Generate Subgoal Images before Act: Unlocking the Chain-of-Thought Reasoning in Diffusion Model for Robot Manipulation with Multimodal Prompts

Fei Ni[1] Jianye Hao[1*] Shiguang Wu[2] Longxin Kou[1] Jiashun Liu[1] Yan Zheng[1]
Bin Wang [2] Yuzheng Zhuang[2]
[1]Tianjin University, China [2]Huawei Noah's Ark Lab, China

Abstract

Robotics agents often struggle to understand and follow the multi-modal prompts in complex manipulation scenes which are challenging to be sufficiently and accurately described by text alone. Moreover, for long-horizon manipulation tasks, the deviation from general instruction tends to accumulate if lack of intermediate guidance from high-level subgoals. For this, we consider can we ***generate subgoal images before act*** *to enhance the instruction following in long-horizon manipulation with multi-modal prompts? Inspired by the great success of diffusion model in image generation tasks, we propose a novel hierarchical framework named as CoTDiffusion that incorporates diffusion model as a high-level planner to convert the general and multi-modal prompts into coherent visual subgoal plans, which further guide the low-level policy model before action execution. We design a semantic alignment module that can anchor the progress of generated keyframes along a coherent generation chain, unlocking the chain-of-thought reasoning ability of diffusion model. Additionally, we propose bi-directional generation and frame concat mechanism to further enhance the fidelity of generated subgoal images and the accuracy of instruction following. The experiments cover various robotics manipulation scenarios including visual reasoning, visual rearrange, and visual constraints. CoTDiffusion achieves outstanding performance gain compared to the baselines without explicit subgoal generation, which proves that a subgoal image is worth a thousand words of instruction. The details and visualizations are available at* https://cotdiffusion.github.io.

1. Introduction

Embodied manipulation focuses on creating generalists that can perceive, reason, and act within complex environments, which sits at the intersection of robotics control, computer vision, and natural language processing [11]. Recent advances in vision-language model (VLM) such as CLIP [38], BLIP [22], enable open-vocabulary visual recognition and show promising potential of robotics closed-loop control by endowing them with the ability to predict robot actions in embodied robotics [5, 10, 18, 51]. However, robotics agents still face significant challenges in following instructions for long-horizon manipulation tasks, especially when the given general instructions are not progressive step-wise prompts, but implicitly contain several subtasks to accomplish. Moreover, complex manipulation scenarios with rich visual contexts are often challenging to be sufficiently and accurately described through text-only prompts, requiring multi-modal prompts to convey the instructions and intentions accurately [21], further increasing the difficulty of instruction following and task completion.

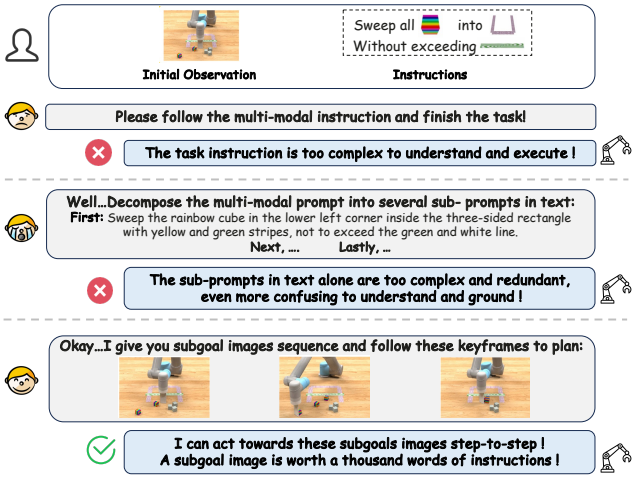

Figure 1. A motivation example of robotics manipulation tasks in multi-modal instructions. The subgoal images are worth a thousand words, inspiring us to propose a novel framework CoTDiffusion to generate goal images step-by-step before act.

For example in Fig. 1, now given a short multi-modal prompt, the robot arm needs to follow the instructions and accomplish the task. The multi-modal instruction is brief but implicitly encapsulates long-horizon multi-step manipulation steps. We have two pathways to solve this task according to the recent embodied manipulation works. The

*Corresponding author: jianye.hao@tju.edu.cn

first paradigm is directly leveraging the powerful VLM to encode prompts and observations into tokens and predict the entire action sequence with the subsequent decoder in an end-to-end manner, like RT-2 [51] or VIMA [21]. However, the compounding small errors over long horizons will lead to catastrophic deviations from the original task instructions due to the lack of intermediate guidance from coherent subgoals. The second paradigm is to incorporate the large language models (LLMs) to explicitly decompose the general prompts into several orderly subgoals with step-wise text instructions, like SayCan [1]. Though LLMs like ChatGPT [37] struggle to support the multi-modal prompts beyond text-only modality, we can seek help from the latest multi-modal LLMs such as LLaVA [25], EmbodiedGPT [33]. Even if we idealistically assume (M)LLM-based decomposition produces well-grounded and rational plans, generated text-only prompts tend to be extremely redundant and complex to accurately specify each subgoal in complex manipulation scene, may even be more confusing than the original general prompt and impose a heavy burden on the robot agents to re-parse and re-understand again.

Considering the phrase 'a picture is worth a thousand words', subgoal images could provide higher expressive capabilities for conveying subtasks compared to sub-prompts with complex text. This motivates us to *generate subgoal images step-by-step before act* to enhance long-horizon instruction following. Fortunately, tremendous success in text-to-image generation sheds light on the possibility of directly translating multi-modal prompts into coherent grounded visual subgoals step by step. Our key insight is to propose a hierarchical framework CoTDiffusion that integrates diffusion model as the high-level visual planner that can understand multi-modal prompts and progressively generate chained subgoal images based on the current visual observations in a chain-of-thought manner. The chained subgoal images serve as visual milestones to anchor the execution of task, further fed into a low-level foundation model for the specific step-wise action planning to achieve the provided goal scene. The subgoal images act as a unified interface bridging the high-level visual planning and low-level action planning, decoupling the instruction understanding and action execution. The foundation module allows the high-level diffusion model to focus solely on instruction understanding and subgoal visualization without overwhelming joint action prediction training. Conversely, the high-level visual planner frees the low-level foundation model from ambiguous prompt understanding and long-horizon planning. By providing coherent subgoal images as visual landmarks, the requirements for the foundation model are reduced to basic single-object manipulation primitives.

The key challenge to enabling CoTDiffusion to progressively generate subgoal images in a chain-of-thought manner lies in tracking the generated subgoal's progress on task prompts. In other words, for coherent subgoal generation, diffusion model needs to know 'which step this subgoal image has reached' and 'which step needs to reach in the next subgoal image' based on task prompts. For this, we design a triple alignment module comparing the cross-frame visual contrasts between the generated subgoal image and initial observation, and align the progress back to prompts to anchor the stage of generation chains. Specifically, we use a coarse-to-fine training pipeline. First, the aligned module is coarsely pre-trained to predict residual mask patches between subgoal images for aligning spatial semantics, focusing on salient differences rather than pixel details in textures or colors. The semantic alignment module is then integrated into the diffusion model for step-wise image generation with fine-grained pixel reconstruction. Additionally, bi-directional generation and frame concatenation mechanism further enhance subgoal image fidelity and instruction following. The contributions of this work are as follows:

- We propose a hierarchical framework CoTDiffusion that the high-level diffusion model translates the multi-modal prompts into coherent subgoal images in a chain-of-thought manner as visual milestones to anchor the low-level foundation model executing, enhance the instruction following on long-horizon manipulation tasks.
- We design a semantic alignment module to capture the semantic relevance between visual subgoals and prompt, progressively tracking the progress of generated images along the coherent generation chain to unlock the chain-of-thought reasoning capability for multi-modal prompt.
- The experiments on various long-horizon manipulation tasks empirically demonstrate that CoTDiffusion enjoys outstanding performance gain than prior methods by explicitly generating goal images step by step before act.

2. Related Work

2.1. Diffusion Models for Text-to-Image Generation

Recently, diffusion models have emerged as a powerful paradigm for high-fidelity text-to-image synthesis [9, 17, 35, 42, 45]. Models such as DALL-E 2 [39], Imagen [44] and Stable Diffusion [43] have demonstrated impressive success in generating realistic images from textual descriptions. The text instructions are tokenized as specific conditions and injected into the denoising process for controllable image generation. Moreover, some recent works have begun to explore image generation with multi-modal prompts, incorporating diverse inputs beyond just text. UniDiffuser [3] treats both image and text as sequential token streams for diffusion generations and Versatile diffusion [49] employs a multi-flow design to tackle multi-modal prompts. However, directly applying existing text-to-image approaches to complex robotics manipulation poses challenges in instruction understanding and following. Prompts

in manipulation tasks are often too high-level and general to effectively ground and reason, making the generated images struggle to accurately follow instructions and ground to the given scenarios, not sufficient to serve as goal images to offer fine-grained guidance for manipulation. Our work employs a semantic alignment module for multi-stage cross-modal alignment to achieve better grounding of generated images and unlock the chain-of-thought reasoning abilities of diffusion model for long-horizon tasks.

2.2. Robotics Manipulation and Control

With the development of VLMs for robotics, such as R3M [34], VIP [29], and LIV [30] pre-trained general visual representations for robotic perception, RT-2 [51], Gato [41] and PaLM-E [10] can convert instruction and observation into tokens and apply large transformers to predict action tokens in an end-to-end manner for robot manipulation tasks. While these works mostly focus on language-conditioned manipulation, more complex manipulation scenarios with multi-modal prompts introduced by VIMA [21] are also gaining increasing attention. Accurately instruction following becomes more challenging with general multi-modal prompts, especially in long-horizon manipulation tasks. It is not trivial to directly borrow the recent works like SayCan [1] and EmbodiedGPT [33] to utilize MLLMs to decompose multi-modal instructions into step-wise text prompts. Even if we assume step-wise instructions can be rationally decomposed, they often require complex and redundant language to accurately specify subtasks and may be more confusing to understand. Inspired by the motto 'show, don't tell', our work aims to translate the general multi-modal prompts into coherent subgoal images instead of text step by step, serving as visual landmarks to enhance the instruction following in long horizon.

2.3. Chain-of-Thought Reasoning

Chain-of-Thought [48] reasoning refers to the general strategy of solving multi-step problems by decomposing them into a sequence of intermediate steps, showing great success in guiding the model to think step-by-step for enhancing instructing following and reasoning accuracy. It has recently been applied extensively in a variety of problems such as mathematical reasoning [8, 24], program execution [36, 40], commonsense or general reasoning [23, 48]. However, chain-of-thought visual reasoning has not been well explored for robotics manipulation. Inspired by prior explorations into CoT reasoning, we introduce the novel concept of chaining conditional image generations step by step to guide long-horizon manipulation. Our work employs a semantic alignment module to anchor the generated subgoal image in the entire reasoning chain, unlocking the chain-of-thought reasoning capabilities for the diffusion model.

3. Method

We propose CoTDiffusion, a hierarchical framework that integrates the diffusion model as the high-level module to decompose multi-modal prompts in a chain-of-thought manner and progressively generate chained subgoals images step by step to guide the underlying foundation model for long-horizon tasks. Specifically, we first discuss the overview pipeline of CoTDiffusion in Sec. 3.1, and then introduce the pretraining of the coarse semantic alignment to capture the step-wise spatial information in Sec. 3.2. Then we discuss how to further fine-grain the CoT reasoning and grounded generation capabilities of the diffusion model in Sec. 3.3 and introduce the details of goal image conditioned foundation model for action planning in Sec. 3.4.

3.1. Pipeline Overview

The overall pipeline of CoTDiffusion is presented in Fig. 2, decoupling the visual planning and action planning. Given the initial observation x_0 and a multi-modal prompt $\mathcal{P}$ as task instruction potentially needs to be reached by N subgoal steps, robots are required to learn a policy conditioned on the prompt and accomplish the specified long-horizon task. Our model is capable of understanding these general instructions and leverages diffusion model implicitly to decompose them into subgoal images chain $\tau_x = \{x_i\}_{i=1}^N$. The precisely generated goal images x_i then guide the foundation model's planning for inferring the action sequences $\tau_a^i = \{a_{i,t}\}_{t=1}^T$ with the subgoal horizon length T. Let p_Θ model this hierarchical decision-making framework. With bi-level modules, p_Θ can be factorized into visual planning p_ϕ, and action planning p_ψ. Under the Markovian assumption, the overall framework can be formulated as:

$$p_\Theta(\{\tau_a^i\}_{i=1}^N|\mathcal{P}, x_0) = \underbrace{\left(\prod_{i=1}^N p_\phi(x_i|\mathcal{P}, x_0)\right)}_{\text{visual planning}} \underbrace{\left(\prod_{i=1}^N \prod_{t=1}^T p_\psi(a_{i,t}|x_i)\right)}_{\text{action planning}} \tag{1}$$

The visual planning module consists of a multi-modal encoder to understand the complex task instructions and conditioned diffusion model which implicitly decomposes the task into progressive chained subgoals images. The standardized subgoal images serve as a unified interface bridging the visual planning and action planning modules. With the decoupled foundation model for action prediction, high-level visual planning focuses solely on comprehending instructions and visualizing subgoals without confusion from joint action prediction training. Conversely, the visual planning module frees the action planning module from ambiguous prompt parsing and long-horizon planning by providing coherent keyframes as visual landmarks, further reducing the capabilities requirements of the foundation model to basic single object manipulation skills.

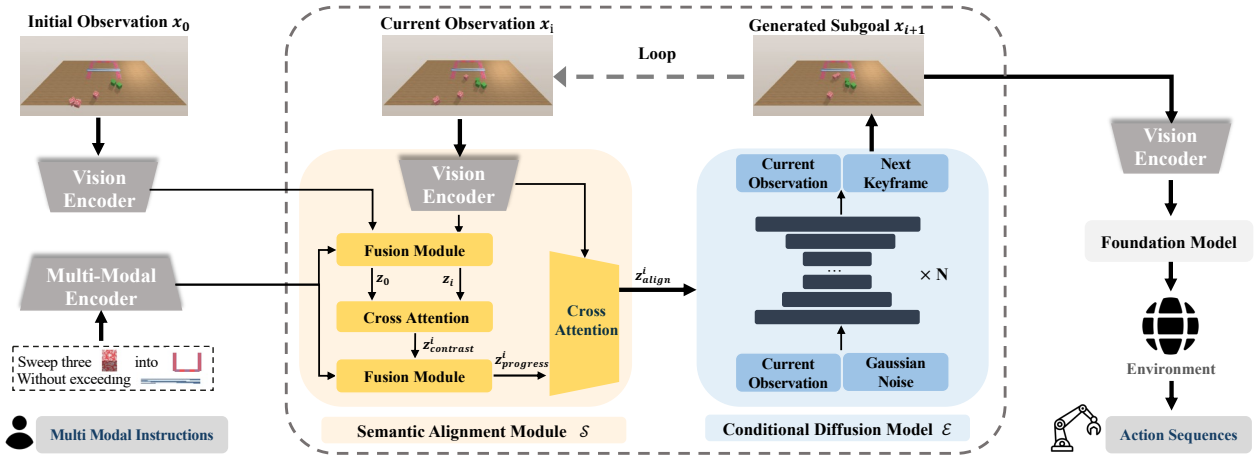

Figure 2. Method overview: CoTDiffusion consists of a multi-modal encoder and vision encoder $\mathcal{V}$, semantic alignment module $\mathcal{S}$, conditional diffusion model $\mathcal{E}$ and foundation model $\mathcal{F}$ for action planning. The prompt and observation tokens are combined and fed into the semantic alignment module to identify the current reasoning chain step, providing progressive guidance for the diffusion model to generate the next subgoal image. The generated keyframes are further fed to the foundation model which predicts action sequences to achieve the imagined goal scene and this recursive process repeats in a receding horizon control loop until the task is finished.

3.2. Pre-training Coarse Semantic Alignment

For the challenge of explicit decomposing step-wise sub-prompts from multi-modal prompts, the diffusion model struggles to generate sequentially coherent keyframes that incrementally advance the prompt instructions and execution progress. The diffusion model fails to identify the stage of current generated keyframes onto the whole chain-of-thought progress and the generated sub-goals images may exhibit repetitions, skipping, or even backtracking to visited subgoals. To alleviate this dilemma, we design a semantic alignment module to empower diffusion model to track the progress of generated keyframes into the entire chain.

Triple Alignment Architecture Since explicit decomposition for the multi-modal prompt is intractable, the initial observation x_0 and prompt $\mathcal{P}$ remain fixed across the progressive generation of diffusion model. Thus, tracking the progress critically relies on extracting semantic information from the generated subgoal x_i. To this end, we design a triple alignment module denoted as $\mathcal{S}(x_0, x_i, \mathcal{P})$ to compare current generated subgoal x_i against initial observation x_0 and prompts $\mathcal{P}$, serving as an information bottleneck to capture the key semantic cues about the subgoal-prompt matching to locate the current chain stage and deliver to the diffusion model for next subgoal generation. Specifically, we first capture visual representation from both the initial and current observation using a shared vision encoder $\mathcal{V}$ and concatenate with prompt tokens $\mathcal{P}$. Then they are refined through fusion module which consists of several self-attention blocks separately to obtain attention tokens z_0 and z_i aligned to the prompts. The cross-attention between z_0 and z_i aggregates a contrast representation z^i_{contrast} and then concatenated with prompt tokens into another fusion module to capture the progress tokens z^i_{progress}, including rich context about the advancement and completion of the prompted task between the two subgoals. Finally, z^i_{progress} and the current visual tokens x_i and then fed into another cross-attention pass, which visual tokens x_i serve as keys and progress tokens z^i_{progress} as queries to infer the final aligned tokens z^i_{align}. The aligned token encapsulates cues about 'which subgoal has reached now' and 'which subgoal should reach next', further injected into the diffusion model as precise semantic guidance to steer the generation of next subgoal. Through multi-stage cross-frame attention mechanisms, the triple alignment module can capture the cross-modal semantic correlation and achieve dynamic grounding of generated subgoal image into the multi-modal prompts to identify progress along the CoT reasoning chain.

Masked Patch Prediction It is not trivial to directly integrate the designed triple alignment module without pre-training into the diffusion model for joint training from scratch. Simultaneously optimizing semantic alignment and pixel-level reconstruction such as texture, color, shape, and spatial transformations overwhelms joint learning. Thus, we propose a two-stage coarse-to-fine approach decoupling semantic alignment pretraining from diffusion model fine-tuning, illustrated in Fig. 3. First, we pre-train coarse semantic alignment focused only on spatial correlations, without pixel generation. Considering the adjacent keyframe contrasts reveal object manipulations as mask residuals patch $\hat{m}_i = x_{i+1} - x_i$, we can utilize these cross-frame spatial semantics to provide coarse alignment supervision. Regions with no visual changes likely correspond to stationary background elements or objects not involved in the current subtask operation. Contrastively, masked areas indicate spatial layout modifications concluding the rich spatial semantic information between keyframes. The sequence

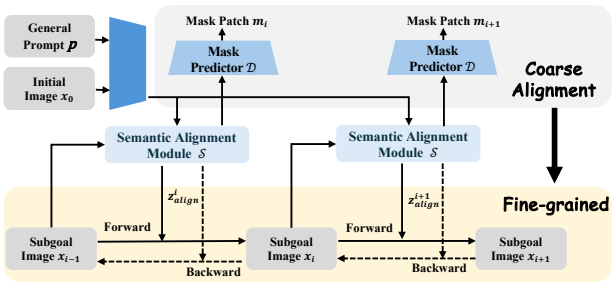

Figure 3. Two phase of coarse-to-fine alignment pipeline.

of mask residuals patch $\{\hat{m}_i\}_{i=1}^{N-1}$ can be viewed as implicitly decomposing the prompt into spatial manipulation steps across visual frames that can be used for alignment. So we fed the aligned tokens into a mask predictor $\mathcal{D}$ to decode the mask region and set modeling mask prediction as a pretext task. Mask prediction encourages the model to infer these salient spatial semantics and changes to incrementally track prompt completion and focus regions. The process can be formulated as: $\{m_i\}_{i=1}^{N} = \sum_{i=1}^{N} \mathcal{D}(m_i|z^i_{\text{align}})\mathcal{S}(z^i_{\text{align}}|x_0, x_i, \mathcal{P})$.

By factorizing the alignment and generation objectives, the alignment module can develop basic visual reasoning skills through patch-level mask prediction before integrated into the diffusion model for controllable image synthesis.

3.3. Fine-grained Diffusion Training

In the second stage, we integrate the pretrained coarse alignment module into the diffusion model training for fine-grained image generation. To further enhance instruction following, we design a bi-directional generation that reconstructs the current frame from the aligned token, providing an additional learning constraint. We also employ a frame concatenation mechanism to enhance the consistency across generated keyframes and the fidelity of generation.

Bi-directional Aligned Generation Building upon coarse pretraining, we integrate the alignment module into the diffusion model for fine-grained training with the pixel-level supervision of ground truth keyframes, illustrated in Fig. 3. Here we propose bi-directional aligned generation, where the aligned token z^i_{align} not only guides forward prediction but also reconstructs the current frame through backward passing. In the forward pass, based on current subgoal x_i, z^i_{align} provides semantic conditioning to generate the next subgoal image x_{i+1}. In the backward pass, the same aligned token can guide diffusion model to reconstruct the current subgoal x_i from the next subgoal image x_{i+1}, acting as an additional constraint. The training objective can be formulated as:

$$L = \mathbb{E}_{x_i \in D}[\underbrace{\|\hat{x}_i - \mathcal{E}\left(x_{i-1}, z^i_{\text{align}}, \mathcal{P}\right)\|}_{\text{Forward Generation}} + \underbrace{\|\hat{x}_{i-1} - \mathcal{E}\left(x_i, z^i_{\text{align}}, \mathcal{P}\right)\|}_{\text{Backward Generation}}] \tag{2}$$

Bi-directional generation compels tighter instruction grounding, as accuracy is required for both progressive and reverse keyframe prediction. This dual-direction linkage enhances visual coherence and sequence controllability compared to forward-only training, improving the chain-of-thought capabilities of diffusion model.

Frame Concat Mechanism Generating consistent and coherent subgoal images requires considering both semantic guidance to match prompts and visual grounding into observations. The aligned tokens z^i_{align} and image tokens x_i are both fused into the denoising process in a classifier-free manner, as shown in Eq. (2). However, merely injecting current observation tokens as condition into noise model struggles to provide sufficient guidance to prevent distortions conflicting with the current observation, which is critical for precisely grounded robotics manipulations. Inspired by VDT[28], we employ frame concat mechanism to directly take current observation frame as the component of input rather than only condition for the next subgoal image generation, illustrated in Fig. 2. The current observation frame concatenated with the standard Gaussian noise gives a strong visual continuity prior, then fed into diffusion model to denoise altogether. For the training, we split the corresponding frame from the denoising frame and supervised it to the ground truth subgoal image. The rich visual context from the concated current frame enables coherent denoising in diffusion model further guarantees consistency with grounded observation and enforces coherence and smoothness across chain-of-thought visual planning.

3.4. Goal-conditioned Policy Model

The final component in our framework is the low-level policy model for action planning, generating an action trajectory τ^i_a when given observation trajectory τ^i_x from visual planning. The policy model can be parameterized as an image-conditioned planner that infers the action $a_{i,t}$ given the current observation $x_{i,t}$ and the generated subgoal image g_i: $\tau^i_a = \{a_{i,t}\}_{t=1}^{T} \sim \prod_{t=1}^{T} p_\psi(a_{i,t}|x_{i,t}, g_i)$. The policy model is trained by imitation learning in an end-to-end manner, using paired subgoal images and expert trajectories. Thanks to the explicit subgoal generation from high-level visual planner, the low-level policy model does not need to master complex multi-step manipulation skills in long-horizon. Therefore, we simplify the policy model architecture and implement it as a simple MLP to only possess basic single-object manipulation primitives, which greatly reduces the burden of policy model training. Although we do not seek improvement through a more powerful policy model in this paper, CoTDiffusion is flexible to incorporate any goal-conditioned design like transformers or diffusion policies. For more details of low-level policy model, please refer to Appendix E.3.

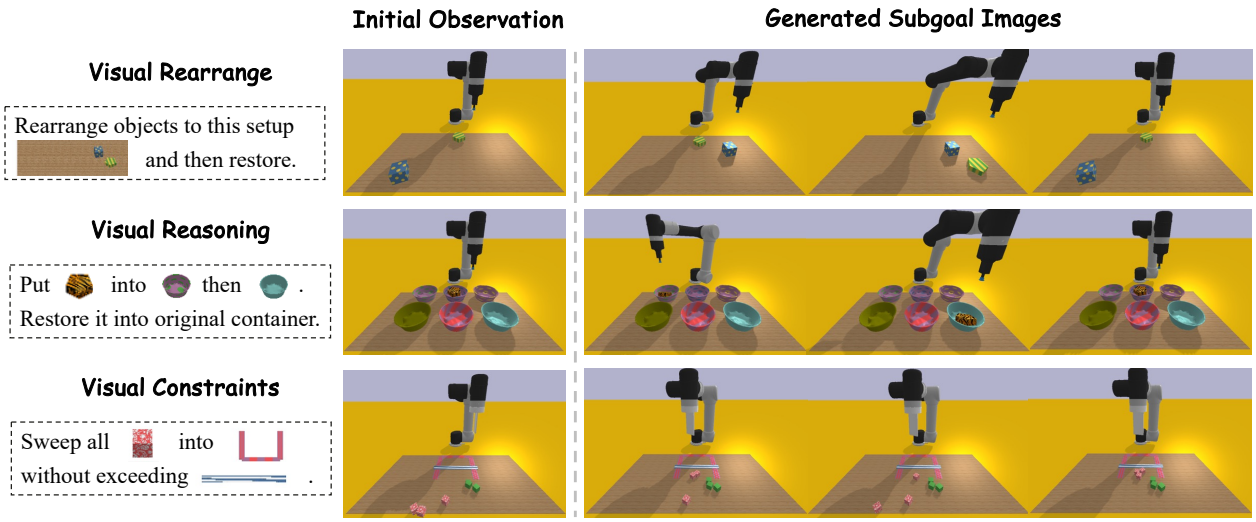

Figure 4. Visualization of CoTDiffusion in three typical long-horizon tasks with multi-modal prompts in VIMA-BENCH.

4. Experiments

4.1. Experiment Setup

Benchmark & Tasks We conduct evaluation on VIMA-BENCH, a benchmark suite for multimodal robot learning, which is built on the Ravens robot simulator [50]. VIMA-BENCH supports extensible collections of objects and textures to compose multi-modal prompts and to procedurally generate a large number of tasks. VIMA-BENCH can generate large quantities of expert trajectories via scripted oracle agents, and the ground truth keyframes that can leveraged for the training of CoTDiffusion. VIMA-BENCH contains various tasks ranging from simple object manipulation to multi-object manipulation and we select three representative long-horizon manipulation tasks in VIMA-BENCH- *visual rearrangement*, *visual constraints*, and *visual reasoning*, which represent different levels of complexity and requirements for embodied manipulation. These tasks represent different levels of complexity and requirements for embodied manipulation and can test the ability to understand spatial relationships and perform precise object manipulations. For more details regarding the simulation benchmark and tasks setting, please refer to Appendix A .

Evaluation Metric The evaluation metrics we use in the experiments cover three aspects as following:

• *Image Fidelity* - measures the quality and realism of the generated keyframe images. We choose Fréchet Inception Distance (FID) [15] as a useful metric.

• *Instruction Following* - evaluates how well the step-by-step keyframes cover the complex instructions. We use CLIP-based text-image similarity scores [14] between the prompt and keyframes to quantify instruction alignment.

• *Task Completion* - tests the end-to-end utility by executing the embodied manipulation task. The final success rate on long-horizon tasks can be a fair metric.

4.2. Baselines

There are several existing methods for constructing robot manipulation policies conditioned on complex prompts, which we use as baselines in our experiments:

- **Gato**[41], a generalist agent that can solve tasks from multiple domains where tasks are specified by prompting the model with the observation and action subsequence. For fair comparisons, we provide the same conditioning manner with multi-modal prompts.
- **Flamingo**[2], a vision-language model that learns to generate textual completion in response to multimodal prompts. We borrow the architecture in [21] which adapts it to support decision-masking by replacing the output layer with robot action heads.
- **VIMA**[21], the first transformer-based generalist robot agent that processes manipulation tasks with multimodal prompts. VIMA adpots an object-centric approach to flatten all the observation and prompts into object tokens sequence and predicts motor actions autoregressively and demonstrates SOTA performance on VIMA-BENCH.
- **SuSIE**[4], the concurrent work for visual planning in manipulation tasks on the CALVIN [31], incorporating pretrained image-editing model and the low-level goal-conditioned policy. For fair comparison, we finetune the model with the same oracle data with CoTDiffusion.

4.3. Quantitative Results of Success Rate

We begin by comparing the performance of CoTDiffusion and baselines to solve long-horizon tasks in visual rearrange, visual reasoning, and visual constraints. The performance comparisons shown in Tab. 1 demonstrate CoTDiffusion significantly outperforms other baselines in success rate. The baselines can be divided into two kinds of planners, including *abstract planner* and *visual planner*. *Abstract planner* like Gato, Flamingo and VIMA directly

map general prompts to subsequent actions in an end-to-end manner. Gato and Flamingo gets low success rates on long-horizon tasks without explicit subgoal generation to correct the accumulative deviation errors from the instructions. In contrast, *Visual planner* like SuSIE and CoTDiffusion can generate intermediate goal images to guide the action planning, which can enhance instruction following for long-horizon tasks via visual planning. However, SuSIE lacks intrinsic chain-of-thought reasoning capabilities, struggling to generate logically progressing subgoal sequences as coherent plans from general prompts. To enable comparison, we grant SuSIE the manually decomposed prompts into privileged step-wise sub-prompts which are unavailable in fair settings, denoted as 'SuSIE + sub-prompts'. Nonetheless, its performance still lags CoTDiffusion, which can implicitly align generated images with multi-modal prompts through learned correspondence between visual signals and language semantics. One more potential reason to restrict the performance of SuSIE is that provided sub-prompts are challenging to perfectly encapsulate complex instructions. In contrast, CoTDiffusion develops intrinsic chain-of-thought reasoning and alignment for generated subgoal images for flexible visual planning directly from the raw multi-modal prompts, without the need for pre-decomposed sub-prompts. The comparisons highlight the importance of chain-of-thought capabilities in visual planning, especially the role of implicit semantic alignment for instruction following long-horizon manipulation tasks.

Methodology	Rearrange	Reasoning	Constraints	Overall
Gato	6.4 ± 1.3	2.5 ± 0.4	25.2 ± 3.1	11.4 ± 1.6
Flamingo	17.5 ± 1.6	3.0 ± 0.5	36.1 ± 4.2	18.9 ± 2.1
VIMA	43.1 ± 3.3	38.2 ± 4.4	67.2 ± 5.2	49.5 ± 4.3
SuSIE	2.7 ± 0.3	3.1 ± 0.6	24.3 ± 5.9	10.0 ± 2.3
+sub-prompts	37.7 ± 6.2	39.0 ± 4.5	52.3 ± 7.0	43.0 ± 5.9
CoTDiffusion	$\mathbf{59.0 \pm 1.7}$	$\mathbf{51.7 \pm 2.6}$	$\mathbf{83.1 \pm 4.7}$	$\mathbf{64.6 \pm 3.0}$

Table 1. The evaluations of success rates on three typical long-horizon tasks with multi-modal prompts.

4.4. Further Analysis

Robustness to Insufficient Perception Rich visual observations from diverse views are crucial for complex robot manipulation tasks. Restricted perspectives limit rich representations used for action planning, degrading performance on long-horizon multi-object manipulation. We evaluate different methods in single-view including the top and front views provided from VIMA-BENCH. As shown in Tab. 2, visual planning approach demonstrates greater robustness to limited observations, with a smaller performance drop in single-view compared to abstract methods. We attribute this to two potential reasons: First, accurate and grounded subgoal images generated in visual planners provide supplemental visual context, which can partly compensate for the insufficient perception to aid robustness under single-view. Second, by providing coherent subgoal images as visual landmarks, the requirements for the low-level action planner are reduced to basic single-object manipulation primitives in short horizon, with less reliance on rich visual perceptions. The experiments demonstrate that CoTDiffusion enjoys better robustness to restricted perception than abstract planners, highlighting the benefits of hierarchical framework decoupled visual planning and action planning.

Methodology	Multi-View	Single-View	Performance Drop
Gato	11.4 ± 1.6	6.5 ± 1.6	4.9(42% ↓)
Flamingo	18.9 ± 2.1	12.0 ± 2.4	6.9(36.2% ↓)
VIMA	49.5 ± 4.3	34.9 ± 3.4	14.6(29.4% ↓)
SuSIE	10.0 ± 2.3	7.9 ± 2.0	2.1(21.0% ↓)
+sub-prompts	43.0 ± 5.9	35.6 ± 6.1	7.4(18.8% ↓)
CoTDiffusion	$\mathbf{64.6 \pm 3.0}$	$\mathbf{56.0 \pm 2.4}$	**8.6**(**13.2%** ↓)

Table 2. The evaluations of performance drop of different methods on single-view and multi-view from VIMA-BENCH.

Table 3. Quantitative comparisons of FID betweeen methods on all three tasks, including visual rearrange, reasoning and constraints.

Methodology	Rearrange	Reasoning	Constraints
SuSIE [4]	19.7	18.3	23.2
- w/o finetune	38.4 ↑	33.0 ↑	52.7 ↑
- w/o stepwise prompt	32.6 ↑	28.1 ↑	34.7 ↑
CoTDiffusion (ours)	**11.3**	**10.8**	**8.6**
- w/o coarse pre-train	23.1 ↑	16.2 ↑	18.9 ↑
- w/o bi-direction align	18.7 ↑	15.0 ↑	15.8 ↑
- w/o frame concat	15.4 ↑	13.7 ↑	12.1 ↑

Fidelity of Image Generation Here we conduct further analysis of *visual planners* by comparing goal image quality against ground truth keyframes with FID as evaluation metrics. As Tab. 3 shows, CoTDiffusion still achieves much better fidelity than SuSIE even though SuSIE has got improved after fine-tuning on the same datasets in VIMA-BENCH. Moreover, SuSIE can only understand short instructions of single-object manipulation tasks, struggling to handle long-horizon instructions due to the lack of chain-of-thought reasoning capabilities to handle multi-modal prompts. With step-wise sub-prompts decomposed in advance, the performance of SuSIE gets largely raised but still underperforms CoTDiffusion, which has no need to explicitly decompose the general prompts and can generate subgoal images in an implicit chain-of-thought manner. Additionally, ablating coarse pretraining and bi-directional generation degrades performance, validating their benefits. The coarse-to-fine semantic alignment training training allows developing spatial reasoning prior to synthesis. Frame concatenation further guides coherent denoising by providing rich context information as visual priors to ground the current observation and enhance the fidelity of generation.

Accuracy of Instruction Following We evaluate instruction following accuracy via CLIP similarity between gen-

erated keyframes and general prompts, normalized by the CLIP score between ground truth ultimate goal image and prompts. The results in Fig. 5 demonstrate that CoTDiffusion enjoys better instruction following capabilities. Without chain-of-thought reasoning abilities, SuSIE struggles to follow instructions when given general multi-modal prompts, let alone generate subgoal images with smooth progressions. Progressively providing privileged prompts step by step improves SuSIE in instruction following and gradual advancement towards ultimate goals. However, text prompts decomposed by rules are not always perfect to convey the original multi-modal prompt, so the reliance on these prompts restricts the performance of SuSIE. In contrast, the coarse alignment pretraining and bi-directional generation can assist the diffusion model in tracking the progress of generated keyframes throughout the entire chain and generate sequenced keyframes incrementally advancing prompt instructions. Moreover, the implicit decomposition maintains semantic coherence between subgoals without overly large jumps and the smoothness and continuity between visual milestones make them more reachable by downstream foundation models with just single-object manipulation capabilities. Additionally, we observe that the bi-directional generation may impedes the diffusion model training if without coarse semantic pretraining. As the initial observation and prompt remain fixed across different generation steps, the align tokens at various stage tend to so similar that may confuse simultaneous progressive generation and reconstruction training.

Generalization across Tasks We evaluate the generalization ability in three levels with increasing difficulty: placement generalization which randomizes the novel placement of objects (L1), object generalization which provides the objects with novel attributes (L2), and task combinatorial generalization which complexes the prompts with extra novel instruction (L3). As Fig. 6 shows, CoTDiffusion exhibits strong zero-shot generalization on unfamiliar objects, colors, and shapes at different placements by explicitly visualizing them into coherent subgoals grounded on new concepts. This avoids re-grounding concepts from multi-modal prompts in the low-level foundation model for

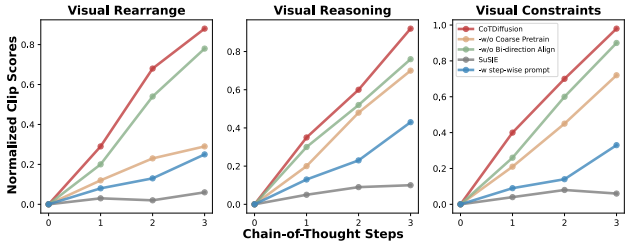

Figure 5. The normalized CLIP scores for each generation step, reflecting the step-wise accuracy of instruction following.

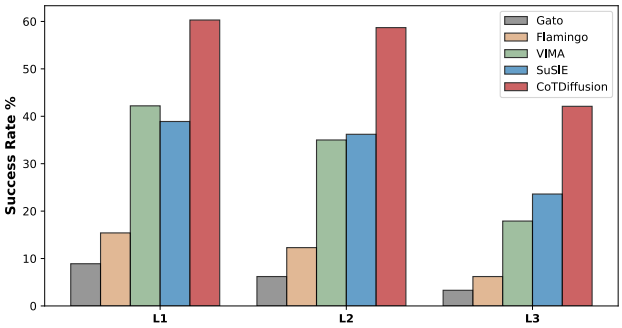

Figure 6. The evaluation on various generalization levels.

the placement and object. For L3 generalization, taking an example, we complex the prompt such as 'rearrange ... then rotate/twist/stack ...' to demand extra tasks such as rotating, twisting degrees, or stacking objects. CoTDiffusion achieves outstanding gain in the zero-shot performance of combinatorial tasks. When CoTDiffusion visualizes novel concepts into goal images, the foundation model can still accomplish the stack by simply achieving the provided subgoal, with no need to inherently understand novel skills like stack. This demonstrates the power of 'a image is worth a thousand words' - the subgoal images facilitate the generalization of the foundation model to unseen objects and novel combinatorial task while the foundation model simply achieves them with reuse of simple known primitives.

5. Conclusion

We presented CoTDiffusion, a hierarchical framework that integrates diffusion model as high-level module to translate the general multi-modal prompts into coherent subgoal images, serves as the visual milestones to anchor the low-level foundation model to plan action sequences, termed as 'generate subgoal images before act'. With the coarse-to-fine training for semantic alignment module, CoTDiffusion can identify the progress of generated subgoals images along reasoning chains, unlocking the chain-of-thought reasoning capabilities of diffusion model for long-horizon manipulation tasks. The experiments cover various long-horizon manipulation scenarios in VIMA-BENCH, and CoTDiffusion show the strong instruction following and outstanding performance gain compared to existed methods without visual planning. Incorporating commonsense knowledge from pre-trained MLLM like GPT-4V provides an avenue for more generalizable and promising reasoning in CoTDiffusion, which leaves as our future work.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant Nos. 92370132) and the Xiaomi Young Talents Program of Xiaomi Foundation.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 2, 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. *NeurIPS*, 2022. 6, 5

[3] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023. 2

[4] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 6, 7, 5

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 1

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 5

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*. 8

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1, 3

[11] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. 1

[12] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: Infinite resolution action detection transformer for robotic manipulation. *arXiv preprint arXiv:2306.17817*, 2023. 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[18] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1

[19] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5

[20] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020. 8

[21] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022. 1, 2, 3, 6, 5

[22] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1

[23] Zhengzhong Liang, Steven Bethard, and Mihai Surdeanu. Explainable multi-hop verbal reasoning through internal monologue. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. 3

[24] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 12

[28] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: An empirical study on video diffusion with transformers. *arXiv preprint arXiv:2305.13311*, 2023. 5

[29] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022. 3

[30] Yecheng Jason Ma, William Liang, Vaidehi Som, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. *arXiv preprint arXiv:2306.00958*, 2023. 3

[31] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. 6, 5

[32] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3): 7327–7334, 2022. 8, 11

[33] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023. 2, 3

[34] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 3

[35] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[36] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021. 3

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[40] Scott Reed and Nando De Freitas. Neural programmer-interpreters. *arXiv preprint arXiv:1511.06279*, 2015. 3

[41] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. 3, 6, 5

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 6

[44] 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. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 6, 12

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 5

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 3

[49] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 2

[50] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR, 2021. 6, 1

[51] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning*, 2023. 1, 2, 3