Wang<sup>3</sup>  
William Yang Wang<sup>1</sup>

<sup>1</sup>University of California, Santa Barbara, CA, USA  
{yujielu,zhiyuchen,wanrongzhu,wangwilliamyang}@ucsb.edu

<sup>2</sup>University of California, Los Angeles, CA, USA  
lupantech@gmail.com

<sup>3</sup>University of California, Santa Cruz, CA, USA  
xwang366@ucsc.edu

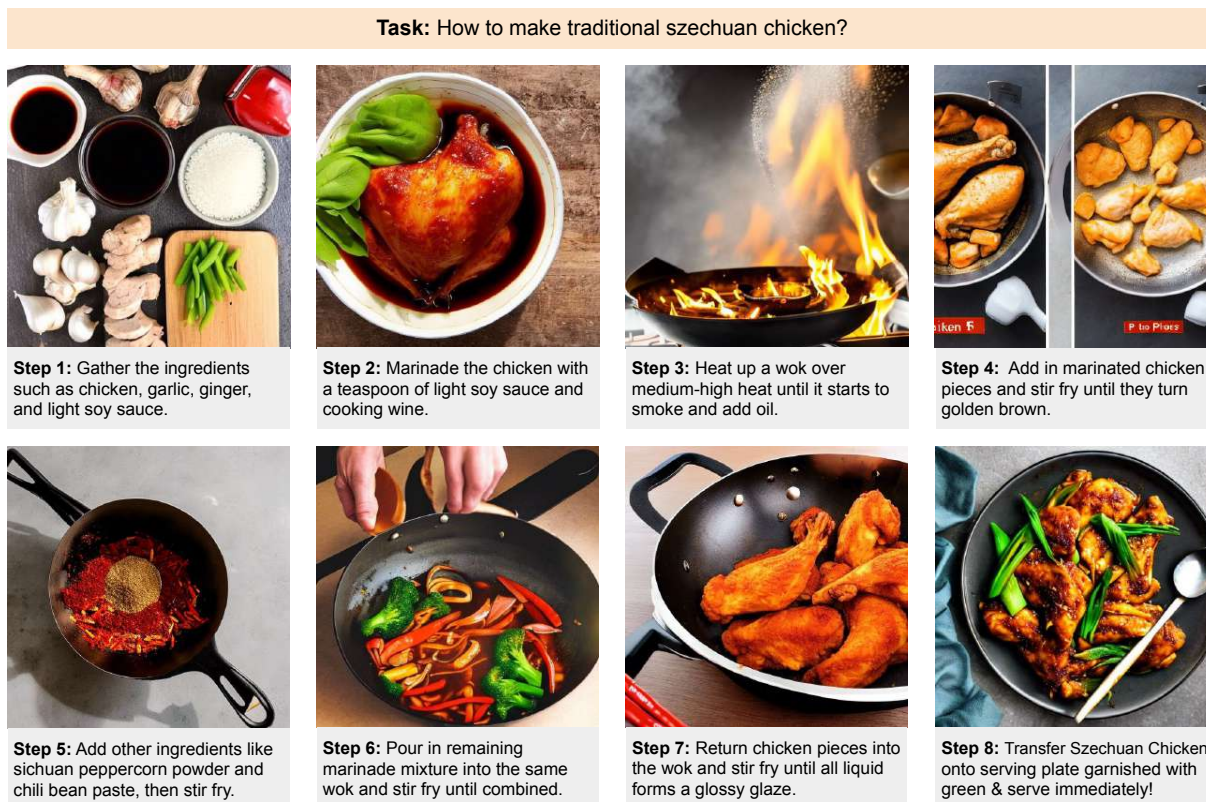


Figure 1: Our dual Text-Image Prompting (TIP) model generates coherent and authentic *multimodal procedural plans* with multiple steps towards a high-level goal, providing useful guidelines in task completion.

## Abstract

Embodied agents have achieved prominent performance in following human instructions to complete tasks. However, the potential of providing instructions informed by texts and images to assist humans in completing tasks remains underexplored. To uncover this capability, we present the multimodal procedural planning (MPP) task, in which models are given a high-level goal and generate plans of paired text-image steps, providing more complementary and informative guidance than unimodal plans. The key challenges of MPP are to ensure the informativeness, temporal coherence, and accuracy of plans across modalities. To tackle this, we propose Text-Image Prompting (TIP), a dual-modality prompting method

that jointly leverages zero-shot reasoning ability in large language models (LLMs) and compelling text-to-image generation ability from diffusion-based models. TIP improves the interaction in the dual modalities using Text-to-Image Bridge and Image-to-Text Bridge, allowing LLMs to guide the textual-grounded image plan generation and leveraging the descriptions of image plans to ground the textual plan reversely. We collect WIKIPLAN and RECIPEPLAN as a testbed for MPP. Our results show compelling human preferences and automatic scores against unimodal and multimodal baselines on WIKIPLAN and RECIPEPLAN in terms of informativeness, temporal coherence, and plan accuracy. <sup>1</sup>

<sup>1</sup>Our code and data are provided in supplemental materials.

# 1 Introduction

Recent advances in embodied (Huang et al., 2022; Anderson et al., 2018) and conversational (Qiu et al., 2021) agents achieve prominent performance in task completion as humans by following instructions informed by texts and images. However, to what extent the models can provide useful guidelines for humans to complete the task remains underexplored. To uncover this, we propose the multimodal procedural planning task (as shown Figure 1). The task aims to generate goal-conditioned (e.g. “How to make traditional szechuan chicken”) text (e.g. “a teaspoon of light soy sauce” explain how to marinate chicken in Step 2) and image (e.g. help identify the ingredients “chicken, garlic, ginger, and light soy sauce” in Step 1) plans as useful guidelines to assist humans in task completion.

Previous work (Huang et al., 2022) explored the generation of procedural plans in text-only form. In contrast, we generate both text and image plans, which provide guidance to perform tasks that acquire complementary information from multimodal contexts. Generating plans in both text and image form poses new challenges since the generated plans should: a) be *informative* enough in both the text and image modalities, b) obey commonsense temporal *coherence*, such as the order of steps, and c) achieve high plan *accuracy*, indicating the complementary and alignment among multimodal plans.

Despite significant progress (Kojima et al., 2022; Song et al., 2022) in the development of large language models (LLMs), they are unable to generate images. Existing text-to-image (T2I) models can generate high-quality images conditioned on textual instructions (Ramesh et al., 2022; Rombach et al., 2022; Brooks et al., 2022). However, they are limited in their ability to generate images that require complex text comprehension, such as temporal reasoning (e.g. “learn basic surf safety *before* hitting the waves”) and physical reasoning (e.g. “*pick up* the wine glass”). Additionally, generating text and image plans separately using LLMs and T2I models results in inconsistency and incoherence between the two modalities.

In this paper, we propose Text-Image Prompting (TIP), a novel dual-modality prompting framework that jointly leverages the capabilities of LLMs and T2I models for multimodal procedural planning. We first generate vanilla text plans by querying LLMs (Kojima et al., 2022) for step-by-step proce-

dures. To generate textual-grounded image plans, we devise the Text-to-Image Bridge (T2I-B), which elicits the complex language comprehension abilities of LLMs to assist T2I models in generating informative and text-aligned image plans. We generate visual-grounded text plans using the Image-to-Text Bridge (T2I-B), which verbalizes the image plans and injects them back into LLMs to assist the text plans revision, thereby improving informativeness. The temporal coherence of the generated plans is improved considering the context of both text and image. Benefiting from dual-modality prompting, generated plans are complementary and aligned across text and image modalities.

To address the lack of suitable datasets for evaluating multimodal procedural planning, we collect the WIKIPLAN and RECIPEPLAN datasets for benchmarking the task. We empirically evaluate the effectiveness of TIP on WIKIPLAN and RECIPEPLAN in a zero-shot setting and compare it with various baselines. Our results demonstrate that TIP generate plausible multimodal plans that are informative, temporally coherent, and accurate. Our work highlights the potential of combining knowledge from LLMs and T2I models to uncover multimodal zero-shot planning capabilities. Our main contributions are as follows:

- We introduce and benchmark the multimodal procedural planning task using our collected WIKIPLAN and RECIPEPLAN datasets.
- We propose Text-Image Prompting (TIP), a dual-modality prompting approach that elicits procedural knowledge jointly from LLMs and T2I models, enabling visual-grounded text plans and textual-grounded image plans.
- We show that TIP substantially improves performance in terms of textual and visual informativeness, temporal coherence, and plan accuracy.

## 2 Related work

**Procedural Planning** This task (Zhang et al., 2020; Chang et al., 2020) has gain much attention in various aspects, including robotics (Tellex et al., 2011; Jansen, 2020; Brohan et al., 2022), vision-and-language navigation (Anderson et al., 2018), conversational assistants (Ilievski et al., 2018; Qiu et al., 2021, 2022; Yang et al., 2022a), and animation (Zhao et al., 2022). Recent work is extended to the multimodal scenarios (Wu et al., 2022; Song et al., 2022; Wang et al., 2022c). In this work,

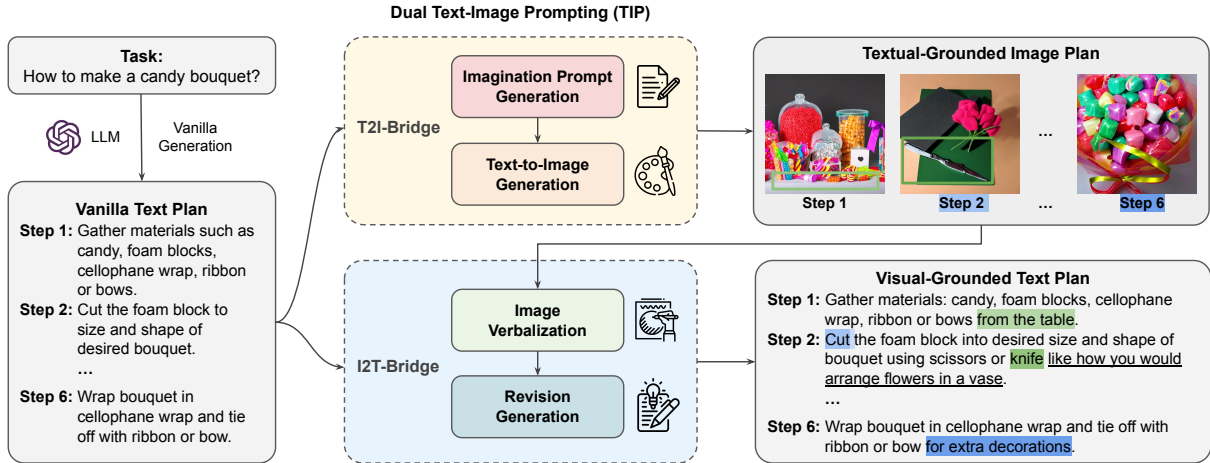


Figure 2: The vanilla text plan is generated using LLM. Our Text-Image Prompting (TIP) generates the textual-grounded image plan using T2I-Bridge (Fig. 3 Left) and the visual-grounded text plan using I2T-Bridge (Fig. 3 Right). The colors blue and green highlight the improved grounding in text and image respectively.

we explore the multimodal procedural planning that generates goal-conditioned text and image sequences grounded in a multimodal context.

**Multimodal Generative Models** Diffusion models (Ramesh et al., 2022; Rombach et al., 2022) have shown remarkable abilities in generating high-quality images given text prompts. However, generating images with desired semantics requires proper prompts. To get more controllable generations, researchers have used large language models (LLMs) to expand input prompts with rich contextual knowledge. InstructPix2Pix (Brooks et al., 2022) combines the knowledge of GPT-3 and Stable Diffusion to generate large-scale examples of image editing as training data. In turn, large-scale models based on transformers (Li et al., 2022; Wang et al., 2022a) exhibit incredible ability in image captioning, describing the given image using natural language. Our work empowers the visual generation with reasoning from LLMs to enable zero-shot multimodal procedural planning.

### 3 Our Approach

#### 3.1 Problem Definition

We formulate multimodal procedural planning as a conditional text and image sequence generation problem. Given a high-level goal  $\mathcal{G}$  in natural language form, the model generates a sequence of low-level steps  $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ . Each step  $s_i$  in the sequence is represented by a paired text  $t_i$  and image  $v_i$  at timestep  $i$ . The text plan  $\{t_1, t_2, \dots, t_n\}$  and image plan  $\{v_1, v_2, \dots, v_n\}$  are both intended to be informative in their respective modalities and

complementary across modalities. The final multimodal procedural plans ( $\mathcal{S}$ ) is the combination of the text plan and image plans, which describe the procedure of completing the high-level goal.

#### 3.2 Method Overview

We first elicit the zero-shot step-by-step reasoning ability in large language models (LLMs) to generate a vanilla text-only plan (left part in Figure 2). To enable grounding in multimodal context, we propose Text-Image Prompting (TIP), a dual-modality prompting method (middle part in Figure 2) upon LLMs and multimodal generative models: (1) Text-to-Image Bridge (T2I-B): we generate the visual imaginative prompt that translates the complex textual instructions (vanilla plan in Figure 3) into explicit scene descriptions (prompt in Figure 3) for text-to-image models. (2) Image-to-Text Bridge (I2T-B): we verbalize the image plan with the image captioning model for generating prompts (red highlighted template in Figure 3) that elicit the revision ability of LLMs with awareness of context. Figure 2 depicts how TIP implements multimodal procedural planning by connecting LLMs and multimodal generative models (Image Caption Model, Text-to-Image Model) with our T2I-B and I2T-B, grounding the image plans in textual context and the text plan in visual context respectively (right part in Figure 2).

#### 3.3 Vanilla Text Plan Generation

We first elicit procedural knowledge of LLM to generate vanilla text plan using Zero-shot Chain-of-Thought (Kojima et al., 2022) that does not re-

quire heavy human-engineered few-shot examples. Specifically, we leverage InstructGPT (Ouyang et al., 2022) to generate a goal-conditioned step-by-step procedure with the template “[*TEMPLATE*] Task: [*GOAL*]?”. [*TEMPLATE*] represents the hand-crafted template to extract the procedural knowledge from LLM. We extend the template “Let’s think step by step” (proposed in (Kojima et al., 2022)) as “What’s the step-by-step procedure of” for procedural planning. Then we replace the input slot [*GOAL*] with the given task name  $T$  (the high-level goal description) as the prompt  $P$  to be fed into the LLM. The LLM then outputs goal-conditioned subsequent steps  $\mathcal{W} = \{t_1, t_2, \dots, t_n\}$  using greedy decoding as our initial textual plan, which is conditioned only on the task name  $T$  in zero-shot generation manner.

### 3.4 Textual-Grounded Image Plan Generation with Text-to-Image Bridge

Our Text-to-Image Bridge (T2I-B) in Figure 3 leverages LLM to bridge the gap between the language understanding capabilities of LLM and the ability of language-conditioned image generation in the text-to-image model. T2I-B elicits visual imagination in LLM to generate explicit scene description (imagination prompt) for text-to-image model conditioned on the vanilla plan.

**Imagination Prompt Generation** We encourage LLM to revise the prompt that already processes the physical or temporal meaning residing in the original textual plan. To access this, for each step, we use the prompt  $P_{t2i}$  “[*STEP*] [*T2I-B*]” that concatenates the original generated textual plan at step  $i$  and the Text-to-Image Bridge template. [*STEP*] represents one of the subsequent steps generated from LLMs. For [*T2I-B*], we use the trigger sentence similar to “What do I need to draw in the picture to describe the above text?”. With this Text-to-Image Bridge guided prompt  $P_{t2i}$ , the text-to-image model then generates the textual grounded image at each timestep to compose the final sequence of visual plan  $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$ .

**Text-to-Image Generation** We exploit the Stable Diffusion (Rombach et al., 2022) model to generate RGB images at  $512 \times 512$  resolution. Figure 4 provides examples of text-to-image generation with and without our T2I-B. Benefiting from the existing knowledge in LLMs, the text-to-image models can generate semantically relevant and high-fidelity images based on the already processed prompt.

### 3.5 Visual-Grounded Text Plan Generation with Image-to-Text Bridge

To enhance the completeness, alignment, and knowledge exchange between the generated text and image plans, we propose revising the vanilla text plan using the textual-grounded image plan.

**Image Verbalization** To complete this, we first transfer the visual plan into a natural language format and then inject it into LLM. We implement this by generating captions for each visual plan. Given the image  $v$ , the captioning model BLIP (Li et al., 2022) generates captions, which transfer the visual knowledge into textual descriptions. For each generated visual plan  $v_i$  at each timestep  $i$ , we generate such pairwise caption with  $caption = G(v, Desc)$ , where  $Desc$  denotes the task description for unified vision and language models, “what does the image describe” in our case. With these image captions, we can further transfer the visual-grounded information into LLM and revise our textual plan.

**Revision Generation** To ground the textual plan in visual context, we use the verbalized description of the visual plan to concatenate with our Image-to-Text Bridge template similar to “Let’s revise the procedure using the captions”. Concretely, we concatenate the initial textual plan, the captions of the visual plan, and the Image-to-Text Bridge template as the prompt  $P_{i2t}$  “Step-by-step Procedure: [*INITIAL*] Captions: [*CAPTION*] [*I2T-B*]”. In this way, we elicit the reasoning ability of LLMs to ground the textual plan in verbalized visual context, as depicted in Figure 3 To this end, our generated multimodal plan is bi-directional grounded by connecting the abilities of LLMs and multimodal generative models.

## 4 Multimodal Procedural Planning Dataset Collection

Our datasets are collected and repurposed from WIKIHOW<sup>2</sup> and RECIPEQA (Yagcioglu et al., 2018) due to their temporal relatedness among texts and images. We collect WIKIPLAN by crawling the household “how to” articles from WIKIHOW and then repurpose them into a multimodal procedural planning dataset by formulating the article title as the task name and content as the textual steps, with the pictures as the visual steps. We collect RECIPEPLAN from RECIPEQA dataset for multimodal procedural planning by sequencing all the given text-image pairs as the text and image

<sup>2</sup><https://www.wikihow.com>

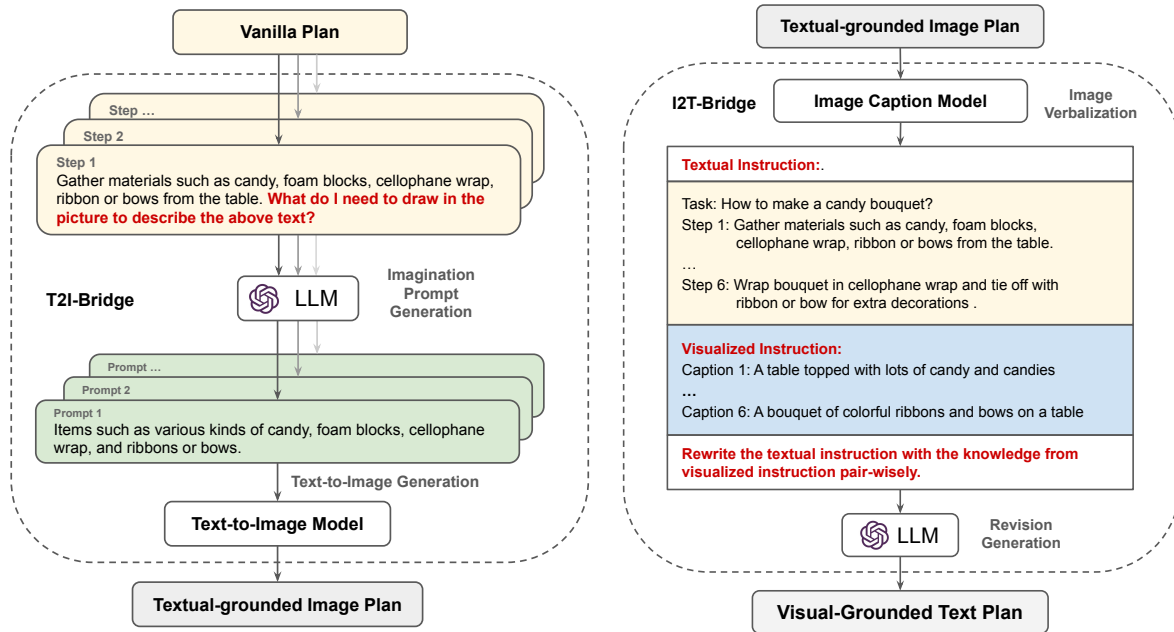


Figure 3: (a) Left: T2I-B elicits visual imagination in LLM to generate explicit scene description (imagination prompt) for text-to-image model conditioned on the vanilla plan. (b) Right: I2T-B injects verbalization of the image plans to foster revision generation of visual-grounded text plans with awareness of multimodal context.



(a) w. T2I-B (b) w.o. T2I-B (c) w. T2I-B (d) w.o. T2I-B

Figure 4: Text-to-image generation showcases with (a) (b) on “before hitting the waves, read up on ocean safety tips and know the rules of the beach” and (c) (d) on “put down the wine glass” with or without T2I-B.

plan correspondingly, with the main title as the task name. We conduct zero-shot experiments on 1,000 distinct, randomly sampled tasks from each dataset. More dataset details are in Appendix C.

## 5 Experiments

### 5.1 Evaluation Metrics

We conduct head-to-head comparisons using Amazon Mechanical Turk (AMT) platform (details can be found in Appendix D.1) on four aspects: (1) Textual Informativeness: the text plans contain the necessary information to complete the task, (2) Visual Informativeness: the image plans contain the necessary information to complete the task, (3) Temporal Coherence: the multimodal plans meet the temporal commonsense requirements, such as the order in which the steps occur, (4) Planning Accuracy: whether referring to the multimodal plans can successfully assist task com-

pletion.

In addition, we measure semantic relevance between predicted text plans and reference text plans using Word Mover’s Distance (WMD) (Kusner et al., 2015), Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). We evaluate image plans using FID (Heusel et al., 2017) and CLIPScore (Hessel et al., 2021; Radford et al., 2021). We calculate Caption-Sentence-BERT (Cap-S) and Text-Sentence-BERT (Text-S) scores by comparing predicted image and text plans respectively to reference text using S-BERT. The average of these yields the All-Sentence-BERT (ALL-S) score for multimodal plans. Evaluations are conducted at a procedure level.

More details of evaluation can be found in Appendix D.

### 5.2 Baselines

As we are exploring a novel task, that require generating the image and text plans conditioned on the high-level goals, no existing models can be directly applied. To validate the effectiveness of our designed dual bridges, we devise several intuitive baselines for comparison: (1) ImageRef + OFA/BLIP-Caption: use image sequences from each article/recipe as the image plans references (ImageRef), and extract captions of image refer-

Dataset	Ours vs. Model	Textual-Informativeness			Visual-Informativeness			Temporal Coherence			Plan Accuracy		
		Win(↑)	Tie	Lose(↓)	Win(↑)	Tie	Lose(↓)	Win(↑)	Tie	Lose(↓)	Win(↑)	Tie	Lose(↓)
WIKIPLAN	Image Ref + OFA-Caption	<b>63.34</b>	18.38	18.27	<b>60.63</b>	20.45	18.92	<b>61.95</b>	21.03	17.02	<b>61.99</b>	19.40	18.61
	Image Ref + BLIP-Caption	<b>62.70</b>	18.70	18.60	<b>61.26</b>	21.18	17.56	<b>62.22</b>	20.78	17.00	<b>62.29</b>	18.28	19.43
	Text Ref + DALLE	<b>62.61</b>	20.34	17.06	<b>59.88</b>	22.38	17.74	<b>60.53</b>	22.08	17.40	<b>61.19</b>	22.07	16.74
	Text Ref + Stable-Diffusion	<b>62.58</b>	19.82	17.60	<b>60.25</b>	21.16	18.58	<b>60.68</b>	22.38	16.94	<b>61.73</b>	20.56	17.72
	Text-Davinci-002 + Stable-Diffusion	<b>60.68</b>	21.56	17.76	<b>59.90</b>	20.41	19.70	<b>60.22</b>	22.99	16.79	<b>60.41</b>	21.53	18.06
	Text-Davinci-003 + Stable-Diffusion	<b>62.32</b>	19.82	17.86	<b>60.29</b>	20.85	18.85	<b>61.10</b>	22.17	16.73	<b>61.48</b>	20.29	18.23
RECIPEPLAN	Image Ref + OFA-Caption	<b>64.51</b>	18.29	17.20	<b>62.39</b>	20.18	17.43	<b>62.74</b>	20.40	16.86	<b>63.66</b>	19.19	17.15
	Image Ref + BLIP-Caption	<b>64.81</b>	18.58	16.61	<b>62.29</b>	19.60	18.11	<b>62.70</b>	20.72	16.58	<b>62.90</b>	19.08	18.02
	Text Ref + DALLE	<b>61.16</b>	20.15	18.69	<b>59.60</b>	20.60	19.80	<b>60.04</b>	20.48	19.48	<b>62.11</b>	19.21	18.68
	Text Ref + Stable-Diffusion	<b>61.31</b>	19.81	18.87	<b>60.49</b>	20.37	19.14	<b>60.37</b>	20.33	19.31	<b>62.38</b>	18.81	18.81
	Text-Davinci-002 + Stable-Diffusion	<b>62.50</b>	19.33	18.17	<b>60.59</b>	18.12	21.29	<b>61.24</b>	21.13	17.63	<b>62.30</b>	17.38	20.31
	Text-Davinci-003 + Stable-Diffusion	<b>62.65</b>	19.26	18.09	<b>61.10</b>	20.00	18.90	<b>61.46</b>	20.60	17.94	<b>62.85</b>	18.75	18.40

Table 1: Percentages of multimodal procedural planning results of TIP that are better than, tied with, or worse than baselines, on sampled 200 distinct tasks from each dataset.

Dataset	Model	Text Plan				Image Plan		Multimodality Plan			Step Length
		WMD	S-BERT	ROUGE-L	METEOR	FID ↓	CLIP ↑	Cap-S	Text-S	ALL-S	Avg.
WIKIPLAN	Image Ref + BLIP-Caption	0.78	0.35	0.06	0.04	-	0.71	<u>0.36</u>	0.41	<u>0.39</u>	8.26
	Image Ref + OFA-Caption	0.86	0.27	0.07	0.06	-	0.71	<u>0.27</u>	<u>0.48</u>	<u>0.37</u>	8.26
	Text Ref + DALLE	0.68	<u>0.76</u>	<u>0.28</u>	<u>0.12</u>	<u>47.39</u>	0.74	0.33	0.26	0.29	8.26
	Text Ref + Stable-Diffusion	0.68	0.76	0.28	0.12	56.64	0.73	0.34	0.26	0.30	8.26
	Text-Davinci-002 + Stable-Diffusion	<u>0.87</u>	0.65	0.10	0.06	61.17	0.50	0.33	0.25	0.28	4.70
	Text-Davinci-003 + Stable-Diffusion	0.86	0.67	0.11	0.08	57.87	0.70	0.33	0.27	0.30	6.68
	<b>TIP (Ours)</b>	<b>0.90</b>	<b>0.67</b>	<b>0.12</b>	<b>0.09</b>	<b>48.82</b>	<b>0.78</b>	<b>0.34</b>	<b>0.28</b>	<b>0.31</b>	6.75
	Image Ref + BLIP-Caption	0.77	0.37	0.08	0.05	-	0.64	0.42	<u>0.56</u>	<u>0.49</u>	6.93
Image Ref + OFA-Caption	0.82	0.40	0.09	0.10	-	0.64	0.43	0.48	0.46	6.93	
RECIPEPLAN	Text Ref + DALLE	0.21	0.59	0.10	0.09	<u>53.55</u>	0.63	0.46	0.40	0.43	6.93
	Text Ref + Stable-Diffusion	0.21	0.59	0.10	0.09	54.58	0.61	<u>0.48</u>	0.40	0.44	6.93
	Text-Davinci-002 + Stable-Diffusion	0.84	0.63	0.11	0.10	60.11	0.49	0.44	0.33	0.38	5.17
	Text-Davinci-003 + Stable-Diffusion	0.85	0.68	0.12	0.13	60.07	0.73	0.42	0.35	0.38	6.82
	<b>TIP (Ours)</b>	<b>0.86</b>	<b>0.68</b>	<b>0.13</b>	<b>0.14</b>	<b>58.68</b>	<b>0.79</b>	<b>0.43</b>	<b>0.36</b>	<b>0.40</b>	6.94

Table 2: Automatic evaluations on 2,000 distinct tasks from WIKIPLAN and RECIPEPLAN. Image Ref and Text Ref baselines use image and text title references from the dataset. Our TIP uses Text-Davinci-003 and Stable-Diffusion as the LLM and T2I model. We underline and bold highest score of models with and without reference baselines.

ences as text plans using image caption models (e.g., OFA or BLIP) (2) TextRef + DALLE/Stable-Diffusion: use titles from each article/recipe as text plans references (TextRef), and use text-to-image models (e.g., DALLE or Stable-Diffusion) to generate images as image plans conditioned on text references (3) Text-Davinci-002/003 + Stable-Diffusion: given the high-level goal, first prompt the LLMs (e.g., Text-Davinci-002/003) to generate text plans, and then use text-to-image models to generate image plans conditioned on the text plans (4) Text-Davinci-003 (Step-based) + Stable-Diffusion: all the variants are in default procedure-based, however this variant instead of generating the plan at the procedure level, it generates one step of the text plans at a time iteratively. These variants are using LLMs and T2I models separately without collaboration.

### 5.3 Quantitative Analysis

**Human Evaluation Results** We conduct Win-Tie-Lose Comparison between TIP and the baselines over WIKIPLAN and RECIPEPLAN. Averaged results from 200 tasks rated by 3 crowdsourcing per example are reported in Table 1. Across four aspects, TIP receives consistently higher preferences, outperforming the baselines over the winning ratio by over 60%. In terms of textual informativeness, the unimodal baselines (Image Ref + OFA-Caption and Image Ref + BLIP-Caption) is slightly worse than the unimodal text reference-based baseline (Text Ref + Stable-Diffusion and Text Ref + DALLE) and multimodal baselines (Text-Davinci-003 + Stable-Diffusion and Text-Davinci-002 + Stable-Diffusion). This is mainly due to the other baselines either directly leveraging the textual information from the reference or the rich text-based knowledge in LLMs. In terms of visual informativeness, the multimodal baselines (Text-Davinci-003 + Stable-Diffusion and Text-Davinci-002 +

Text-to-Image Bridge Template	Alignment		Image-to-Text Bridge Template	Alignment	
	WIKIPLAN	RECIPEPLAN		WIKIPLAN	RECIPEPLAN
<i>What do I need to draw in the picture to describe the above text?</i>	<b>0.9625</b>	<b>0.9595</b>	<i>Rewrite the textual instruction with the knowledge from visualized instruction pair-wisely.</i>	0.7644	0.6945
What do you see in the figure?	0.9366	0.9397	Based on the visual caption, can you revise the step-by-step procedure according to the paired captions?	<b>0.8011</b>	0.6205
Describe what the picture corresponding to the text should have.	0.9070	0.9181	Revise each step according to the visual imagination.	0.6921	0.7329
Let's think about what we need to visualize to present the above idea.	0.8986	0.8941	Let's revise the procedure using the captions.	0.6155	<b>0.7691</b>
<i>Describe something irrelevant to the above text.</i>	0.5598	0.5325	<i>What's the procedure that disobey the captions?</i>	0.5079	0.5902
<i>What do you usually draw?</i>	0.5350	0.4826	<i>Provide an interesting procedure to be irrelevant with the captions.</i>	0.1519	0.163

Table 3: Robustness check of various templates used in both Text-to-Image Bridge and Image-to-Text Bridge over WIKIPLAN and RECIPEPLAN dataset. The *italic* templates are misleading examples. Our Text-Image Prompting model chooses the template with averaged best multimodal alignment, highlighted in purple.



(a) Multimodal procedural plan generated by baseline Text-Davinci-003 + Stable-Diffusion.



(b) Multimodal procedural plan generated by our Text-Image Prompting (TIP).

Figure 5: Improved grounding in textual and visual context are highlighted in pink and green respectively. Red texts indicate reasoning of physical action in image plan generation.

Stable-Diffusion) can not achieve on-par results with textual reference-based baseline. We hypothesize this is due to the lack of visual knowledge injected into LLMs. The performance gain of TIP over multimodal baselines (Text-Davinci-003 + Stable-Diffusion and Text-Davinci-002 + Stable-Diffusion) imply the importance of grounding our

multimodal plans in a multimodal context.

**Automatic Evaluation Results** In Table 2, TIP achieves consistent improvement over baselines (without Ref), and even surpasses the baselines using reference from the dataset on RECIPEPLAN. This further confirms our superiority in generating multimodal plans with semantic correctness and

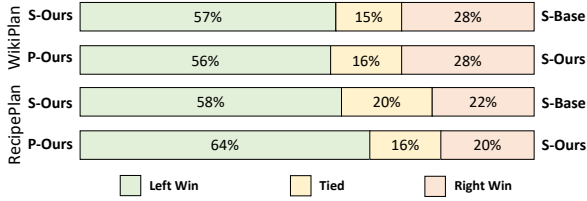


Figure 6: Step-based (S) vs. Procedure-based (P) Win-Tie-Lose over Plan Accuracy.

Plan w.o. I2T-B	WIKIPLAN	RECIPEPLAN
	Avg. Textual	Avg. Textual
Imagination Prompt	0.341 (-18.4%)	0.363 (-14.1%)
Image Verbalization	0.261 (-37.5%)	0.273 (-35.4%)

Table 4: Ablation of Image-to-Text Bridge on TIP over text plan generation.

Model	WIKIPLAN		RECIPEPLAN	
	FID ↓	Align ↑	FID ↓	Align ↑
DALLE	119.03	0.77	83.27	0.64
+ Text-to-Image Bridge	<b>117.02</b>	<b>0.79</b>	<b>67.64</b>	<b>0.78</b>
Stable Diffusion	129.13	0.74	88.17	0.62
+ Text-to-Image Bridge	<b>119.74</b>	<b>0.78</b>	<b>84.37</b>	<b>0.78</b>

Table 5: Ablation of Text-to-Image Bridge with Text Ref + Stable-Diffusion on single-step image plan generation.

alignment. Notice that Text Ref baselines directly use the title from the dataset, which is a summarized version of the main content. These are used as golden reference in automatic evaluations).

**Template Robustness** In Table 3, we compare various similar templates for T2I-B and I2T-B against misleading templates. The Alignment is measured with CLIP (Radford et al., 2021) to capture the similarity between given text/image and conditionally generated image/text. The poor alignment of misleading templates and similar alignment of various bridge templates prove the robustness of the template choice in the experiments.

**Single Bridge Effect** We report the effects of Text-to-Image Bridge and Image-to-Text Bridge of TIP in Table 4. The performance drop indicates that the text plan without condition on visual information is vulnerable in text-only planning quality. In Table 5, we also observe using T2I-B brings obvious improvement over both FID score and Alignment of image plans. These results further indicate that the key ingredient of our proposed method TIP is that LLMs and multimodal generative models will collaboratively generate multimodal procedural plans benefiting from our designed dual bridges.

**Step-based or Procedure-based** We explore our procedure-based method (P-Ours) against the step-based TIP (S-Ours) and step-based Text-Davinci-003 + Stable-Diffusion (S-Base). The main difference from step-based approach is the plan is generated one step at a time. Since LLMs achieve promising long text reasoning capability, we generate the full procedure for efficiency consideration. We report the head-to-head comparison results in Figure 6. The procedure-based method achieves 60% win rate over the step-based TIP. We observe this is partially due to the instinct of LLMs to repeat input texts and is less clear to understand the full intent of generation expectation. Thus the procedure-based method usually achieves better planning quality at the very beginning.

## 5.4 Qualitative Analysis

In Figure 5, we compare the performance of TIP to baselines in multimodal procedural planning. TIP generate image plans that are grounded in the textual context. With the help of LLMs reasoning in the temporal dimension, we transfer this ability to image generation, conditioning on the revised prompts of LLMs. This allows digestion of the temporal and complex reasoning present in the text plan and directly indicates what needs to be depicted in the image. The highlighted steps of image plans correctly visualize the scene described in the textual context. For example, at Step 2, instead of only showing the vegetables, ours show an image of a person washing the produce thoroughly. TIP also generate text plans that are better grounded in the image plan. The text plan correctly refers to the objects in visual input, such as “liquid form” and “blended mixture”, and also complements the visual context, such as “extract juice from it”. This indicate the potential for uncovering multimodal reasoning capabilities in LLMs. More comparisons are provided in Appendix E.1.

## 6 Conclusions

We introduce the Multimodal Procedural Planning task that aims to generate goal-conditioned text and image subsequences and benchmark models’ performance with our curated testbed WIKIPLAN and RECIPEPLAN. Our proposed dual-modality prompting framework enables plausible multimodal procedural plan generation.



## References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEE Evaluation@ACL*.
- Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. In *6th Annual Conference on Robot Learning*.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.
- Tuhin Chakrabarty, Arkady Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2022. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *OpenReview Preprint*. Preprint under review.
- Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. 2020. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.
- Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.
- Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*.
- Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.
- Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6629–6640, Red Hook, NY, USA. Curran Associates Inc.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.
- Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*.
- Vladimir Ilievski, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2018. Goal-oriented chatbot dialog management bootstrapping with transfer learning. In *IJCAI*.
- Peter Jansen. 2020. Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4412–4417.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). *arXiv preprint*.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. [From word embeddings to document distances](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.
- Liang Qiu, Yuan Liang, Yizhou Zhao, Pan Lu, Baolin Peng, Zhou Yu, Ying Nian Wu, and Song-Chun Zhu. 2021. Socaog: Incremental graph parsing for social relation inference in dialogues. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human utility. In *The 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.
- Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2022. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *arXiv preprint arXiv:2212.04088*.
- Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.
- Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.
- Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. 2020. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*.
- Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1507–1514.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.
- Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.
- Qingyun Wang, Manling Li, Hou Pong Chan, Lifu Huang, Julia Hockenmaier, Girish Chowdhary, and Heng Ji. 2022c. Multimedia generative script learning for task planning. *arXiv preprint arXiv:2208.12306*.
- Zhenhailong Wang, Manling Li, Ruochen Xu, Luwei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022d. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*.
- Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022e. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*.
- Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2022. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4525–4542.
- Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.
- Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368.
- Shiquan Yang, Rui Zhang, Sarah Erfani, and Jey Han Lau. 2022a. An interpretable neuro-symbolic reasoning framework for task-oriented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4918–4935.
- Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022b. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3081–3089.
- Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. 2021. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*.
- Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aavek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.
- Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842.
- Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with wikihow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639.
- Yizhou Zhao, Liang Qiu, Wensi Ai, Pan Lu, and Song-Chun Zhu. 2022. Triangular character animation sampling with motion, emotion, and relation. *arXiv preprint arXiv:2203.04930*.
- Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810.
- Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815.

## A Background

Text-to-image generation is a task that synthesizes images from text prompts. DALL-E 2 (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022) are state-of-the-art text-to-image models developed on top of diffusion models conditioned on input texts. Some early work in text-to-image models trains generative adversarial networks (GANs) (Goodfellow et al., 2014) on image captioning datasets (Xu et al., 2018; Zhu et al., 2019; Tao et al., 2020; Zhang et al., 2021; Ye et al., 2021) to generate images conditioned on textual descriptions. Other work follows the VQ-VAE (Van Den Oord et al., 2017) framework and trains autoregressive transformers that take both the text and image as sequences of tokens (Ramesh et al., 2021; Ding et al., 2021; Gafni et al., 2022). However, these methods are struggling to generate photorealistic images. Motivated by the remarkable progress of diffusion models in generating images with fidelity (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020), recent work has applied them to text-to-image generation with auxiliary text encoders (Rombach et al., 2022; Nichol et al., 2021; Gu et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). (Wang et al., 2022e) propose the first large-scale text-to-image prompt dataset, DiffusionDB, which enables a new research direction of prompt engineering to construct better prompts. (Chakrabarty et al., 2022) use GPT-3 to generate a detailed textual description with rich visual metaphors to prompt the DALL-E 2 model.

Recently, there has been a trend of using large language models (LLMs) like GPT-3 (Brown et al., 2020) to transfer visual knowledge in order to improve their capabilities in downstream natural language processing (NLP) and multimodal tasks. For example, images and videos can be translated into captions which further instruct a language model to generate contextual descriptions (Wang et al., 2022d; Zeng et al., 2022) or answer knowledge-based visual questions (Yang et al., 2022b). Instead of being prompted with textual descriptions, language models can extend to vision-language settings through text generation controlled by visual features (Cho et al., 2021; Tsimpoukelli et al., 2021; Su et al., 2022; Zhu et al., 2022; Wang et al., 2022b; Alayrac et al., 2022).

## B Method Details

### B.1 Configurations

The experiments using Text-Davinci and DALLE are conducted with OpenAI API on January 2023. We use BLIP w/ ViT-B and CapFilt-L<sup>3</sup> and OFA-base from huggingface demo<sup>4</sup>.

### B.2 Details of Module Outputs

We showcase all the details of the outputs of each module for the example task “How to make a candy bouquet” in Figure 7. In addition, we showcase examples on “How to make peppermint patty pudding shot” and “How to make bartending for beginners simple margarita” in Figure 8. We visualize the output details of each module in Figure 7 and Figure 8.

## C Dataset Details

### C.1 RECIPEPLAN

**Data Repurpose** RECIPEQA was proposed in (Yagcioglu et al., 2018) that provide four tasks (Textual Cloze, Visual Cloze, Visual Ordering, Visual Coherence) for multimodal machine comprehension of cooking recipes. This dataset contains question-answer pairs generated from copyright-free recipes. Each of them is under a license, which is provided in each data JSON file. We collect RECIPEPLAN by repurposing the test dataset from RECIPEQA that relates to the Visual Ordering task as the sequence generation task for the multimodal procedural planning evaluation testbed. We use recipe instructions as textual plan reference and their paired images as visual plan reference.

**Dataset Statistics** We visualize two examples of our repurposed RECIPEPLAN for multimodal procedural planning in Figure 9. We also show the word-cloud distribution of task name and textual plan reference in Figure 11.

### C.2 WIKIPLAN













**Raw Data Collection** To facilitate research on learning to generate procedural planning in a multimodal setting, we have constructed the large-scale WIKIPLAN dataset collected from the WIKIHOW website<sup>5</sup>, which is under an Attribution-Noncommercial-Share Alike 3.0 Creative Commons License.. This website provides a wide range

<sup>3</sup><https://github.com/salesforce/BLIP>

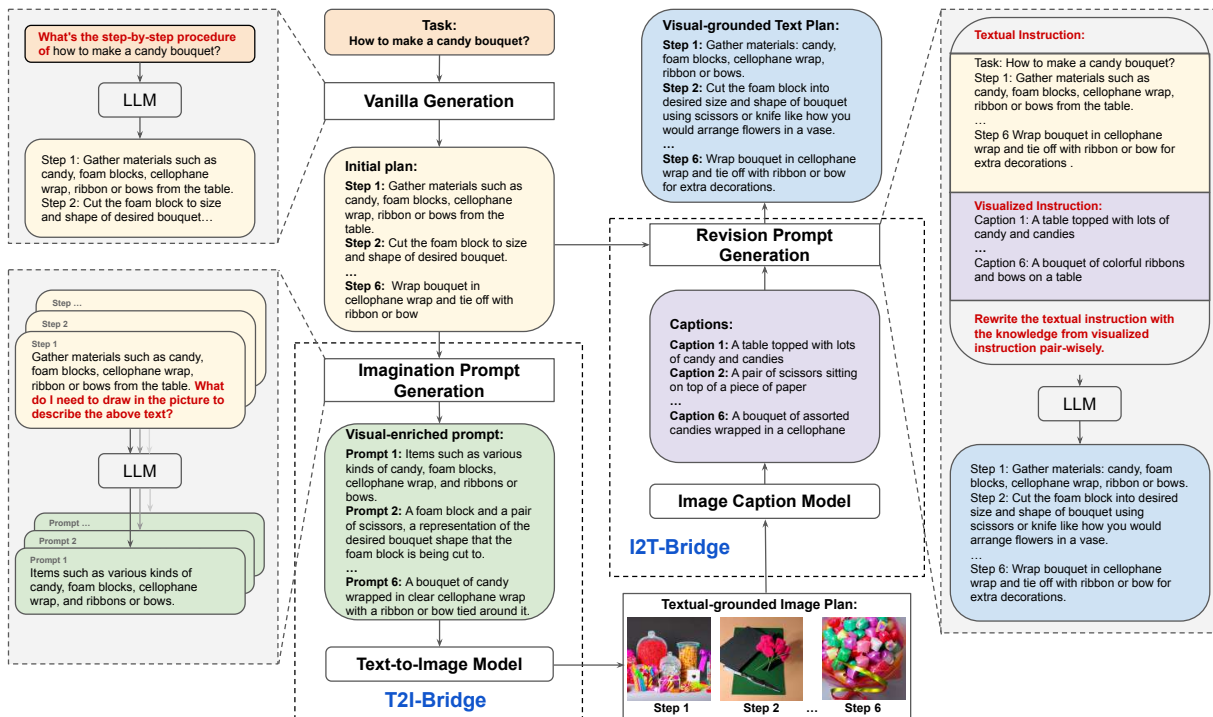
<sup>4</sup><https://huggingface.co/OFA-Sys/ofa-base>

<sup>5</sup><https://www.wikihow.com/>

**Task:** How to make a candy bouquet?

Vanilla Text plan	Vanilla Image Plan	Imagined Prompt	Textual-Grounded Image Plan	Verbalization	Visual Grounded Text plan
<b>Step 1:</b> Gather materials: candy, foam blocks, cellophane wrap, ribbon or bows.		Items such as various kinds of candy, foam blocks, cellophane wrap, and ribbons or bows.		A table topped with lots of candy and candies	<b>Step 1:</b> Gather materials such as candy, foam blocks, cellophane wrap, ribbon or bows from the table.
<b>Step 2:</b> Cut the foam block to size and shape of desired bouquet.		A foam block and a pair of scissors, a representation of the desired bouquet shape that the foam block is being cut to.		A pair of scissors sitting on top of a piece of paper	<b>Step 2:</b> Cut the foam block into desired size and shape of bouquet using scissors or knife like how you would arrange flowers in a vase.
<b>Step 3:</b> Insert wooden skewers into each piece of candy.		Wooden skewers inserted into each piece of candy, arrows pointing to the skewers to indicate that they are being inserted into the candy.		A group of toothbrushes sitting on top of a table	<b>Step 3:</b> Insert wooden skewers into each piece of candy as if they were lollipops on top of a blue table.
<b>Step 4:</b> Arrange the pieces of candy onto the foam in a pleasing pattern.		A bouquet of candy in a pleasing pattern. The candy should be arranged on top of a foam base. The candy should be arranged in an aesthetically pleasing pattern, such as alternating colors, sweet and		A bouquet of colorful candies in a vase	<b>Step 4:</b> Arrange the pieces of candy onto the foam in an interesting pattern like playing with colored rocks on plate and cupcakes and candies for your bouquet decoration ideas.
<b>Step 5:</b> Secure pieces with hot glue if necessary.		A bouquet of candy pieces that are being held together with hot glue, a person holding a hot glue gun, to show that the pieces of candy are being secured with hot glue.		A pile of different colored candies and lollipops	<b>Step 5:</b> Secure pieces with hot glue if necessary just like arranging colorful ribbons and bows around the bouquets centerpiece.
<b>Step 6:</b> Wrap bouquet in cellophane wrap and tie off with ribbon or bow.		A bouquet of candy wrapped in clear cellophane wrap with a ribbon or bow tied around it.		A bouquet of assorted candies wrapped in a cellophane	<b>Step 6:</b> Wrap bouquet in cellophane wrap and tie off with ribbon or bow for extra decorations.

(a) Full example outputs.



(b) Visualization of outputs from each module.

Figure 7: Output Table Details of each module of TIP for Multimodal Procedural Planning. The T2I-B leverage the complex language comprehension and zero-shot reasoning ability of LLMs to improve text-to-image generation. Reversely, the I2T-B injects visual knowledge via verbalization of the visual plans to generate a visually-grounded and complementary textual plan.

**Task:** How to make peppermint patty pudding shot?

Vanilla Text plan	Vanilla Image Plan	Imagined Prompt	Textual-Grounded Image Plan	Verbalization	Visual Grounded Text plan
<b>Step 1:</b> Combine 1/2 cup of peppermint schnapps, 1/4 cup of Bailey's Irish cream, and 2 tablespoons of chocolate syrup in a bowl or blender.		A bowl or blender filled with 1/2 cup of peppermint schnapps, 1/4 cup of Bailey's Irish cream, and 2 tablespoons of chocolate syrup.		Two cups of hot chocolate with spoons on a table	<b>Step 1:</b> Place three cups of chocolate pudding on a flat surface and add 1/2 cup of peppermint schnapps, 1/4 cup of Baileys Irish Cream, and 2 tablespoons of chocolate syrup.
<b>Step 2:</b> Add 4 ounces of cold heavy whipping cream to mixture and blend until smooth.		A bowl with the mixture in it, a spoon in the mixture, a measuring cup with 4 ounces of heavy cream, a blender, and a glass with the peppermint patty pudding shot.		A table topped with three different types of desserts.	<b>Step 2:</b> Using a spoon or blender mix all ingredients until smooth.
<b>Step 3:</b> Place 4 small glasses on a flat surface and fill with pudding shot mix.		Four small glasses on a flat surface, each filled with a pudding shot mix. You should also include any relevant ingredients, such as peppermint extract, liqueur, and cream, to complete the pudding.		A table topped with three shot glasses filled with liquid.	<b>Step 3:</b> Pour the mixture into four small glasses and top each with whipped cream.
<b>Step 4:</b> Chill for at least 30 minutes before serving.		A picture of a tray of peppermint patty pudding shots in the refrigerator, with a timer counting down the 30 minutes.		A tray filled with chocolate cups and candy canes.	<b>Step 4:</b> Add a candy cane for garnish on each glass and chill for at least 30 minutes before serving.

**Task:** How to make bartending for beginners simple margarita?









Vanilla Text plan	Vanilla Image Plan	Imagined Prompt	Textual-Grounded Image Plan	Verbalization	Visual Grounded Text plan
<b>Step 1:</b> Gather the following ingredients: tequila, triple sec (or Cointreau), lime juice, and simple syrup.		A glass of margarita with a lime wedge garnish, and next to it illustrate the ingredients needed to make the margarita - a bottle of tequila, bottle of triple sec or Coint		A bottle of orange juice next to limes and lime slices.	<b>Step 1:</b> Gather the following ingredients: tequila, triple sec (or Cointreau), lime juice, and simple syrup.
<b>Step 2:</b> Fill a shaker with ice cubes.		A shaker filled with ice cubes.		A glass filled with ice and lime slices.	<b>Step 2:</b> Fill a shaker with ice cubes.
<b>Step 3:</b> Pour in 1 ounce of tequila, 1/2 ounce of triple sec or Cointreau, 1/4 ounces of fresh lime juice, and 1/2 ounce of simple syrup into the shaker filled with ice cubes.		A shaker filled with ice cubes, a bottle of tequila, a bottle of triple sec or Cointreau, a bottle of fresh lime juice, and a bottle of simple syrup.		A glass of lemonade next to a bottle of lemonade.	<b>Step 3:</b> Pour in 1 ounce of tequila, 1/2 ounce of triple sec or Cointreau, 1/4 ounces of fresh lime juice into the shaker filled with ice cubes and shake well until all ingredients are combined and chilled (about 20 seconds).
<b>Step 4:</b> Shake well until all ingredients are combined and chilled (about 20 seconds).		A person shaking a cocktail shaker filled with ice cubes, limes, tequila, and triple sec. You could also add a timer set to 20 seconds to illustrate how long the drink should be shaken.		A person pouring a drink into a glass.	<b>Step 4:</b> Add half an ounce of simple syrup before straining the mixture into a glass rimmed with salt or sugar-lime mix (optional).
<b>Step 5:</b> Strain the mixture into a glass rimmed with salt or sugar-lime mix (optional).		A glass with a rim lined with either salt or a sugar-lime mix and a pitcher or shaker filled with a mixture of margarita. A spoon or strainer.		A glass of water with a lime next to it.	<b>Step 5:</b> Garnish with a wedge of lime or orange slice before serving cold margarita in your desired glass!
<b>Step 6:</b> Garnish with a wedge of lime or orange slice before serving cold margarita to your guests!		A glass filled with a cold margarita, a wedge of lime or orange slice on the side of the glass, and possibly some straws.		A glass of orange juice with a straw and lime.	<b>Step 6:</b> Enjoy your delicious homemade drink!

Figure 8: Full example of outputs details of TIP for Multimodal Procedural Planning.

of how-to articles related to everyday life topics, which are collaboratively written by its users and reviewed by experts. We crawled each article, collecting the task title, URL, introduction, topics, and steps. Each step includes a brief textual action, a detailed context, and an illustration image. Our

raw dataset consists of 30,026 examples across 19 categories and 2,062 topics. We plan to release the raw data in the hopes of pre-training models for procedural planning and knowledge reasoning.

**Quality Control** To improve the evaluation of different baselines, we further selected five categories

Task: How to make baileys hot chocolate mason jar.							
<b>Step 1: Gather Your Supplies.</b> Wide mouth, pint sized mason jars Hot chocolate powder Mini-marshmallows Twine Baileys mini bottles - one for each jar you plan to make.	<b>Step 2: Wash Your Mason Jars.</b> Make sure the jars are clean and free of debris. Dry thoroughly before moving on to the next step.	<b>Step 3: Add Hot Cocoa Mix.</b> Pour the cocoa mix into the jar. I added 2 bags of powder so it can fill up to a little less than half of the jar.	<b>Step 4: Marshmallows.</b> I'm a bit marshmallow fanatic, so I'm friendly with the amount I add. Feel free to add more or less, depending on your preference. Then, close the jar.	<b>Step 5: Wrap Twine Around the Lid.</b> Cut a long piece of twine. Firmly wrap the twine several times around the lid, being careful not to let it slip down. Once you have it firmly wrapped around, tie a knot. Leave remaining twine to attach mini bottle.	<b>Step 6: Tie the Mini Baileys Bottle.</b> With the remaining twine, tie a simple knot around the mini bottle of Baileys. I would double knot as well, around the mini bottle to be sure it is firm and tightly attached to the mason jar. Cut any excess twine.	<b>Step 7: This Is What One Bottle Looks Like!</b> Only a few items needed, to make a fast, unique gift! A person receiving this gift would only need to add hot water or milk and pour in some yummy Baileys to their liking!	<b>Step 8: Perfect for the Holidays!</b> This idea is a great gift for any occasion, but it's a great idea for the holidays coming up. You can even add a greeting card or any other embellishment you would like!
							






Task: How to make vegan lavender shortbread cookies.						
<b>Step 1: Prepare the Ingredients.</b> Ingredients (for 20 cookies): 1/2 cup (115g) margarine at room temperature (you can also use coconut oil); 1 cup (130g) all-purpose white flour; 1/4 cup (30g) powdered sugar; ...	<b>Step 2: Chop the Lavender and Cream the Margarine.</b> Preheat oven to 350°F / 180°C (gas mark 4) and prepare a cookie tray. Line the tray with parchment paper. In a small bowl sift together the flour and salt. ...	<b>Step 3: Add the Dry Ingredients.</b> Add the powdered sugar to the large bowl and beat until smooth and creamy. I like to sift my powdered sugar before adding it to the bowl. ...	<b>Step 4: Form a Log.</b> At this point the mixture should resemble to wet sand. Continue mixing with your fingers until you form a ball of dough. Place the dough between two cling film rectangles and shape it into a log. ...	<b>Step 5: Cut the Cookies.</b> Remove the log from the freezer and discard the cling film. Flour the working surface. Trim the rounded edges of the log. With superficial slits, divide the log into 5mm-thick cookies. ...	<b>Step 6: Refrigerate and Bake the Cookies.</b> Place the shortbread cookies on the parchment paper-lined tray making sure you space the cookies at least 5 cm apart. Place the tray in the refrigerator for 15 minutes. ...	<b>Step 7: Enjoy!</b> Store the cookies in an airtight container at room temperature for up to 1 week. Note: For this recipe I used margarine. You can also use vegan butter or solidified coconut oil, refined or unrefined. ...
						

Figure 9: Two examples in the RECIPEPLAN dataset.

that feature temporal actions and high-fidelity visual contexts: *Food and Entertaining*, *Hobbies and Crafts*, *Home and Garden*, *Pets and Animals*, and *Sports and Fitness*. In order to reflect common tasks in real-life scenarios, articles with fewer than three steps or more than 22 steps were excluded, as well as articles with images of a dimension size of fewer than 400 pixels. Each category was balanced with 200 examples. To further ensure high quality, we conducted a quality control in which well-trained human annotators reviewed the dataset and manually revised the examples if there was any wrong or inappropriate content.

**Dataset Statistics** Finally, our WIKIPLAN dataset consists of 1,000 examples across 5 categories and 370 different topics. Three examples of the dataset are illustrated in Figure 10, which include muddling mint leaves for a cocktail, encouraging a cat to eat, and becoming a better football player. Each example is composed of a title, introduction, related topics, and a list of detailed steps with visual aids. The word-cloud distributions of the task titles and step text are visualized in Figure 12, which

demonstrate the diversity of topics and words.

## D Evaluation Details


### D.1 Crowdsourcing Human Evaluation






We manually ensure no personal information is collected and no offensive content is presented during human evaluations. The hourly wage paid to participants is estimated at \$12. And the total amount spent on participant compensation is \$1958.

We average the results from 3 annotators for each example. Given the high-level goal (task name) for each assignment, we want the annotators to compare two generated text and image sequences in terms of *Textual-Informativeness*, *Visual-Informativeness*, *Temporal Coherence* and *Plan Accuracy*. Before going to the question, we let the annotators read the instructions below:

**Instruction:** Given the Task (e.g, Task: How to muddle), please compare two sequences of steps Sequence 1 and Sequence 2, and determine which sequence is better in terms of four aspects:

- Textual-Informativeness: whether the textual

<b>Task: How to Muddle</b>				
<b>Introduction:</b> Muddling is a bartender's technique for releasing flavor from solid cocktail ingredients. The basic idea is as simple as crushing the fruit or herb, but there are devils in the details ready to make your cocktail bitter or unpleasant. Learn how to do it the right way, and you won't have to muddle through on your own intuition.				
<b>Topics:</b> Spirits and liqueurs, Cocktails				
<b>Step 1: Choose a gentle muddling tool.</b> A small, flat wooden tool is ideal, such as the end of a wooden spoon or a French rolling pin (without handles). Plastic or hard rubber tools will also work, but require a delicate touch. Avoid muddlers with teeth, since these tear up the leaves too much. A tough herb such as rosemary needs more breaking down. Follow the instructions for fruit instead. For instance, try using the wide, flat end of a wooden spoon. That will crush whatever you're muddling, without destroying it.	<b>Step 2: Put the ingredients in a sturdy glass.</b> Do not muddle in a delicate glass that may chip or break when struck. If the cocktail also calls for fruit, cucumber, or other non-herb ingredients, muddle them separately for best results. Granulated sugar tears into the ingredients as you muddle. This can be overkill for soft herbs, so add it with the fruit instead, or dissolve it in a few drops of water and add it to the cocktail separately.	<b>Step 3: Press and twist lightly.</b> Mint, basil, and other soft leaves release bitter flavors when crushed or torn apart. Just press down gently with the blunt tool while you twist your wrist, then release. Do this two or three times. Use your dominant hand to muddle while you hold the glass steady with your other hand. You only need to lightly bruise the leaves to release the oils and aroma.	<b>Step 4: Finish the drink.</b> The leaves are ready when lightly bruised, but still intact. You should be able to smell the herb, since the goal of muddling is to release flavorful, aromatic oils. You can leave the herbs in the final cocktail for presentation or strain them out, as desired.	
				

<b>Task: How to Encourage Your Cat to Eat</b>				
<b>Introduction:</b> Cats are finicky creatures, and they may go off food when they are sick or getting older, for instance. If your cat suddenly stops eating, you should definitely take it to the vet for a checkup, as it may have an underlying problem. Otherwise, you can work on the cat, environment, and food to encourage your cat to eat, as well as use a few tricks when switching foods.				
<b>Topics:</b> Cats, Feeding cats				
<b>Step 1: Feed the cat alone.</b> If there are other animals in the house, the cat may feel uncomfortable eating, as it tries to compete for food. If you section off a room where you can feed just that cat, it can increase the chances of the cat eating.	<b>Step 2: Give smaller meals.</b> Smaller meals throughout the day may seem counter-intuitive, but it can actually encourage your cat to eat. If your cat doesn't have a large appetite, it may be overwhelmed by a large bowl of food.	<b>Step 3: Offer comfort and attention.</b> Some cats want attention when they're eating. Try stroking your cat and talking soothingly to it while it's eating. You can also try a bit of petting if your cat is near the bowl but not eating. This tactic won't work with every cat, though, so if your cat seems disturbed by the attention, leave it alone.	<b>Step 4: Wipe the cat's nose.</b> If the cat has been sick, it may not be smelling very well. Wiping the cat's nose and trying to remove discharge may help it smell better. In turn, it may be more interested in food because it can smell it.	<b>Step 5: Hand feed the cat.</b> A cat who hasn't been interested in food may be more inclined to eat if you hand feed it. For canned food, you can put a little on your finger and offer it to the cat. For dry food, place a bit in your hand, and hold it out for the cat to eat.
				






<b>Task: How to Be a Better Softball Player</b>				
<b>Introduction:</b> Softball is a fun and thrilling sport that requires a lot of hard work and focus. If you feel like you aren't living up to your full potential as a softball player, consider re-evaluating your techniques as you bat, throw, and catch. Additionally, take a moment to think about the demands of your current position, and how you can adjust your practice schedule to hone your skills on the field. With enough focus and dedication, you'll start to see your softball skills improve!				
<b>Topics:</b> Team sports, Softball				
<b>Step 1: Keep your elbow higher so you can throw farther.</b> Pitch the ball with your arm positioned above your shoulder. While the throwing motion will eventually bring your arm to waist or hip level, try to keep your elbow at a higher starting level. When you create the round throwing motion with your arm and elbow, a higher starting point will create more momentum for the ball. Make sure that you're using a fielding glove when you practice, and not a fielding mitt.	<b>Step 2: Thrust your palm forward before letting go with your fingers.</b> Aim to hold the ball in a four-seam grip, where your pointer, middle, and ring finger are positioned behind the edge of the seam you're choosing to hold. When you're holding the ball, don't rest it against your palm. Instead, use palm as a source of momentum as you start to throw, then lift up your fingers quickly to release the ball. One of the most important elements of this type of throw is timing. Practice releasing your fingers at the exact moment your palm hits so that the throw can get more speed.	<b>Step 3: Practice throwing balls heavier than 6 ounces (0.17 kg) to improve your strength.</b> Build muscle in your arms by practicing with softballs that are heavier than what you plan on using in an actual game. Keep your throwing technique consistent, even if your arm feels a bit more tired. When you play with a normal-sized ball, your arm will feel much more adjusted to the throw.	<b>Step 4: Move your arm in a fluid motion to give the pitch momentum.</b> Continue throwing the ball with your usual technique, using a circular motion as your arm swings over your shoulder to waist or hip level. As you swing downwards, don't slow your movement; instead, keep your arm moving swiftly when you release the ball. If you do this, you're more likely to release a more powerful pitch. If you slow down your arm at the end of the throw, you can lose pitch speed.	<b>Step 5: Eye your target before you release the ball.</b> Whether you're pitching or fielding, make firm eye contact with where you want to throw the ball. Don't let your gaze waver; even if your aim is not always consistent, you're much more likely to have success if you throw with a target in mind. As you throw, keep your shoulders and hips angled in the direction of this target. Be sure to square your shoulders in a specific direction whenever you throw the ball. Your shoulders help serve as an invisible guide to which direction the ball goes.
				

Figure 10: Three examples in our curated WIKIPLAN dataset.

sequence (the sequence of texts) contains the amount of information needed to complete the

task.

- Visual-Informativeness: whether the visual



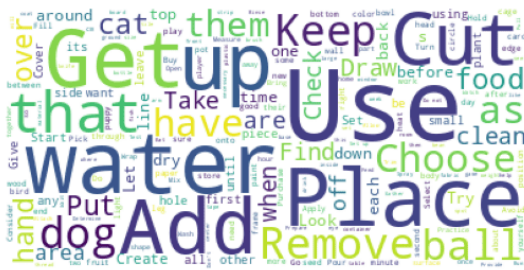


(a) RecipeQA step text



(b) RecipeQA task title text

Figure 11: Word cloud distributions of the task title and step text in the RECIPEPLAN dataset.



(a) Wikihow step text



(b) Wikihow task title text

Figure 12: Word cloud distributions of the task title and step text in the WIKIPLAN dataset.

sequence (the sequence of images) contains the amount of information needed to complete the task.

- Temporal Coherence: whether the multimodal sequence (the paired sequence of texts and images) meets the temporal commonsense requirements, such as a step occurring before another step instead of after.
- Plan Accuracy: whether the multimodal sequence (the paired sequence of texts and images) can successfully complete the task.

To be concrete, the annotators were asked to choose one from the two sequences by 1 - *Sequence 1 is better*, 2 - *Tie*, and 3 - *Sequence 2 is better*. We provide the multimodal plans being annotated as follows:

**Task: How to Get Kids to Eat Healthy.**

**Sequence 1:**

Step 1: Talk to your kids about the importance of eating healthy and make sure that nutritious food is accessible and visible in the house.

Visual Plan at Step 1: [Image]

Step 2: Set rules or guidelines for what is allowed and not allowed in terms of snacks and meals., involve children in grocery shopping and meal preparation as much as you can, lead by example by practicing healthy eating habits yourself.

Visual Plan at Step 2: [Image]

Step 8: Encourage better dietary decisions at mealtimes by sitting down together with them at the table filled with various types of wholesome foods.

Visual Plan at Step 8: [Image]

**Sequence 2:**

Step 1: Talk to your kids about the importance of eating healthy.

Visual Plan at Step 1: [Image]

Step 2: Set rules or guidelines for what is allowed and not allowed in terms of snacks and meals.

Visual Plan at Step 2: [Image]

Step 8: Reward your child for trying new foods or making healthier choices.

Visual Plan at Step 8: [Image]

We show the paired visual plan at each step below the text plan.

Please refer to our Amazon Mechanical Turk human evaluation interface for head-to-head comparison on WIKIPLAN and RECIPEPLAN in Figure 14 and Figure 13 respectively.

## Overview

Thank you for helping us with our research! Please read the below instruction before you go to the question.

### Instruction:

Given the **Task** (e.g, Task: How to muddle), please compare two sequences of steps (**Sequence 1** and **Sequence 2**) and **determine which sequence is better** in terms of four aspects:

- **Textual-Informativeness:** whether the textual sequence (the sequence of texts) contain the amount of information needed to complete the task.
- **Visual-Informativeness:** whether the visual sequence (the sequence of images) contain the amount of information needed to complete the task.
- **Temporal Coherence:** whether the multimodal sequence (the paired sequence of texts and images) meet the temporal commonsense requirements, such as a step occurs before another step instead of after.
- **Plan Accuracy:** whether the multimodal sequence (the paired sequence of texts and images) can successfully complete the task.

## Question

\$(task\_txt)

### Sequence 1:

Task: How to Get Kids to Eat Healthy

Step 1. Talk to your kids about the importance of eating healthy and make sure that nutritious food is accessible and visible in the house.	Step 2. Set rules or guidelines for what is allowed and not allowed in terms of snacks and meals; involve children in grocery shopping and meal preparation as much as you can, lead by example by practicing healthy eating habits yourself.	Step 3. Plan healthy meals together with your kids using their favorite ingredients or recipes incorporating more fruits, vegetables, lean proteins, whole grains gradually without forcing it on them suddenly.	Step 4. Make sure that nutritious food is accessible and visible in the house so they can grab it when they are hungry instead of unhealthy options like chips and candy bars	Step 5. Involve children in grocery shopping and meal preparation as much as you can to make them understand where food comes from	Step 6. Lead by example by practicing healthy eating habits yourself	Step 7. Reward your child for trying new foods or making healthier choices with a variety of fruits & vegetables on a cutting board	Step 8. step 8 Encourage better dietary decisions at mealtimes by sitting down together with them at the table filled with various types of wholesome foods
							

### Sequence 2:

Task: How to Get Kids to Eat Healthy

Step 1. Talk to your kids about the importance of eating healthy.	Step 2. Set rules or guidelines for what is allowed and not allowed in terms of snacks and meals.	Step 3. Plan healthy meals together with your kids using their favorite ingredients or recipes.	Step 4. Incorporate more fruits, vegetables, lean proteins, whole grains into your child's diet gradually without forcing it on them suddenly.	Step 5. Make sure that nutritious food is accessible and visible in the house so they can grab it when they are hungry instead of unhealthy options like chips and candy bars	Step 6. Involve children in grocery shopping and meal preparation as much as you can to make them understand where food comes from	Step 7. Lead by example by practicing healthy eating habits yourself	Step 8. Reward your child for trying new foods or making healthier choices
							

1. **Select an Option:** [Please only consider texts in this option.] Which sequence contains richer Textual-Informativeness? The richer the textual sequence, the more information in the textual sequence (Step 1, Step 2, ...) is for completing the given **Task**.

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

2. **Select an Option:** [Please only consider images in this option.] Which sequence contains richer Visual-Informativeness? The richer the visual sequence, the more information in the visual sequence (the image below each Step) is for completing the given **Task**.

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

3. **Select an Option:** [Please consider both texts and images in this option.] Which sequence has more reasonable Temporal Coherence? In other words, the reasonable multimodal (text and image) sequence obeys commonsense knowledge in real-world that certain steps should happen before others in the given **Task**.

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

4. **Select an Option:** [Please consider both texts and images in this option.] Which sequence can achieve higher Plan Accuracy? In other words, can the multimodal (text and image) sequence more successfully complete the given **Task**?

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

Figure 13: Amazon Mechanical Turk Platform. Questions Layout for Human Raters for Win-Tie-Lose Comparison on WIKIPLAN

## E More Results

### E.1 Showcases

We show more cases in Figure 15- 18 comparing our Text-Image Prompting with powerful baselines Text-Davinci-003 + Stable-Diffusion.

### E.2 Word Cloud

In comparison with the word cloud distribution of the ground truth, we also show the word cloud of the baselines and TIP on WIKIPLAN and RECIPEPLAN.

### E.3 Failure Cases

In Figure 19, we showcase failure generation. For example, the state of the almond stays unchanged in Figure 19a, we suppose this is due to no explicit awareness of previous state change. In Figure 19b, at step 2, the generated image plan, though complemented with the text plan, loses authenticity in that the clock should not appear in a pan with carrots.

### E.4 Limitations

Relying on the LLMs to reason over complex text for text-to-image models though improving the quality, still remains a large gap with human perfor-

## Overview

Thank you for helping us with our research! Please read the below instruction before you go to the question.

### Instruction:

Given the **Task** (e.g, Task: How to muddle), please compare two sequences of steps (**Sequence 1** and **Sequence 2**) and **determine which sequence is better** in terms of four aspects:

- **Textual-Informativeness:** whether the textual sequence (the sequence of texts) contain the amount of information needed to complete the task.
- **Visual-Informativeness:** whether the visual sequence (the sequence of images) contain the amount of information needed to complete the task.
- **Temporal Coherence:** whether the multimodal sequence (the paired sequence of texts and images) meet the temporal commonsense requirements, such as a step occurs before another step instead of after.
- **Plan Accuracy:** whether the multimodal sequence (the paired sequence of texts and images) can successfully complete the task.

## Question

\$(task\_txt)

Sequence 1:

Task: How to make cheese explosion caps.

Step 1. Preheat the oven to 350°F and prepare a baking sheet lined with parchment paper or foil sprayed lightly with non-stick cooking spray.	Step 2. Cut off the tops of mini sweet peppers, removing seeds from inside each pepper cap.	Step 3. Fill each pepper cap with approximately 1 tablespoon of cheese.	Step 4. Place filled mini peppers on the prepared baking sheet, making sure they are not touching one another.	Step 5. Bake in preheated oven until cheese is melted and golden brown, approximately 10 minutes or longer depending on your desired level of doneness/crispyness of the outside layer of cheese around the pepper caps (optional: sprinkle parmesan cheese over top near end of cooking).	Step 6. Remove from oven and let cool slightly before serving warm
					

Sequence 2:

Task: How to make cheese explosion caps.

Step 1. Preheat oven to 350°F.	Step 2. Cut the caps of mini sweet peppers and remove seeds from inside.	Step 3. Fill each pepper with cheese, about 1 tablespoon for each pepper cap.	Step 4. Place filled mini peppers on a baking sheet lined with parchment paper or foil sprayed lightly with non-stick cooking spray, making sure they are not touching one another.	Step 5. Bake in preheated oven until cheese is melted and golden brown, approximately 10 minutes or longer depending on your desired level of doneness/crispyness of the outside layer of cheese around the pepper caps (optional: sprinkle parmesan cheese over top near end of cooking).	Step 6. Remove from oven and let cool slightly before serving warm
					

1. **Select an Option:** [Please only consider texts in this option.] Which sequence contains richer Textual-Informativeness? The richer the textual sequence, the more information in the textual sequence (Step 1, Step 2, ...) is for completing the given **Task**.

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

2. **Select an Option:** [Please only consider images in this option.] Which sequence contains richer Visual-Informativeness? The richer the visual sequence, the more information in the visual sequence (the image below each Step) is for completing the given **Task**.

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

3. **Select an Option:** [Please consider both texts and images in this option.] Which sequence has more reasonable Temporal Coherence? In other words, the reasonable multimodal (text and image) sequence obeys commonsense knowledge in real-world that certain steps should happen before others in the given **Task**.

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

4. **Select an Option:** [Please consider both texts and images in this option.] Which sequence can achieve higher Plan Accuracy? In other words, can the multimodal (text and image) sequence more successfully complete the given **Task**?

- 1 - Sequence 1 is better  2 - Tie  3 - Sequence 2 is better

Figure 14: Amazon Mechanical Turk Platform. Questions Layout for Human Raters for Win-Tie-Lose Comparison on RECIPEPLAN.

mance. This is mainly restricted by the pre-training gap between LLMs and text-to-image models. To solve this, further work should explore the finetuning stage that how to inject this language reasoning into the multimodal generation models.

In addition to the model-side limitations, the dataset is limited in that not all the possible multimodal plans are provided and their quality is hard

to validate. Due to the lack of perfect metrics in evaluating the text-image sequences, the research in multimodal procedural planning maybe difficult to scale up. Future work should explore this promising direction and furthermore lead LLMs and T2I models better multimodal procedural planners.



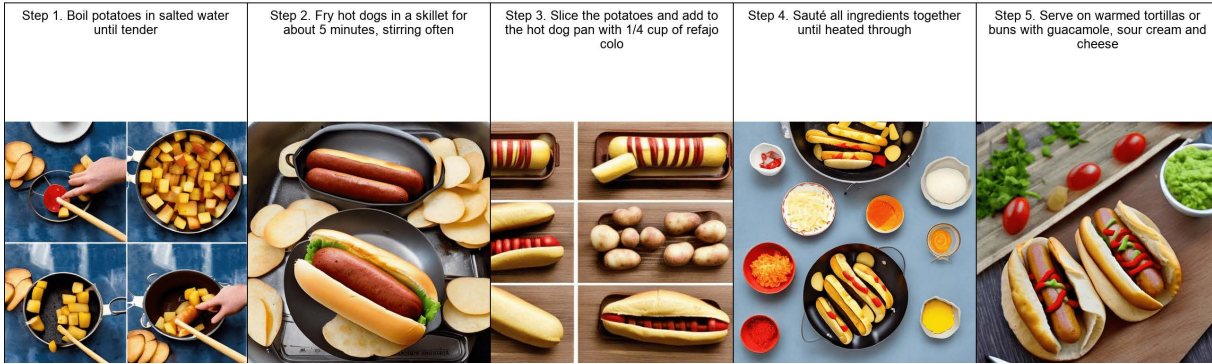
(a) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)



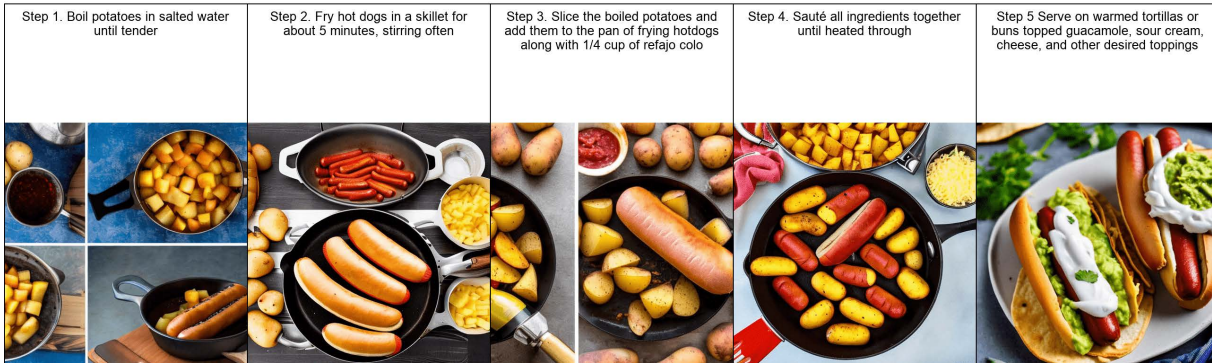
(b) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)

Figure 15: More showcases of plan comparisons on RECIPEPLAN.

Task: How to make hot dogs and potatoes slachipapas with refajo colo.



Task: How to make hot dogs and potatoes slachipapas with refajo colo.



(a) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)

Task: How to make peppermint patty pudding shot.



Task: How to make peppermint patty pudding shot.



(b) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)

Figure 16: More showcases of plan comparisons on RECIPEPLAN.  
10951

Task: How to Use Oregano Oil

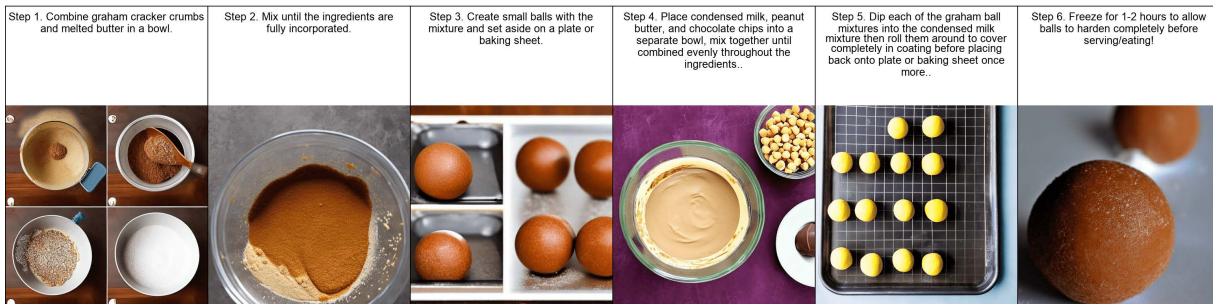


Task: How to Use Oregano Oil



(a) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)

Task: How to Make Graham Balls



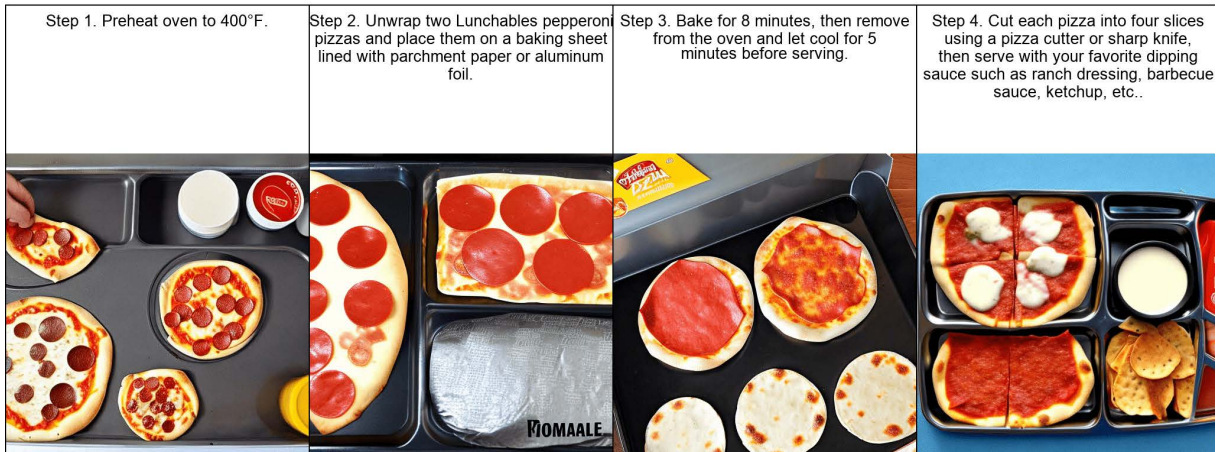
Task: How to Make Graham Balls



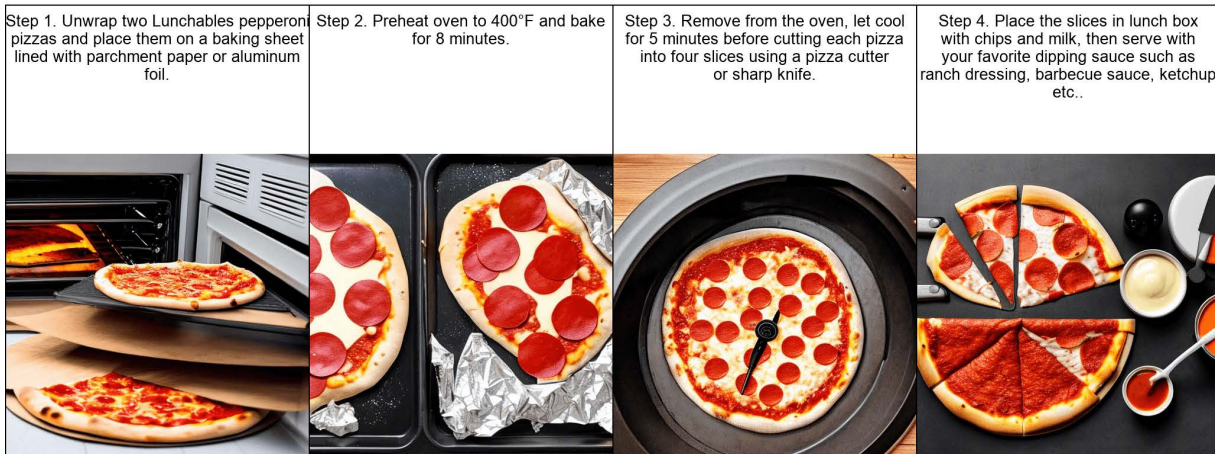
(b) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)

Figure 17: More showcases of plan comparisons on WIKIPLAN.

Task: How to Make Pizza Lunchables

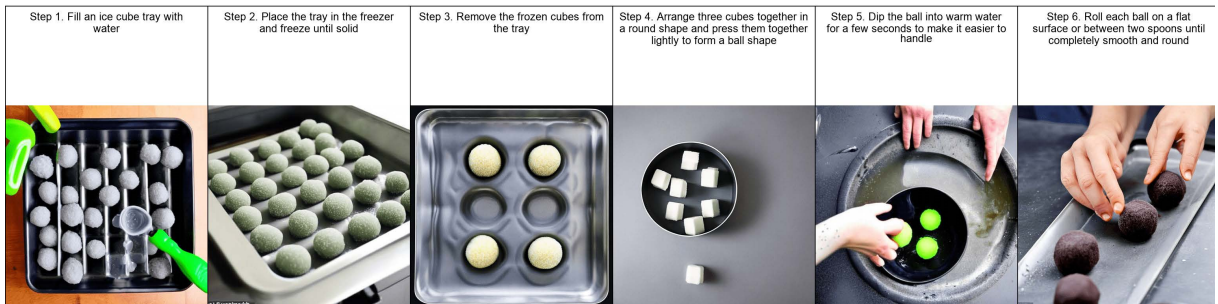


Task: How to Make Pizza Lunchables



(a) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)

Task: How to Make Ice Balls



Task: How to Make Ice Balls



(b) Text-Davinci-003 + Stable-Diffusion (Top) vs. TIP (Bottom)






Figure 18: More showcases of plan comparisons on WIKIPLAN.

Task: How to make homemade raw almond butter.

Step 1. Place almonds in a bowl and soak overnight	Step 2. Preheat oven to 350°F and spread soaked almonds on a baking sheet	Step 3. Roast for 10-15 minutes, stirring occasionally until lightly browned	Step 4. Remove from oven and let cool completely	Step 5. Place cooled roasted almonds in food processor or high-speed blender; blend until creamy, stopping to scrape down sides as needed	Step 6. Add salt (optional) and continue blending until desired consistency is reached
					

(a) Failure multimodal plans generated by our Text-Image Prompting (TIP).

Task: How to make carrot and swede potch mash.

Step 1. Peel and chop carrots and swedes into small cubes	Step 2. Boil the vegetables in a saucepan with boiling salted water for 15-20 minutes until tender	Step 3. Place the cut up carrots into a wooden bowl before adding them to the pan of boiling water along with the swede cubes	Step 4. Drain off excess liquid from the cooked vegetables	Step 5. Mash the cooked vegetables with butter, salt and pepper to taste
				

(b) Failure multimodal plans generated by our Text-Image Prompting (TIP).

Figure 19: We showcase failure cases of our Text-Image Prompting on generating multimodal plans on both datasets.