RoboDreamer: Learning Compositional World Models for Robot Imagination

Siyuan Zhou [1] **Yilun Du** [2] **Jiaben Chen** [3] **Yandong Li** [4] **Dit-Yan Yeung** [1] **Chuang Gan** [5]

https://robovideo.github.io/

Abstract

Text-to-video models have demonstrated substantial potential in robotic decision-making, enabling the imagination of realistic plans of future actions as well as accurate environment simulation. However, one major issue in such models is generalization – models are limited to synthesizing videos subject to language instructions similar to those seen at training time. This is heavily limiting in decision-making, where we seek a powerful world model to synthesize plans of unseen combinations of objects and actions in order to solve previously unseen tasks in new environments. To resolve this issue, we introduce RoboDreamer, an innovative approach for learning a compositional world model by factorizing the video generation. We leverage the natural compositionality of language to parse instructions into a set of lower-level primitives, which we condition a set of models on to generate videos. We illustrate how this factorization naturally enables compositional generalization, by allowing us to formulate a new natural language instruction as a combination of previously seen components. We further show how such a factorization enables us to add additional multimodal goals, allowing us to specify a video we wish to generate given both natural language instructions and a goal image. Our approach can successfully synthesize video plans on unseen goals in the RT-X, enables successful robot execution in simulation, and substantially outperforms monolithic baseline approaches to video generation.

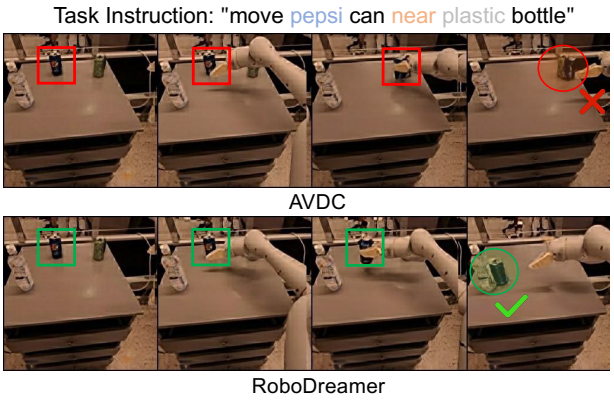

Figure 1: **Compositional Action Specification.** When existing text-to-video models (AVDC (Ko et al., 2023)) are given unusual combinations of language instructions, they are unable to synthesize videos that align accurately with these descriptions. RoboDreamer factorizes the generation compositionally, enabling generalization to novel combinations of language.

1. Introduction

Text-to-video models(Ho et al., 2022; Singer et al., 2022; Villegas et al., 2022) have seen extensive development in the setting of AI content generation, where models can generate high-quality videos given short text descriptions of motions. Such models have recently been applied in robotics, demonstrating significant potential in the development of policies, dynamic models, and planners (Du et al., 2023b; Ajay et al., 2023; Yang et al., 2023b). However, while natural language commands found in content generation typically focus on the global motion of a scene, natural language actions in robotics revolve around precise spatial rearrangements between objects.

Such commands, such as "move pepsi can near plastic bottle." remain challenging for existing models. As shown in Fig. 1, existing methods generate a video where pepsi can is placed near green can, failing to accurately capture the specified object relationship. Furthermore, these challenges become even more pronounced in scenarios where language instructions deviate from those encountered during training time, especially in reinforcement learning datasets where

[1]Hong Kong University of Science and Technology [2]Massachusetts Institute of Technology [3]University of California, San Diego [4]University of Central Florida [5]University of Massachusetts Amherst. Correspondence to: Chuang Gan <ganchuang1990@gmail.com>.

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

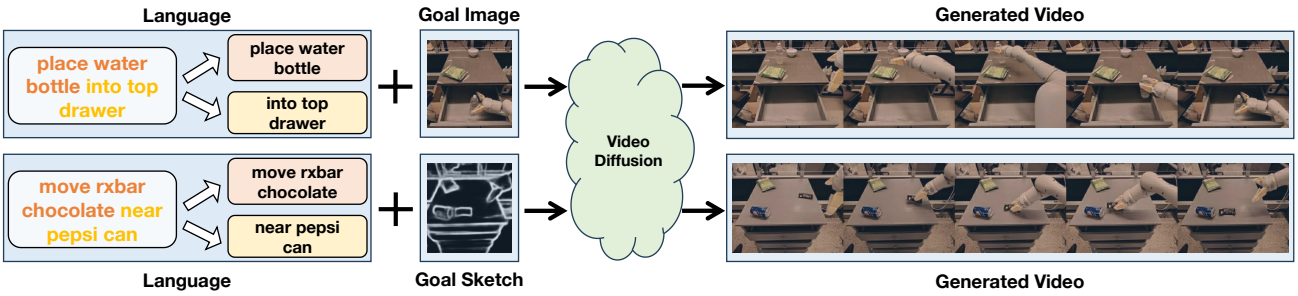

Figure 2: **Compositional World Models.** Given language instructions and multimodal instructions such as goal images and sketches, our approach factorizes the generation into a composition of diffusion models conditioned on inferred components. This enables our approach to generalize to both new combinations of language and multimodal input.

the data are scarce and natural language instructions are highly biased.

In response, we introduce RoboDreamer, a compositional world model capable of *factorizing the video generation process* by leveraging the *inherent compositionality of natural language*. This model is designed to equip conventional text-to-video generation systems with the ability to perform compositional reasoning. By utilizing a text parser, we dissect language instructions into a set of primitives, isolating actions and the spatial relationships between objects. These parsed components then serve as distinct conditions for a compositional set of diffusion models, enhancing the ability of each model to capture the nuances of spatial relationships among objects. Additionally, by decomposing a text goal into a set of text components, our approach naturally generalizes to new combinations of language as long as each parsed component is in distribution. This is crucially important in robotics, where there is a lack of systematic data covering all possible actions in an environment and a need to be able to generalize to new unseen actions.

While natural language is one representation to specify tasks a robot should accomplish, it is high-level and abstract, making it difficult to convey the precise nuances of motion over target goal configurations. Other modalities of information, such as a goal image provide much more detailed information on the final goal we wish to achieve. We illustrate how RoboDreamer also enables us to compose across *multimodal specifications* to flexibly specify goals at inference time. In contrast to prior work leveraging video models for robotics, we further condition video generation on multimodal task specification goal images and goal sketches. These modalities, particularly rich in spatial information, play a crucial role in clarifying ambiguities inherent in task execution instructions. Goal sketches, in particular, offer an intuitive and user-friendly means for spontaneous and on-the-fly task expression, akin to language instructions.

To compose these multimodal specifications together in RoboDreamer, we similarly factorize generation into a set of models jointly conditioned on language components as well as other multimodal components (Figure 2). We illustrate how this approach enables us to richly specify and generate videos given a large set of specified conditions, enabling us at inference time to combine larger and new combinations of both language and multimodal specifications, even if such paired conditions are not available at training time. Prior approaches, such as ControlNet (Zhang et al., 2023) introduce an additional encoder upon pre-trained text-to-image models to tackle this challenge, but this requires the availability of paired data across language and multimodal inputs and is limited at inference time to a similar combination of inputs that are seen at training time.

Our contributions are three-fold.

- We introduce RoboDreamer, a compositional world model capable of factorizing the video generation process by leveraging the inherent compositionality of natural language.
- We illustrate how RoboDreamer also allows us to combine multimodal information, enabling us to combine goal information from images with natural language.
- We empirically demonstrate that RoboDreamer achieves strong alignment with tasks under multimodal instructions and promising results when deploying on robot manipulation tasks.

2. Background

We first provide background information on how text-conditioned video generation can serve as a world model for robotics. As an initial step, we introduce formalism on how text-conditioned video generation can be used for planning and how we can implement such video level planning in Section 2.1. We then discuss how to execute video plans

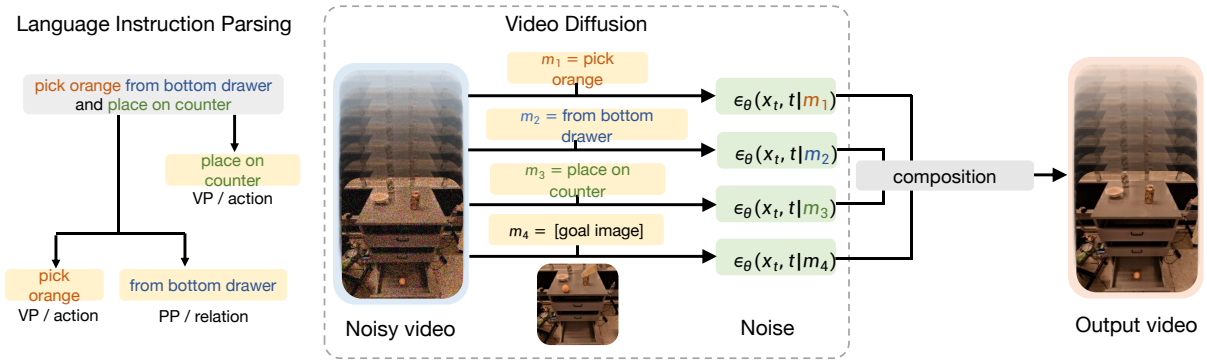

Figure 3: **Overall framework of RoboDreamer.** On the left, We leverage the natural compositionally of language to parse instructions into components like action phrases and relation phrases. On the right, we show how RoboDreamer composes multiple components.

in Section 2.2 and subsequently use videos in a close-loop manner for task completion.

2.1. Planning with Text-Conditioned Video Generation

We formulate task planning as a text-conditioned video generation problem using the Unified Predictive Decision Process (UPDP) abstraction from (Du et al., 2023b). Formally, a UPDP is defined as a tuple $\mathcal{G} = \langle \mathcal{X}, \mathcal{C}, H, \rho \rangle$, where $\mathcal{X}$ denotes the observation space of images, $\mathcal{C}$ denotes the space of textual task descriptions, $H \in \mathcal{N}$ is a finite horizon length, and $\rho(\cdot|x_0, c) : \mathcal{X} \times \mathcal{C} \to \Delta(\mathcal{X}^H)$ as a conditional video generator which synthesizes a video given a text description c and starting observation x_0. Given a UPDP $\mathcal{G}$, we then use a trajectory-task conditioned policy $\pi(\cdot|\{x_h\}_{h=0}^H, c) : \mathcal{X}^{H+1} \times \mathcal{C} \to \Delta(\mathcal{A}^H)$ to infer executable actions from synthesized videos.

Under this decision process, decision-making simply corresponds to learning ρ, which synthesizes videos of future image states given a natural language instruction c. This enables us to convert planning directly into a text-to-video generation problem. To implement this generation problem, we use the video diffusion model and use the base source code from (Ko et al., 2023).

2.2. Executing Videos Plans

Given a synthesized video plan $\tau = [x_1, \ldots, x_H]$, we follow (Du et al., 2023b) and use an inverse dynamics model $\pi(\cdot)$ to infer actions to execute to realize the video plan. The policy takes as input two adjacent image observations x_t and x_{t+1} in the synthesized video τ and outputs an action a to execute. We sequentially execute inferred actions starting from x_1 to x_{H-1} in the video.

To account for intermediate estimation error from predicting actions using an inverse dynamics model, we predict videos in a close-loop manner, where we periodically regenerate new video plans and execute actions based on this new plan.

3. RoboDreamer

In this section, we describe our proposed RoboDreamer in detail. To construct a compositional world model, RoboDreamer first uses a text parser to convert the compositional structure of language into a set of shared components in Section 3.1. Given these shared components, we illustrate how we can use compositional generation to guarantee generalization to novel combinations of these components in Section 3.2. Finally, we illustrate how such compositional components can be extended to multimodal conditions such as images or sketches for detailed compositional specification of synthesized plans in Section 3.3.

3.1. Text Parser

In contrast to many existing applications of text-to-video models in AI content creation, in the robotics setting, we are interested in synthesizing accurate video plans given detailed natural language instructions of actions. To solve robotics tasks, it is important that video models can synthesize actions that precisely rearrange one object with detailed specified relations with respect to nearby objects, including rearrangements not seen at training. However, reasoning and correctly generating videos subject to such object spatial relations is often challenging for existing models, especially on unseen textual descriptions.

To construct models that can more accurately synthesize spatial relations, we propose to decompose each spatial relation phrase into a set of compositional components. In particular, the actions of the tasks usually correspond to the verb phrase in the language and the object spatial relations correspond to the prepositional phrase after the verb phrase. Thus, given a text action instruction L, we decompose the instruction into a set of verb and propositional phrases l_i

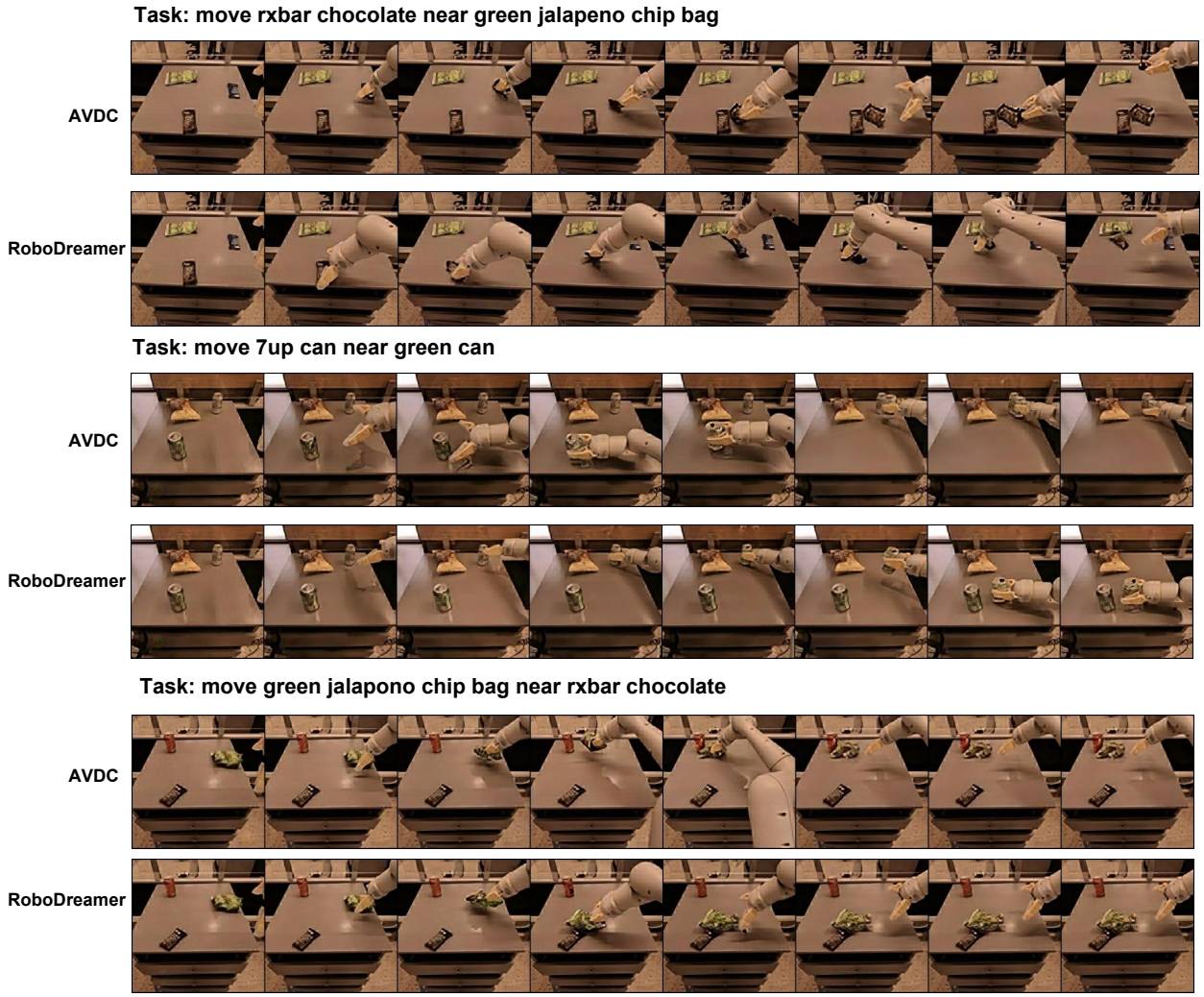

Figure 4: **Zero-Shot Video Generation.** Given novel combinations of natural language, RoboDreamer is able to substantially more accurately synthesize videos than a single monolithic text-to-video model.

which we use to condition a set of diffusion models on.

Take the task *place water bottle into bottom drawer* as an example. From this sentence, we parse the verb phrase *place water bottle* as the actions of the task and the prepositional phrase *into bottom drawer* as the object spatial relations. We utilize the pre-trained parser (Kitaev et al., 2018) and the rule-based approach to parse language instructions based on such characteristics. We provide a schematic of the text parsing in Figure 3.

3.2. Compositional Generation

Given a natural language instruction L that is parsed into a set of language components $\{l_i\}_{i=1:N}$, we propose to formulate our text-to-video model generative model $p_\theta(\tau|L)$ as a product of individual generative models defined on each parses language subcomponent l_i

$$p_\theta(\tau|L) \propto \prod_{i=1:N} p_\theta(\tau|l_i)^{\frac{1}{N}}. \quad (1)$$

Note that Eqn 1 naturally enables *compositional generalization* – given unseen combinations of natural language instructions L, our probabilistic expression in Eqn 1 will generalize perfectly as long as each parsed components l_i are in distribution. Thus, we can naturally map the compositionally of language into the video space through syntactic parsing.

To train our probabilistic expression in Eqn 1, we can leverage the close connection between diffusion models and EBMs (Liu et al., 2022; Du et al., 2023a), and learn a set of score functions $\epsilon(\tau, t|l_i)$ for each probability density $p_\theta(\tau|l_i)$. The score of the product of the densities in

Algorithm 1 Training

Input: Diffusion Model ϵ_θ, Training Step N
for i in $0, \ldots, N$ **do**
  Get training samples τ_0 and language instructions $L = \{l_i\}$
  Diffusion timestep $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\tau_t \leftarrow \sqrt{\bar{\alpha}_t}\tau_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
  $\mathcal{L}_{\text{MSE}} = \|\frac{1}{|L|}\sum_i \epsilon_\theta(\tau_t, t|l_i) - \epsilon\|$
  Take gradient descent step on $\mathcal{L}_{\text{MSE}}$
end for

Algorithm 2 Inference

Input: Diffusion Model ϵ_θ, Language Instructions $L = \{l_i\}$, Guidance weight w
$\tau_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for i in $T, \ldots, 1$ **do**
  $\epsilon_{\text{uncond}} \leftarrow \epsilon_\theta(\tau_t, t)$
  $\epsilon_i \leftarrow \epsilon_\theta(\tau_t, t|l_i)$
  $\tilde{\epsilon} \leftarrow= \epsilon_{\text{uncond}} + \sum_i w(\epsilon_\theta(\tau_t, t|l_i) - \epsilon_{\text{uncond}})$
  $\tau_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\tau_t - \frac{1-\alpha_t}{\sqrt{1-\hat{\alpha}_t}\tilde{\epsilon}}) + \sigma_t z$
end for
Return: τ^0

Equation 1 then corresponds to the average of score functions $\sum_i \frac{1}{N}\epsilon(\tau, t|l_i)$. We can then train this composite score function using the standard denoising diffusion training objective

$$\mathcal{L}_{\text{MSE}} = \|\frac{1}{N}\sum_i \epsilon(\tau_t, t|l_i) - \epsilon\|^2, \quad (2)$$

where τ_t corresponds to the original video corrupted with t steps of Gaussian noise.

One issue with directly optimizing Eqn 2 is that while the product of the composed distribution is encouraged to model the distribution of videos given text $p(\tau|L)$, each component is not necessarily encouraged to accurately model the distribution of videos given relevant textual information in the text snippet l_i, $p(\tau|l_i)$. To can encourage the score function $\epsilon(\tau_t, t|l_i)$ to capture this objective by also training the score function to denoise a video given only the relevant text snippet

$$\mathcal{L}_{\text{MSE}} = \|\epsilon(\tau_t, t|l_i) - \epsilon\|^2. \quad (3)$$

To unify both objectives, we use a hybrid training objective where given a set of language components $S = \{l_i\}_{i=1:N}$, we randomly a subset S' of M components and train with objective

$$\mathcal{L}_{\text{MSE}} = \|\frac{1}{M}\sum_i \epsilon(\tau_t, t|l_{S_i}) - \epsilon\|^2, \quad (4)$$

Given these learned score functions, at sampling time, we can sample from a novel combination of score functions. We illustrate overall training and sampling algorithms for our approach in Algorithm 1 and 2.

3.3. Multi-modal Composition

In addition to conditioning our generation at training time on a set of language components $\{l_i\}_{i=1:N}$ we can also condition our generation of a set of multimodal instructions M that $M = \{m_i\}_{i=1:K}$. We can express the likelihood of our video generative model $p_\theta(\tau|L, M)$ as the expression

$$p_\theta(\tau|L, M) \propto \prod_{i=1:N} p_\theta(\tau|l_i)^{\frac{1}{N+K}} \prod_{i=1:K} p_\theta(\tau|m_i)^{\frac{1}{N+K}}, \quad (5)$$

corresponding to the product of all models. Note that in the above expression, we can at inference time change the expression to adjust to both a variable number of modalities as well as language, in addition to novel combinations of both modality and language.

Based on the above expression, we can derive a variant of Eqn 4, to train a set of modality and language conditioned score functions so that they accurately model the above expression in Eqn 5 using the objective:

$$\mathcal{L}_{\text{MSE}} = \|\frac{1}{2M}\sum_i \epsilon(\tau_t, t|l_{S_i}) + \frac{1}{2M}\sum_j \epsilon(\tau_t, t|M_{S_j}) - \epsilon\|^2, \quad (6)$$

Once we have this set of score functions, we can then flexibly compose language and modalities in our compositional world model.

Model	Seen	Unseen
AVDC	63.1	46.9
HiP	70.3	50.1
RoboDreamer w/o	85.5	68.8
RoboDreamer	**90.1**	**81.3**

Table 1: **Human Study of Task Instruction Evaluation.** We present human evaluation for object relations and task completion of generated videos on seen and unseen tasks. RoboDreamer outperforms baselines, especially on unseen tasks.

4. Experiments

In this section, we evaluate the proposed RoboDreamer model in terms of its ability to enable generalizable compositional generation. We structure the experiments to answer the following questions:

- **RQ1:** Does RoboDreamer have zero-shot generalization abilities when encountering unseen task instructions? (Section 4.1)
- **RQ2:** Does compositional generation of multi-modal instructions improve spatial reasoning and object relations? (Section 4.1)

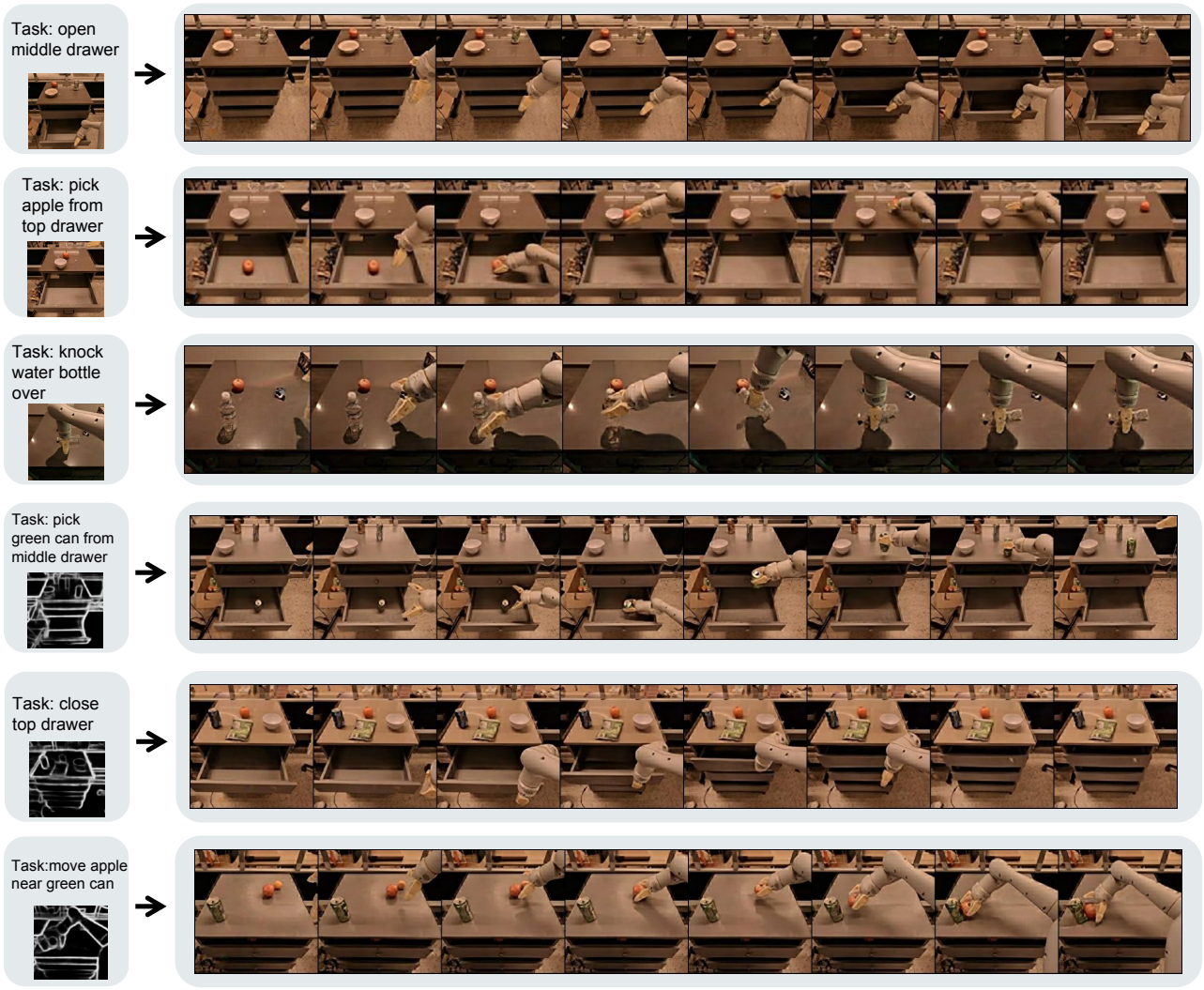

Figure 5: **Multimodal Compositionality.** RoboDreamer is able to compose multimodal inputs such as goal and sketch image conditioning with language instructions and synthesize plausible videos.

- **RQ3:** Can RoboDreamer be deployed on robot manipulation tasks? (Section 4.2)

4.1. Evaluation on Video Generation

In this section, we take a comprehensive evaluation of RoboDreamer's capabilities in two areas: its effectiveness in generating videos from textual instructions through compositional generation and its adeptness in incorporating multimodal conditions.

Experimental Setup. We take the real-world robotics dataset RT-1 (Brohan et al., 2022) to evaluate video generation. The dataset consists of various robotic manipulation tasks, i.e. *pick brown chip bag from middle drawer*. The robot is required to detect the middle drawer, pick a brown chip bag, and then place it on the table. Specifically, we train RoboDreamer on about 70k demonstrations and 500 different tasks. We randomly select language instructions as unseen test cases for evaluation.

Baselines. We compare RoboDreamer with AVDC (Ko et al., 2023), a video generation model for robotics; HiP (Ajay et al., 2023), a latent video diffusion model for robotics; RoboDreamer w/o, our model without text-parsing approach. To make the comparison fair, we only give language instructions to RoboDreamer and all other baselines. For all methods, we use pertaining T5-XXL as text encoder.

Metrics. Some works (Girdhar et al., 2023) have demonstrated that there are still no adequate metrics to evaluate the alignment between video generation and task instructions. As a result, we conduct human evaluation. Each sample is rated by at least three persons for their task completion. The scores are 0, 1, where 0 means the robotic planning in the generated videos is unreasonable or fails to solve tasks and

1 means the robotic planning is executable and succeeds in finishing the tasks.

Implementation Details. The video diffusion model of RoboDreamer is built upon AVDC and Imagen (Ho et al., 2022). We use a spatial-temporal convolution network in each ResNet block of U-Net for efficiency. We introduce the temporal attention layer on the ResNet block. We utilize a three-stage cascaded diffusion model for super-resolution. We use a similar tiling approach to enhance temporal consistency. Since the background of the all frames in one video should be consistent, we concatenate the input condition frame to the all noisy frames before feeding into U-Net.

We use pre-trained models to encode multi-modal instructions. We use the frozen T5-XXL text encoder (Raffel et al., 2020) for processing natural language instructions, which enables us to generate contextual embeddings. We borrow the pre-trained image encoder from VQVAE of Stable Diffusion (Rombach et al., 2022) for goal image and goal sketch instructions. The pre-trained downsampling encoder can be fast to extract spatial information for tasks and enhance efficiency. All modality embeddings will be fed into PerceiverSampler (Jaegle et al., 2021), an architecture designed for general inputs and outputs. The outputs will be integrated into the U-net by introducing a cross-attention layer into the ResNet blocks.

Zero-shot Generalization. We first evaluate that the text-parsing approach and compositional generation method bring zero-shot generalization abilities when facing unseen task instructions. The results, as highlighted in Table 1, reveal a significant enhancement in the model's performance. Through human evaluation, it becomes evident that our approach greatly improves the task alignment of video generation. This underscores the efficacy of RoboDreamer in understanding and executing unseen task instructions, thereby answering **RQ1**. As is illustrated in Figure 4, the baseline method AVDC and HiP fail to accurately infer the spatial relationship between objects, incorrectly placing them in the wrong positions. In contrast, our method successfully deduces these relationships, positioning the objects correctly according to the given instructions. By factorizing textual instructions into primitive components, RoboDreamer could successfully generalize to unseen task instructions by formulating them into combinations of seen components.

Multi-modal Generation. Subsequently, we evaluate the multi-modal-conditioned video generation capabilities of RoboDreamer, focusing on how it leverages visual information to enhance spatial reasoning of video generation. Specifically, we take the final frames as goal images and generate the goal sketches by annotators of ControlNet (Zhang et al., 2023). RoboDreamer's variations were methodically tested: RoboDreamer (t) is given only language description, RoboDreamer (t+i) is given language description and the goal image, and RoboDreamer (t+s) is given language description and goal sketches. As is shown in Table 2, with the additional help of goal image instructions or sketch instructions, RoboDreamer can achieve strong alignment on both human evaluations. This experimental result affirmatively addresses **RQ2**, with RoboDreamer's adeptness at integrating multi-modal instructions, RoboDreamer has a more nuanced understanding and execution of tasks.

Model	Human ↑	FVD ↓
AVDC	46.9	517.1
RoboDreamer (t)	81.3	487.8
RoboDreamer (t+s)	94.7	454.7
RoboDreamer (t+i)	**95.8**	**444.3**

Table 2: **Evaluation on Multi-Modal Generation.** RoboDreamer (t+s) and RoboDreamer (t+i) achieve strong performance on human evaluation and good video quality.

4.2. Evaluation on Robotic Planning

Finally, our examination extends to the practical applicability of RoboDreamer in robotic planning tasks. We investigate whether synthesized videos of RoboDreamer can do robotic planning.

Experimental Setup. We conduct experiments on RLBench (James et al., 2020). The agent captures observations as RGB images using multi-view cameras and controls a robotic arm with seven degrees of freedom (7 DoF) on RLBench. The environment is constructed to mimic real-world conditions, featuring high dimensionality in both observation and action spaces, which presents a significant challenge. There are 74 challenging vision-based robotic learning tasks whose categories vary from tool-using tasks, and pick-and-place tasks to long-term planning tasks. We follow the setting of previous works (Guhur et al., 2023) to use macro-steps, which will make the environment focus more on robot planning. We only consider RGB images from the front camera as observations, which makes RLBench much more challenging. For a fair comparison, we don't add goal images as instructions.

Baselines. We consider three baselines:

- Image-BC, an imitation learning approach given observation, states, and goal description.
- Hiveformer (Guhur et al., 2023): a transformer-based approach that integrates natural language instructions, multi-view scene observations, and a full history of observations and actions.
- UniPi (Du et al., 2023b): a method that utilizes text-to-video models to generate videos and inverse dynamic policy to predict actions based on the videos. We use the open-source text-to-video codebase from (Ko et al.,

Model	lamp off	lamp on	stack blocks	lift block	take shoes	close box	Average
Image-BC	60.1	47.0	0	0	0	82.4	31.6
Hiveformer	81.2	**53.2**	10.6	**28.2**	1.0	90.8	44.2
UniPi	70.6	47.1	7.1	23.3	3.8	94.1	41.0
RoboDreamer	**96.3**	51.9	**18.5**	22.2	**10.5**	**96.3**	**49.3**

Table 3: **Success Rate on RLBench.** The highest success rate demonstrates that the videos generated by RoboDreamer are feasible and executable and help robot planning.

2023) to train models.

Robot Planning. According to the results presented in Table 3, RoboDreamer achieves superior task success rates compared to baseline models even if RoboDreamer is only given observation from single cameras. As expected, Image-BC and Hiveformer are struggling in the long-term tasks that are *stack blocks* and *take shoes*. On the other hand, RoboDreamer achieves a success rate of 15% with the help of predicted future observations. UniPi performs poorly as it does not align with task instructions well. The strong performance of RoboDreamer demonstrates that the synthesized videos are significantly beneficial for robot planning.

5. Related Work

Diffusion Models for Decision-Making Diffusion models have emerged as promising generative models for many decision-making tasks (Chi et al., 2023; Pearce et al., 2023; Zhang et al., 2022; Liang et al., 2023; Huang et al., 2023; Liu et al., 2023b; Zhou et al., 2023). Some works (Janner et al., 2022; Ajay et al., 2022) train diffusion model on low-level state and action space on simulation data. They generate trajectories to do robot planning. While these works are hard to generalize to high-dimensional data like videos, some works (Du et al., 2023b; Ko et al., 2023) formulated the robot planning as a text-to-video generation problem. Most similar to our work, UniPi (Du et al., 2023b) trains a video diffusion model to predict future frames and gets actions with inverse dynamic policy. On the other hand, AVDC (Ko et al., 2023) utilizes pre-trained flow networks to predict the actions. However, previous works are limited in generalization to unseen tasks. Our text-parsing approach and composition abilities can enhance the zero-shot generalization abilities of RoboDreamer.

Compositional Generation Our work is also related to the compositional generation method (Bar et al., 2020; Liu et al., 2022; Yu et al., 2022; Ajay et al., 2023; Yang et al., 2023a; Zhang et al., 2023; Shi et al., 2023; Hu et al., 2024; Du et al., 2023a). The promising solutions are based on those works (Nie et al., 2021; Du et al., 2021; Liu et al., 2021; Gkanatsios et al., 2023) that probabilistically combing different generative models for jointly generating outputs. Most of them works have delved into the realm of compositional text (Deng et al., 2020; Liu et al., 2023a) and image generation (Shi et al., 2023; Liu et al., 2022; Du et al., 2023a). However, when it comes to the domain of text-to-video generation, the extent of exploration of compositionality is comparatively limited. In text-to-video generation, VideoAdapter (Yang et al., 2023a) introduces a novel setting by transferring pre-trained general T2V models to domain-specific T2V tasks with small amount of training data, however, it fails to genelize on unseen tasks or text descriptions. HIP (Ajay et al., 2023) aims to solve long-horizon tasks, however, it composes several expert foundation models. Different from previous research, our approach demonstrates how to construct compositional video-based world models, decomposing the learned probability distribution during training into a series of compositional components. This empowers us to seamlessly generate videos subject to combinations of detailed language and image specifications that are not seen during training.

6. Conclusion

In conclusion, we introduce RoboDreamer, a compositional approach to video generation that generalizes significantly better than prior works in video generation in robotics. By leveraging the natural compositionality of language and integrating multi-modal instructions, RoboDreamer demonstrates significant advancements in generating videos that accurately capture complex spatial relationships and object interactions. Experimental results verify RoboDreamer's capabilities in zero-shot generalization, multi-modal-conditioned video generation, and its potential application in robotic manipulation tasks.

Limitations Although RoboDreamer exhibits strong performance in robot planning tasks, it has several limitations. **(1)** While many robot tasks often use information from multiple cameras, RoboDreamer is limited to single camera views and is unable to consider the multi-camera information. This limits the applicability of RoboDreamer to many robotics tasks that require detailed 3D information. We believe that exploring how to add 3D inductive biases into RoboDreamer to consider multi-camera information is a rich source of future research. **(2)** RoboDreamer generalizes poorly to many real-world images we tested. We believe that existing robotics datasets are still limited in diversity, and

it may interesting to explore joint co-training of our compositional model across both robotics data and existing videos on YouTube to improve generalization. **(3)** The capabilities of video generation models, including ours, are constrained when it comes to moving-camera settings. Addressing these challenges necessitates the approach to stabilizing.

Impact Statement

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

Acknowledgements

This project is in part supported by Cisco, Google, and Amazon research funding. This work is also supported by a Research Impact Fund project (R6003-21) funded by the Research Grants Council of the Hong Kong Government.

References

Ajay, A., Du, Y., Gupta, A., Tenenbaum, J., Jaakkola, T., and Agrawal, P. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022.

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

Bar, A., Herzig, R., Wang, X., Rohrbach, A., Chechik, G., Darrell, T., and Globerson, A. Compositional video synthesis with action graphs. *arXiv preprint arXiv:2006.15327*, 2020.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

Deng, Y., Bakhtin, A., Ott, M., Szlam, A., and Ranzato, M. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.

Du, Y., Li, S., Sharma, Y., Tenenbaum, J., and Mordatch, I. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34: 15608–15620, 2021.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning*, pp. 8489–8510. PMLR, 2023a.

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

Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S. S., Shah, A., Yin, X., Parikh, D., and Misra, I. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

Gkanatsios, N., Jain, A., Xian, Z., Zhang, Y., Atkeson, C., and Fragkiadaki, K. Energy-based models as zero-shot planners for compositional scene rearrangement. *arXiv preprint arXiv:2304.14391*, 2023.

Guhur, P.-L., Chen, S., Pinel, R. G., Tapaswi, M., Laptev, I., and Schmid, C. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pp. 175–187. PMLR, 2023.

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

Hu, H., Chan, K. C., Su, Y.-C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., et al. Instruct-imagen: Image generation with multi-modal instruction. *arXiv preprint arXiv:2401.01952*, 2024.

Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., and Zhu, S.-C. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16750–16761, 2023.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Kitaev, N., Cao, S., and Klein, D. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint arXiv:1812.11760*, 2018.

Ko, P.-C., Mao, J., Du, Y., Sun, S.-H., and Tenenbaum, J. B. Learning to act from actionless videos through dense correspondences, 2023.

Liang, Z., Mu, Y., Ding, M., Ni, F., Tomizuka, M., and Luo, P. Adaptdiffuser: Diffusion models as adaptive self-evolving planners. In *International Conference on Machine Learning*, 2023.

Liu, G., Feng, Z., Gao, Y., Yang, Z., Liang, X., Bao, J., He, X., Cui, S., Li, Z., and Hu, Z. Composable text controls in latent space with ODEs. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16543–16570, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1030. URL https://aclanthology.org/2023.emnlp-main.1030.

Liu, N., Li, S., Du, Y., Tenenbaum, J., and Torralba, A. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Liu, W., Du, Y., Hermans, T., Chernova, S., and Paxton, C. Structdiffusion: Language-guided creation of physically-valid structures using unseen objects. In *RSS 2023*, 2023b.

Nie, W., Vahdat, A., and Anandkumar, A. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021.

Pearce, T., Rashid, T., Kanervisto, A., Bignell, D., Sun, M., Georgescu, R., Macua, S. V., Tan, S. Z., Momennejad, I., Hofmann, K., et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shi, C., Ni, H., Li, K., Han, S., Liang, M., and Min, M. R. Exploring compositional visual generation with latent classifier guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 853–862, 2023.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Yang, M., Du, Y., Dai, B., Schuurmans, D., Tenenbaum, J. B., and Abbeel, P. Probabilistic adaptation of text-to-video models. *arXiv preprint arXiv:2306.01872*, 2023a.

Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., and Abbeel, P. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023b.

Yu, W., Chen, W., Yin, S., Easterbrook, S., and Garg, A. Modular action concept grounding in semantic video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3605–3614, 2022.

Zhang, E., Lu, Y., Wang, W. Y., and Zhang, A. Lad: Language augmented diffusion for reinforcement learning. In *Second Workshop on Language and Reinforcement Learning*, 2022.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Zhou, S., Du, Y., Zhang, S., Xu, M., Shen, Y., Xiao, W., Yeung, D.-Y., and Gan, C. Adaptive online replanning with diffusion models. *arXiv preprint arXiv:2310.09629*, 2023.

A. Experimental Details

A.1. Video Diffusion

We list the details about our method RoboDreamer as follows.

1. Our method RoboDreamer is built upon AVDC (Ko et al., 2023) and Imagen (Ho et al., 2022) and we utilize a three-stage cascaded diffusion model for super-resolution.
2. For video diffusion models, we use 4 ResNetBlock within U-Net and each block is composed of spatial-temporal convolution layers and cross-attention layers with conditioned instructions.
3. We introduce temporal-attention layers in the last block within the encoder of U-Net and the first block within the decoder.
4. The base channel is 128 and the channel multiplier is [1, 2, 4, 8].
5. We train our video diffusion models with 256 batch size and 5e-5 learning rate on about 100 V100 GPUs.
6. We train base video diffusion models with $8 \times 64 \times 64$ videos and then subsequently upsample to $8 \times 128 \times 128$ and $8 \times 256 \times 256$ videos.

A.2. Other Details

1. RT-1 Dataset: RT-1 Dataset has about $70k$ demonstrations with an average length of 44. We sample one every 5 frames. There are about 500 tasks. We list some categories of tasks here: *pick*, *pick ... from ...*, *place*, *open*, *close*, *knock* and *pull*.
2. Inverse Dynamics Model: Inverse dynamics model is trained to predict actions given two adjacent frames and the current state. We use ResNet18 as the backbone followed by an MLP layer. This model is trained using Adam optimizer with a learning rate 1e-4 for 10K steps.
3. RLBench: We use Franka Panda arm with Franka gripper on RLBench. It has seven degrees of freedom (7 DoF) and 8-dimensional action space, along with an additional gripper state.
4. Human Evaluation: Each sample is rated by at least three human raters and we evaluate about 128 samples in total including more than 20 text prompts.

B. Additional Results

B.1. Visualization on RLBench

We visualize the tasks on RLBench in Figure 6.

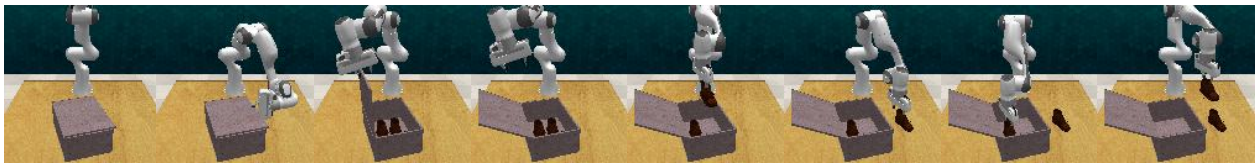

Figure 6: **Visualization on RLBench.**

B.2. More Results on Video Generation

More video generation results are shown in the website (https://robovideo.github.io/).

B.3. IMO Metrics

We use pretrained GroundingDino to detect the bounding boxes of the target objects. The IMO results (Iou of the bounding boxes) are shown in Table 4. The results are consistent with Human Evaluation. This supports the effectiveness of our approach in achieving better alignment through multi-modal inputs.

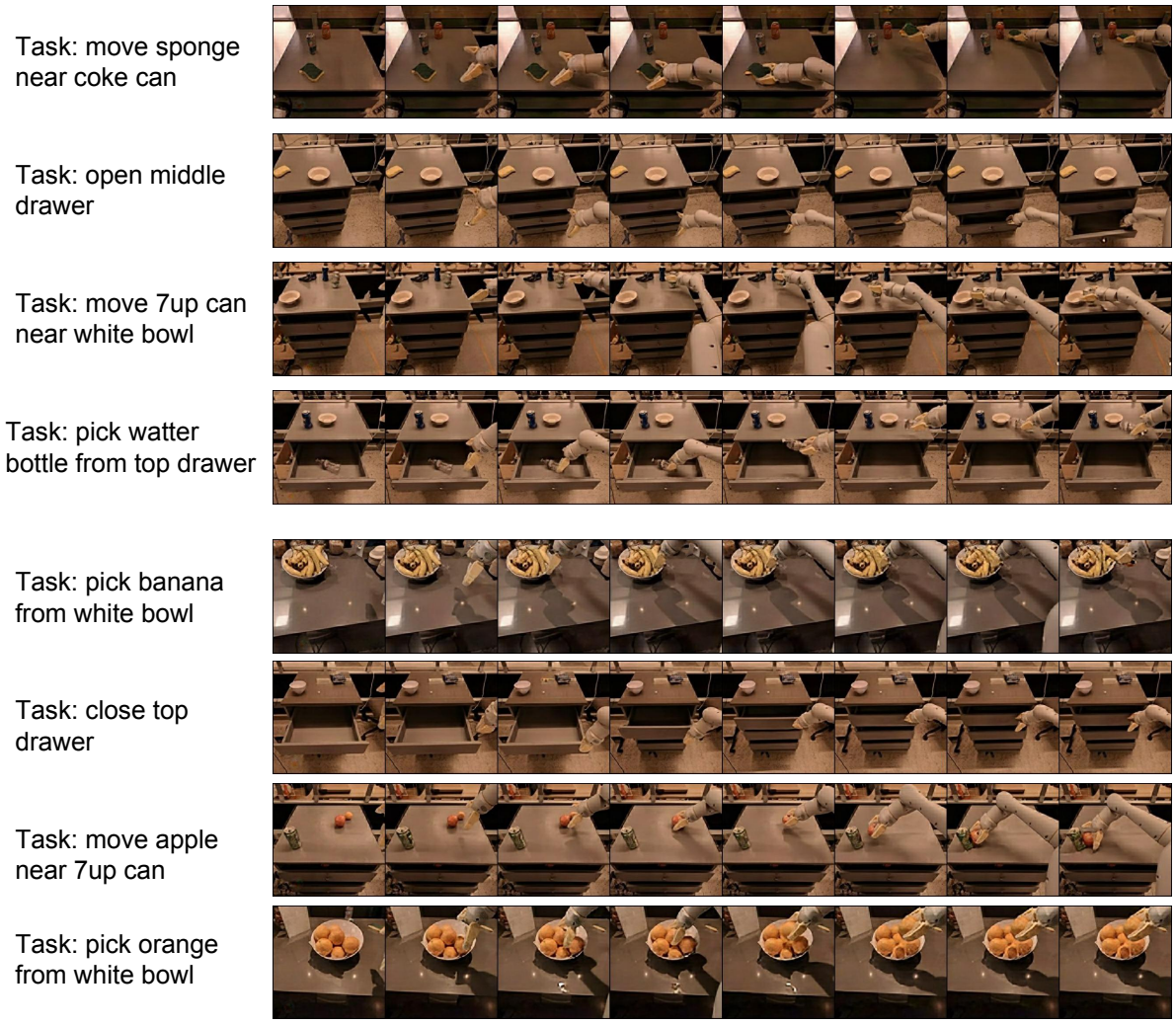

Figure 7: **Video Generation.**

Model	Human ↑	FVD ↓	IMO↑
RoboDreamer (t)	81.3	487.8	63.5
RoboDreamer (t+s)	94.7	454.7	72.5
RoboDreamer (t+i)	**95.8**	**444.3**	**78.1**

Table 4: **Evaluation on Multi-Modal Generation.** RoboDreamer (t+s) and RoboDreamer (t+i) achieve strong performance on human evaluation and good video quality.