DreamArrangement: Learning Language-conditioned Robotic Rearrangement of Objects via Denoising Diffusion and VLM Planner

Wenkai Chen*, *Student Member, IEEE*, Changming Xiao*, *Student Member, IEEE*, Ge Gao, Fuchun Sun, *Fellow, IEEE*, Changshui Zhang, *Fellow, IEEE*, and Jianwei Zhang, *Senior Member, IEEE*

***Abstract*—The capability for robotic systems to rearrange objects based on human instructions represents a critical step towards realizing embodied intelligence. Recently, diffusion-based learning has shown significant advancements in the field of data generation while prompt-based learning has proven effective in formulating robot manipulation strategies. However, prior solutions for robotic rearrangement have overlooked the significance of integrating human preferences and optimizing for rearrangement efficiency. Additionally, traditional prompt-based approaches struggle with complex, semantically meaningful rearrangement tasks without pre-defined target states for objects. To address these challenges, our work first introduces a comprehensive 2D tabletop rearrangement dataset, utilizing a physical simulator to capture inter-object relationships and semantic configurations. Then we present DreamArrangement, a novel language-conditioned object rearrangement scheme, consisting of two primary processes: employing a transformer-based multi-modal denoising diffusion model to envisage the desired arrangement of objects, and leveraging a vision-language foundational model to derive actionable policies from text, alongside initial and target visual information. In particular, we introduce an efficiency-oriented learning strategy to minimize the average motion distance of objects. Given few-shot instruction examples, the learned policy from our synthetic dataset can be transferred to the real world without extra human intervention. Extensive simulations validate DreamArrangement's superior rearrangement quality and efficiency. Moreover, real-world robotic experiments confirm that our method can adeptly execute a range of challenging, language-conditioned, and long-horizon tasks with a singular model. The demonstration video can be found at https://youtu.be/fq25-DjrbQE.**

***Index Terms*—Robotic rearrangement, Denoising diffusion, Prompt-based learning, Vision-language model.**

I. Introduction

FROM the perspective of embodied intelligence, how can we empower the household robots with the capability to discern *how and where they should rearrange messy tabletop objects* especially involving ambiguous human instructions? Comprehensive reasoning and planning across diverse constraints from object geometry, language-conditioned tasks, collision physics, and human preference, pose a significant challenge for autonomous robots operating within varied and unstructured household scenarios, such as automated packaging and sorting in warehouses, kitchen cleaning, and complex assembly tasks in manufacturing. In this work, we study this challenge by introducing human-like imagination and planning ability to the robots in the context of human instructions and prior observations.

Robotic rearrangement can be defined as a canonical task: given a previously unseen environment, the robot needs to rearrange each object into an appropriate pose to form a specified structure following human preference. This paradigm can also encompass a diverse array of activities, such as making a bed, ironing clothes, and cleaning a room. However, we specifically concentrate on investigating tabletop object arrangements, considering this challenging but tractable [1]. Recently, some approaches that leverage large language models (LLMs) have demonstrated a strong generalization for robots to understand complex semantic contexts and generate long-horizon planning for tabletop arrangement task [2]–[4]. However, the goal states of different objects still need to be manually specified in the prompt instructions. Furthermore, to estimate the target states of objects intelligently, some generative work based on VAE [5] and diffusion models [6], [7] has been proposed to endow the robot with human-like imagination, hopefully generating and refining the distribution of object poses. For instance, [6] proposes to utilize DALL-E, a web-scale artificial intelligence-generated content (AIGC) model, to generate a target image that implicitly incorporates various objects the robot observes. Nevertheless, the exclusive reliance on textual input for image generation has proven to be notably unstable and inefficient in real-world robot manipulation, primarily due to the neglect of crucial observational cues. Inspired by this prior work, building a model that conducts observation reasoning first and then imagines goal states intuitively via language is a crucial step towards autonomous robotic rearrangement.

On the other hand, considering functional and stylistic inter-object relationships emerges as a critical dimension for real-world robotic rearrangement [8]. For a given "messy" scenario, a "clean" arrangement should not be deterministic because there exists a plurality of desirable layouts from different human preferences. Thus, beyond the initial phase

∗ Indicates equal contribution.

This research was funded by the German Research Foundation (DFG) and the National Science Foundation of China (NSFC) in the project Crossmodal Learning, DFG TRR-169/NSFC 62061136001.

W. Chen and J. Zhang are with TAMS (Technical Aspects of Multimodal Systems) group, Department of Informatics, Universität Hamburg, Germany.

C, Xiao and C. Zhang are with Institute for Artificial Intelligence, Tsinghua University (THUAI), Beijing National Research Center for Information Science and Technology (BNRist), and Department of Automation, Tsinghua University, Beijing, China.

F. Sun is with Department of Computer Science and Technology, Tsinghua University, Beijing, China.

G, Gao is with Mech-mind Robotics, Beijing, China.

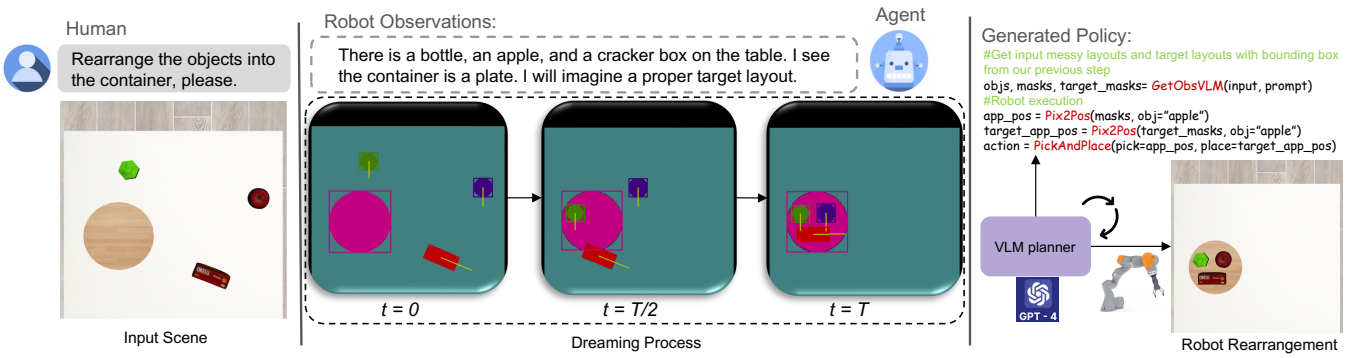

Fig. 1. Overview of the proposed scheme on the robotic rearrangement task when the multi-object structure setting is *containing*.

of estimating the goal poses of objects, the subsequent phase involves reorganizing the global layout of rearranged objects to align with human preferences, often communicated through language instructions. Moreover, to improve the real-world rearrangement efficiency, we also need to reduce the duration cost of the long-horizon manipulation by considering the motion distance of each object as much as possible. Despite notable advancements in learning-based scene synthesis and robotic rearrangement methods [7]–[10], there remains a challenge to meet diverse desiderata in real human-robot cooperation environments.

In this paper, we design a novel robotic arrangement scheme to solve the aforementioned flaws and maximize versatility and adaptivity, where the robot can rearrange objects in different goal poses and structures via language instructions without extra manual intervention. Specifically, we first construct a kitchen-based tabletop arrangement dataset consisting of four different global structures - *horizontal, vertical, circle, and containing*, and two local regularities - *symmetry* and *uniform*, where 22-class objects with different shape scales and texture materials are selected. Given that the input is a messy scene with human language instructions, we propose a transformer-based multi-modal denoising diffusion framework to estimate the goal states of objects by implicitly reasoning multi-object semantic relations.

Furthermore, we treat the planning problem of robotic rearrangement as a long-horizon estimation task by utilizing a frozen vision-language model (VLM) like GPT-4 to bridge connections between language text, visual perception, and robotic action. When prompted with several examples followed by the corresponding rearrangement policy, VLM planners can take in new language instructions and semantic contexts from initial "messy" scenes and predicted "clean" scenes, autonomously generating a new robotic arrangement policy. An example of the whole human-robot arrangement process of the *containing* structure can be visualized in Fig. 1. It describes a task scene in which a household robot needs to place all objects on the table into a container like a plate or box without causing objects collision and penetration. Finally, the proposed scheme is evaluated in both simulation and real robot experiments and compared with several state-of-the-art baselines, demonstrating that it can achieve better rearrangement quality and efficiency for different structure-based rearrangement tasks. The contributions of this paper are described as follows:

1) Considering the differentiated requirements of inter-object relationships and human preference in the robotic rearrangement task, we construct a 2D kitchen rearrangement dataset consisting of a variety of household object scenes with different global structures and local regularities.

2) To generate a high-quality rearranged scene, we propose a transformer-based multi-modal denoising diffusion model, which can effectively reason semantic and geometric relations from diverse objects, and explicitly predict the goal states of objects instructed by contextualized language representation.

3) To obtain the optimal layout in the real world, we propose an efficiency-oriented rearrangement learning strategy, which pursues a minimal average motion distance of objects.

4) Inspired by prompt-based learning, we integrate the generative model with VLMs to formulate a VLM planner, which outputs robot action policies in different arrangement tasks and can be directly deployed into a physical robot.

This paper is organized in the following manner: Sec. II provides a review of literature relevant to our study. The problem we aim to address is detailed in Sec. III. Secs. IV and V discuss our approach and the experimental validation, respectively. The paper is concluded in Section VI.

II. Related Works

A. Language-based Robotic Manipulation

Language is a flexible and instinctive medium, enabling humans to specify tasks, communicate contextual details, and express their intentions. Much work about language-conditioned robot manipulations has been proposed to control a robot by generating low-level policies via reinforcement learning (RL) or imitation learning (IL) [11]–[14]. [15] proposes using language as abstract representations of hierarchical RL framework, demonstrating that the agent can learn compositional tasks like object sorting and multi-object arrangement in a simulation environment. Furthermore, [16] designs a novel RL agent that directly maps language instructions and raw visual input to generate a sequence of actions without requiring intermediate representations and planning procedures. However, language-conditioned RL methods are difficult to deploy into real physical robots due to the challenge of learning the

relationship between language and multimodal sensor data in the unstructured robot environment. To further improve learning efficiency, other researchers adopt the language-conditioned IL approaches, where agents are trained to perform tasks by mimicking the actions demonstrated by a human expert. Focusing on the *containing* task, [17] first proposes a language-conditioned visuomotor policy utilizing unstructured and unlabeled data collected from a teleoperated robot in a physics simulator. [18] further integrates the low-level motion controller into the language-conditioned learning framework. Both test results indicate that the robot can hopefully accomplish long-horizon tasks in the simulation environment. However, these IL-based methods require a large and diverse set of high-quality demonstration data. Acquiring such data on actual robots is a process that demands considerable time and resources. Contrary to prior efforts in language-conditioned research, our work emphasizes the utilization of language instructions to steer the denoising diffusion process, where the target states of objects will be estimated and used for the subsequent robot planning.

B. Large Foundation Models in Robotics

Recently, large foundation models based on language and vision have become a dominant paradigm in solving long-horizon robotic manipulation tasks [2], [19]–[22]. They demonstrate strong few-shot or zero-shot reasoning ability to any text or vision input through just prompting by human instructions [23]. SayCan [20] uses a large language model (LLM) to perform various tasks, where language objectives are destructed into a hierarchical sequence of instructions. These instructions are subsequently fed into skill-oriented value functions and search heuristics to obtain optimal action sequences. Informed by multimodal prompting, Socratic Models [21] exhibits a modular framework to capture multimodal information and leverage LLMs to achieve zero-shot robotic perception and planning. Furthermore, Code-as-Policies [2] adopts the LLMs to generate a policy code of robot action, showing LLMs have a strong programming ability in controlling robots by recomposing perception and controller API functions. Utilizing the capability to generate codes, [22] uses LLMs to integrate 3D value maps into the robotic observation space after inferring affordances and constraints from language instructions, which produces low-level control on the contact-rich manipulation tasks successfully. Nevertheless, the final goal states of each robot task from previous work on LLMs remain predominantly predefined, relying on human expertise or demonstrative guidance encapsulated within the prompt instructions. In contrast, our work adopts the diffusion model to make the robot imagine an appropriate rearranged layout from different objects autonomously. This conceptualization is subsequently incorporated as a visual cue within the VLMs module to facilitate the generation of robotic policy code.

C. Diffusion Models

In the computer vision field, diffusion models have risen to prominence as leading generators of data, distinguished by their ability to accurately model complex distributions and generate a diverse array of high-quality samples. [24], [25]. The concept draws inspiration from the physical phenomenon of diffusion, where particles migrate from regions of higher concentration to lower concentration until a state of balance is achieved. Many applications from diverse domains, such as text, image, audio, and video, demonstrate that diffusion models can significantly improve the quality, realism, and creativity over previous generative models [26], [27]. Especially for the text-to-image diffusion models, their groundbreaking synthesis abilities with input from text description can significantly improve creating efficiency [28], [29]. However, these models offer limited control over the content they generate, primarily achieved through a single text-based input modality. Some techniques have been developed to enhance performance and gain more precise control using various input types, such as contextual layouts and class labels. These techniques strive to finely tune the creation of content by adjusting the generation process following the model that has been pre-trained. [30]–[32]. Taking an example of the inpainting task, [32] proposes a solution to achieve image inpainting successfully by leveraging a pre-trained vision-language model (CLIP), where the inpainting process is guided from a text description along with an ROI mask. Inspired by the recent development of controllable diffusion models, we design our diffusion architecture by considering natural language instruction, designated placement position, and diverse object attributes to generate a clean scene for different initial messy scenes.

D. Tabletop Robot Rearrangement

The objective of an intelligent robotic rearrangement system is to equip robots with the ability to understand their surroundings and interact with humans, thereby achieving precise and efficient object repositioning according to different structures or criteria that reflect human preferences [10]. Various approaches have been explored to tackle this challenge. Typically, [33] proposes to utilize an RL strategy based on the proximal policy optimization (PPO) algorithm to push irregular objects on the table inside a crate, which is hard to generalize to other rearrangement tasks because of the fixed position of the crate on the table. To improve the generalizability, VIMA [34] introduces prompt-based learning to train a multimodal generalist agent, achieving a simple zero-shot robot arrangement setting in the simulation environment. Nevertheless, it is still difficult to deploy in real robot experiments due to the lack of human-designed visual prompts. Moreover, [10] first introduces the concept of semantic structure in the robot arrangement task, which necessitates a robot's ability to understand the relationships between scattered objects and subsequently rearrange them into a spatial structure instructed by human languages. However, the efficiency is compromised by its sequential processing, where the goal state of the current object is estimated only after finishing the arrangement of the previous object. To address this inefficiency, StructDiffusion [7] implements a 3D diffusion-based approach based on the same dataset, achieving a better rearrangement performance. However, we found that the predicted object states for a given structure demonstrate negligible layout adaptability on the table when the initial messy observation and motion distance of

objects between the messy scene and the rearranged scene are not taken into account. This issue largely results from all target object states being derived from predetermined Gaussian noise throughout the denoising diffusion process. Moreover, the inherent design of the dataset presents challenges in enabling scattered objects to form varied structures upon completion of the rearrangement process.

To overcome the limitations in prior work, we first construct a dataset consisting of different semantic structures corresponding to language instructions where the same messy scenes can exhibit different goal scenes. Primarily, we add the *containing* structure in our rearrangement task as it represents an important application in the industrial sorting task. Furthermore, we design a transformer-based multi-modal diffusion architecture to generate goal states of objects according to language commands and prior observation. For the sake of improving real-world rearrangement efficiency, we also add a constraint to minimize the average moving distance of objects. To reduce human intervention when executing such a long-horizon task, we further integrate the proposed target generation network into a VLM module via prompt-based learning, where robot policy codes are generated autonomously.

III. Problem Statement

We introduce DreamArrangemnt, a novel robotic arrangement scheme designed to comprehend diverse human language instructions and the distribution of 2D object scenes including variations in attributes like semantic classes, geometric shapes, and placements of multiple objects, which shows the ability to perform a long-horizon manipulation task autonomously.

We consider the initial tabletop scenes where all objects are scattered in an image coordinate system, starting from the top left corner as the origin. In each messy scene S, we depict a combination of of a table T and objects $\{o_1, ..., o_N\}$. To achieve semantic rearrangement based on human preference, a structure-based language instruction $\mathcal{L}$ (e.g., "rearrange all objects into a circle shape") is also given. To enhance contextual understanding, we further employ approaches from text summarizing (e.g., prompt-based LLM parsing or search-based word dictionary) to decompose the abstract language into specified word tokens $\mathcal{L} \rightarrow (l_1, l_2, ..., l_n)$. This study primarily explores the challenge of generating a language-conditioned clean object scene S^* for a robot $\mathbf{r}$. S^* can be directly used in the planning phase as a visual prompt module in the VLM planner, finally generating a manipulation policy $\mathcal{P}$. We formulate this as an optimization problem to use the robot $\mathbf{r}$ to rearrange a "messy" scene S under a language instruction $\mathcal{L}$ via learning a bijection f of paired objects and minimizing their motion distance, referring to the ground truth "clean" scene $\tilde{S}$:

$$\begin{aligned} f^* = \underset{f}{\operatorname{argmin}} \quad & \mathcal{F}_{arrangement}(S, \mathcal{L}) + \lambda \mathcal{F}_{motion}(S, \mathcal{L}), \\ s.t. \quad & \mathcal{F}_{arrangement}(S, \mathcal{L}) = f(S, \mathcal{L}) - \tilde{S}, \\ & \mathcal{F}_{motion}(S, \mathcal{L}) = f(S, \mathcal{L}) - S, \end{aligned} \tag{1}$$

where λ is the weight hyperparameter of the $\mathcal{F}_{motion}(S, \mathcal{L})$ term. Then the policy can be expressed as:

$$\mathcal{P} = VLM(S, f^*(S, \mathcal{L}), \mathcal{L}), \tag{2}$$

More specifically, each object o in the input scene S is defined by its semantic class $c \in \mathbb{R}^C$, 2D oriented bounding box size $s \in \mathbb{R}^2$, object translation $t \in \mathbb{R}^2$, and object rotation $r \in SO(2)$[1], respectively. Since the *containing* structure is a special semantic scene, we define an additional 'mask' object class m to represent *containers* like plates and tables. Besides, we use type $tp \in \mathbb{R}^T$ instead of c to differentiate different containers. In summary, we denote each scene S as follows:

$$S = \{m_i, ..., o_i, ...\}, m_i = (t_i, r_i, s_i, tp_i), o_i = (t_i, r_i, s_i, c_i). \tag{3}$$

The object semantic class label c_i and container type label tp_i are represented as one-hot vectors of C and T classes, respectively, and the 2D bounding box size s_i is obtained by performing the principal component analysis (PCA) and then computing the positional relation of 4 corners. The values of translation t_i and rotation r_i are characterized by calculating the center position and the orientation angle of the bounding box. To facilitate a stable training process, we further use the normalization operation to make t_i and s_i into the same range of $[-1, 1]$ as r_i.

IV. Methodology

A. Denoising Diffusion Models

Denoising Diffusion Models [24], [35] are a class of generative models that learn data distribution by progressively denoising from a tractable noise distribution. Below, we provide a brief preliminary introduction from a score-based perspective. For more details, please refer to [35]. Given various samples from an unknown data distribution $q_0(x)$, our goal is to train a model capable of generating new samples that mimic the original distribution $q_0(x)$. A critical mechanism employed in this endeavor is Langevin dynamics, a concept borrowed from the domain of physics. This approach can produce samples from a distribution $p_{data}(x)$ when its *score*, defined as its gradient $\nabla_x \log p_{data}(x)$, is known. Starting from x_T of any prior distribution, the Langevin method recursively denoises the data as follows:

$$x_{t-1} = x_t + \alpha_t \nabla_{x_t} \log q_0(x_t) + \beta_t \epsilon, \tag{4}$$

where α_t and β_t are pre-defined step sizes associated with the time step t and $\epsilon \sim \mathcal{N}(0, I)$ is a stochastic term. As T becomes sufficiently large, the final obtained x_0 will converge to a sample drawn from $q_0(x)$.

We aim to train a neural network s_θ to approximate the *score* of the target distribution. The denoising score matching technique [36] is adopted to make the estimation of *score* tractable, with the key insight being to utilize conditional distribution settings. This involves perturbing $x_0 \sim q_0(x)$ with various noise kernels $q_t(x_t|x_0)$ across a spectrum of step parameters $t \sim \mathcal{U}[1, T]$. The original score matching objective of the perturbed distribution $q_t(x)$ can be expressed as $\mathbb{E}_{q_t(x_t|x_0)q_0(x_0)} \|s_\theta(x_t) - \nabla_{x_t} \log q_t(x_t|x_0)\|^2$. As demonstrated in [35], the final optimal network parameter θ^* for this objective should ensure $s_{\theta^*}(x) \approx \nabla_x \log q_t(x)$. Moreover, when employing Gaussian kernels $q_t(x_t|x_0) =$

[1]The first column of the rotation matrix is used.

TABLE I
OBJECT ATTRIBUTES AND SPATIAL STRUCTURES IN OUR DATASET

Entity	Type	Name
Object attributes	class (22)	apple, bear, banana, bowl, box, can, cracker_box, cup, fork, knife, lemon, milk...
	material (3)	YCB texture, metal, wood
	scale ratio (3)	0.8, 1.0, 1.2
Spatial structures	global structure	horizontal, vertical, circle, containing
	local regularity	symmetry, uniformity

$\mathcal{N}(x_0, \sigma_t^2)$ with pre-defined noise levels σ_t, the *score* of the conditional probability density can be analytically derived as $\nabla_{x_t} \log q_t(x_t|x_0) = \frac{x_0 - x_t}{\sigma_t^2}$. Consequently, the unified objective amalgamating all procedural steps is formulated as:

$$L_{score}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[1,T], q_t(x_t|x_0)q_0(x_0)} \lambda_t \left\| s_\theta(x_t) - \frac{x_0 - x_t}{\sigma_t^2} \right\|^2, \tag{5}$$

where λ_t denotes the objective weight, pragmatically set to σ_t^2.

In summary, we need to first optimize the *score* network s_θ to minimize objective Eq. 5. After that, we use the trained model $s_{\theta^*}(x_t)$ to incrementally refine the approximation of $\nabla_{x_t} \log q_0(x_t)$ as per the Langevin dynamics, facilitating an update along the Markov chain with Eq. 4 to generate new samples ultimately.

B. 2D Object Rearrangement Dataset

To facilitate tabletop robotic rearrangement, it's necessary to collect a large object rearrangement dataset which includes different object categories and spatial structures. However, collecting such a dataset involving complex physical interactions in the real world can be time-consuming, labor-intensive, and costly. In this work, we collect a 2D synthetic dataset based on the Mujoco physics simulator, consisting of 2223 clean object scenes. A physics simulator can help us precisely control the position, orientation, scale, and texture of each object and keep each object in the rearrangement scene collision-free and penetration-free. Additionally, it is convenient for us to describe each clean rearrangement structure with high-level language instruction. Specifically, we adopt 22 household object models from the YCB objects and ShapeNet objects as our object database. For each valid clean scene in the dataset, we preprocess it using instance segmentation and extract the oriented bounding box of each object as its explicit representation. The obtained scene $\tilde{S}$ will be regarded as our target output of the generation model. To simulate different messy scenes in our daily lives, we further perturb the target scene on the fly to generate clean-messy pairs and re-associate objects within each category, where the generated messy scenes will serve as the input data.

More importantly, high-level language instruction corresponding to a structure usually conveys different object layouts in the real world. As seen in Fig. 2, when we tell the household robot: "Based on the current messy scene, please rearrange the apple, lemon, orange, and peach into a horizontal shape", the mainstream solutions [5]–[7] will make the robot arrange objects into a centered layout as in clean scene1, which is a common pattern in their training data. However, according

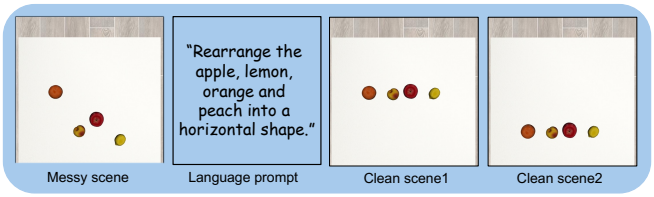

Fig. 2. Comparison of different generation results for clean scenes based on the same messy scene and language prompt.

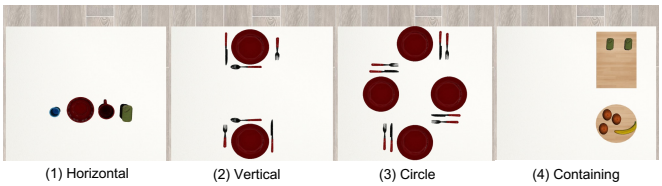

Fig. 3. The four kinds of global structures in our dataset: horizontal, vertical, circle, and containing.

to our human experience, we prefer to arrange the unordered objects into clean scene2, because it can save a large amount of time and effort. Therefore, to avoid the drawback in previous works that diverse initial scenes are arranged into the identical layout given the same instruction, we adopt the technique of data augmentation to enrich layout variations of target scenes with the same structure in our dataset.

Another rearrangement setting is that we want to arrange the same configuration on the table into different structures given different language prompts. We further design four kinds of physically meaningful spatial structures to pair with text descriptions. As shown in Fig. 3, the structures of *horizontal*, *vertical*, and *circle* represent all objects forming a horizontal, vertical, and circle shape globally, respectively. Since the *containing* structure involves the additional 'mask' object class m_i, we describe it as placing different objects in different containers, including plate-like containers and box-like containers. The semantic and geometric parameters of these containers will also be employed in the *containing* rearrangement task.

Moreover, to distinguish the difference of local distribution in real-world table settings, we introduce the concept of *symmetry* and *uniformity* in language instruction. Taking forks, knives, and plates as an example, *symmetry* represents that a pair of knives and forks are placed on varied sides of the plate while *uniformity* denotes that knives and forks are positioned on the same side of the plate. Finally, all object attributes and spatial structures in our dataset are shown in Tab. I.

C. Vision and Language Parsing

Based on our collected synthetic dataset, we further propose a scheme for solving the tabletop rearrangement task as shown in Fig. 4. A messy scene S and a high-level text description $\mathcal{L}$ from human language are taken as the input.

Object Detection and Parameterization: In order to obtain the geometric and semantic attributes of all objects, including o_i and m_i, in the messy scene S, we first adopt the latest Grounded Segment Anything Model (SAM) [37], which has

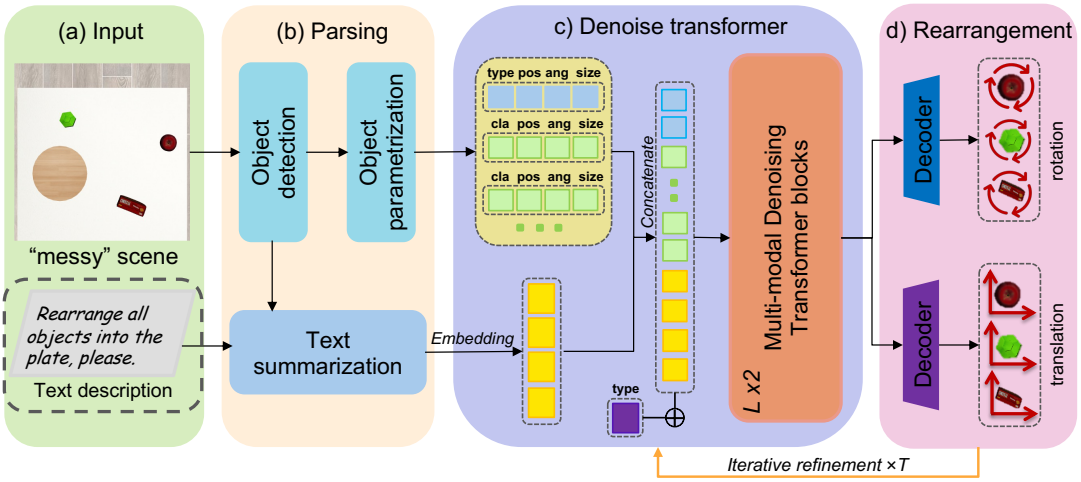

Fig. 4. An overview of the proposed conditional rearrangement diffusion network. (a) We sample the combinations of text descriptions from the humans with the messy observation as input. (b) The parsing process is to obtain explicit object attributes and word tokens from input. (c) We build a denoising diffusion framework with transformer architecture that separately encodes object attributes and word tokens into latent space. (d) To achieve the rearrangement task, the direction of translation and rotation of each instance are iteratively refined during the limited denoising steps T.

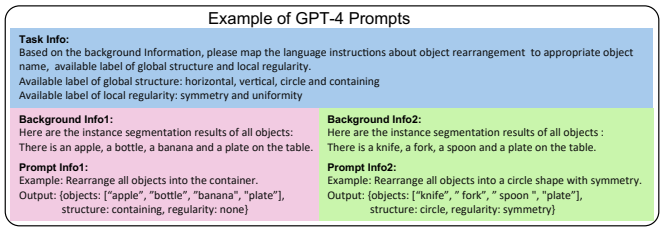

Fig. 5. Example of GPT-4 prompts that map human language instruction and segmentation results from vision parsing to concise word tokens.

shown a strong zero-shot ability for object recognition and instance segmentation. Then the segmented results are subjected to principal component analysis (PCA) to get the oriented bounding box of each object. Specifically, the values of object translation t_i and rotation r_i are derived by averaging all pixel points p_m of the object and establishing the covariance matrix:

$$t_i = \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} p_m, \tag{6}$$

$$r_i = \underset{v,\ \|v\|_2=1}{\operatorname{argmax}} \quad v^T \left[\frac{1}{\mathcal{M}-1} \sum_{m=1}^{\mathcal{M}} (p_m - t_i)(p_m - t_i)^T \right] v, \tag{7}$$

where $\mathcal{M}$ is the number of pixel points in the segmented object and the vector v indicates the projection direction to be searched for. Next, we employ separate neural network layers to encode the geometric and categorical features to obtain the instance embedding for the regular object o_i and the mask embedding for the container object m_i.

Text Summarization: To encode the natural language instruction into implicit representation, we need to distill the most important information and convert it into a condensed form. In this work, we adopt the concept of text summarization to capture the key essence from the text description and visual clues and then stitch them together. For most language-conditioned robotic works [10], [14], they generally need to retrain an extra language model based on a pre-trained CLIP or MiniLM model on their self-deigned task-oriented sentence dataset to achieve text summarization.

To enhance the efficiency and multimodal adaptability of the summarization process, we use prompt-based learning via GPT-4 to achieve contextual understanding and generate word tokens. As the most advanced language model, GPT-4 has a vast knowledge base and linguistic proficiency, allowing it to produce the concise summaries that humans want. An example of prompts in Fig. 5 shows that GPT-4 can learn to produce outputs tailored to our specific mapping tasks by providing prompts that are representative of the summarization task. To enable word embedding, we further use the strategy of label encoding to assign a unique integer to each class of labels in the generated word tokens.

D. Conditional Rearrangement Diffusion Network

The architecture of the proposed conditional rearrangement diffusion network is illustrated in Fig. 4. The transformer structure is employed as it is adept at fusing the information from different modalities. We first encode various scene object attributes and parsed word tokens into latent tokens, which are then processed by the multi-modal transformer. The network outputs translation and rotation predictions for each instance, and a diffusion scheme is adopted to successively refine the pose of each object. Below, we elaborate on each component of our network.

Token Encoder: The input tokens of the transformer include word embedding, mask embedding, and instance embedding. The word embedding represents the language instruction used to specify the target configuration. We map the parsed global structure and local regularity types to learned embeddings, which is conducive to identifying the commonalities of instructions faster during training compared to encoding the whole sentence with language models. Next, as defined in

Sec. III, the attributes of container objects and regular objects contain continuous variables such as translation, rotation, and size, as well as discrete variables like type and class. Similar to [8], we employ positional encodings of certain frequencies and subsequent linear layers to convert t_i, r_i, and s_i into vectors. As for discrete properties, a multilayer perceptron (MLP) is adopted to map one-hot vectors to high-dimensional latent. The above features are concatenated and then processed through an MLP to form the mask embedding and the instance embedding. This object-centric representation encodes each object separately, and 2 sets of specific MLPs are applied for mask and instance, respectively. Furthermore, a learned type embedding $T_{\mathcal{T}}$, which is utilized to distinguish different types of tokens ($\mathcal{W}$ord, $\mathcal{M}$ask, and $\mathcal{I}$nstance), is concatenated to the aforementioned embeddings as follows:

$$\hat{T}_{\mathcal{W},\mathcal{M},\mathcal{I}} = T_{\mathcal{W},\mathcal{M},\mathcal{I}} \oplus T_{\mathcal{T}}^{\mathcal{W},\mathcal{M},\mathcal{I}}, \tag{8}$$

where T and $\hat{T}$ represent the embedding of each modality and the final input token for the transformer, respectively.

Multi-Modal Transformer: We adopt the conventional encoder-only transformer architecture as the backbone. Our multi-modal transformer is a stack of several standard transformer blocks [38], consisting of the multi-head self-attention module and the position-wise feed-forward module. The self-attention mechanism helps the model enact the interactions between multiple objects, which allows it to be regarded as a fully connected graph structure. Besides, the language token and the mask token serve as conditional constraints and affect posture prediction through attention calculation. In the end, we build our network decoder as a two-layer MLP to output the denoised direction of t_i and r_i for each instance.

Efficiency-oriented Rearrangement Learning: To improve the interpretability of the network in our rearrangement task, we reparameterize the original score-based model s_θ to $\epsilon_\theta = \sigma_t^2 s_\theta$ as our multi-modal denoising diffusion model, indicating that the optimization of Eq. 5 evolves into a noise prediction problem. For the forward process during model training, we add sampled Gaussian noise with a specific standard deviation to the translation and rotation parameters of each object in the clean scene to formulate noisy S_t, which allows for the generation of various perturbed scenes with different levels of noise for one clean scene. After that, a reversed denoising process is learned by projecting S_t to the clean scene manifold via noise prediction.

As formulated in Eq. 1, we also want to minimize the motion distance between the initial messy scene and the rearranged clean scene. For most denoising diffusion works based on high-dimensional image space, it is typically presumed that the original image constitutes the nearest projection to its version perturbed by noise. However, each object instance's pose information in our task is low-dimensional data. This discrepancy suggests that with the introduction of different noise levels, the optimal projection target for a messy scene might not necessarily be consistent with the initially intended clean scene. Especially when applied to practical applications, such as food preparation or tabletop arrangement, the efficiency cost is enormous if persistently converting diverse messy scenes into the same specific layout.

Thus, we propose several techniques to ensure that the rearranged scene shares more similarities with the initial messy scene. First, we re-associate instances within the same class. Taking a language instruction as an example: *"Please rearrange all small boxes into a circle shape"*, we re-establish the pairing relationship p among all boxes between the current messy scene S and the target clean scene $\tilde{S}$ by computing their Earth Mover's Distance [39]:

$$\text{EMD} = \min_{p} \frac{1}{n} \sum_{i=1}^{n} \left\| t_i - \tilde{t}_{p(i)} \right\|_2^2, \tag{9}$$

where n is the number of instances in the scene, t and $\tilde{t}$ represent translation parameters of S and $\tilde{S}$, respectively. During training, we choose $\tilde{t}_{p^*(:)}$ with the optimal pairing relationship p^* instead of $\tilde{t}_{:}$ to construct $\tilde{S}$, which hopefully encourages a more efficient movement during rearrangement.

Second, as shown in Fig. 2, horizontal and vertical structures possess a certain degree of ambiguity. Drawing on the principles of least squares approximation from statistical analysis, we further propose to pan the clean scene along the relevant axis. This is to ensure that the average position of all instances in the optimal target scene $\tilde{S}^*$ aligns with the average position of all instances in the messy scene S. Through this procedure, we analytically guarantee minimal movement during the arrangement process, which can be formalized as:

$$\tilde{t_i^v}^* = \tilde{t_i^v} + \frac{1}{n} \sum_{i=1}^{n} \left(t_i^v - \tilde{t_i^v} \right) \text{ or } \tilde{t_i^h}^* = \tilde{t_i^h} + \frac{1}{n} \sum_{i=1}^{n} \left(t_i^h - \tilde{t_i^h} \right), \tag{10}$$

where t_i^v and t_i^h represent the coordinates of the vertical and the horizontal axis, respectively. We operate on the vertical axis for the horizontal structure and on the horizontal axis for the vertical structure.

Inference: During inference, we pursue the typical diffusion scheme shown in Eq. 4, where the learned $\epsilon_{\theta^*}(S_t)$ is asymptotically proportional to $\nabla_{S_t} \log q_0(S_t)$ as t declines. Given a messy scene S with attributes extracted, we treat it as S_T with a specific time step T and recursively predict the layout of the "cleaner" scene. We update the translation and rotation parameters of each instance and iterate continuously. α_t and β_t in Eq. 4 are designed to decrease as denoising progresses. The final attained S_0 is our rearrangement of the messy scene.

E. VLM Programming as Planner

Recently, much work [2], [40] from LLM-based robotic manipulation has demonstrated that language models have the potential to directly generate code snippets by parametrizing object states and the robot controller API. However, they all have to define the goal object states manually to finish different manipulation tasks. In this work, we use the conditional diffusion model to imagine different goal object states and integrate them as prompts into a multimodal vision-language model (VLM) to generate programming policy. In practice, we use the OpenAI GPT-4 model as our VLM cornerstone. Moreover, the generated output from the GPT model is expected to be valid Python code that covers programs from visual perception, language parsing, and denoising diffusion to robotic control.

```
import numpy as np
from vision_utils import get_obj_masks, PCA
from language_utils import LLM_parsing
from diffusion_utils import diffusion_pipeline
from robot_utils import pixel2pos, pick_and_place
```

Fig. 6. Statements of python APIs in our rearrangement task.

Robot Planning Example of GPT-4 Prompts

Scene Info:
You are an autonomous household robot activated in a domestic kitchen environment. Your primary location is by a table typically used for object arrangement and food preparation.

Now I give you the scene information called 'messy.png' and related task instructions. Please hierarchically utilize known API functions to compose robot planning process.

Prompt Info:

```
Example:
'messy.png', Rearrange an apple, a banana, a lemon into a plate container.
Output:
obj_names, obj_masks = get_obj_masks('messy.png')
obj_bboxes = PCA(obj_masks)
word_token, specified_objects = LLM_parsing('Rearrange … container.')
target_boxes = diffusion_pipeline(obj_bboxes, word_token)

##specified_objects = ['apple', 'banana', 'lemon', 'plate']
For j in range(len(specified_objects)-1):
    pick_pose = pixel2pos(obj_bboxes[specified_objects[j]])
    place_pose = pixel2pos(target_bboxes[specified_objects[j]])
    pick_and_place(pick_pose, place_pose)
```

Fig. 7. An example of GPT-4 prompts that generate a planning policy based on human language instruction and visual information of the messy scene when the robot executes the object rearrangement task.

Specifically, we first need to define our own Python function libraries that can inform the GPT model of which APIs are available and provide type hints on how to use these APIs. Fig. 6 shows all statements that can be imported into our rearrangement task. Furthermore, a few demonstration examples are used as prompts to instruct the GPT model to present contextual understanding and few-shot learning ability. Fig. 7 gives an example that directly outputs executable planning code comprising the capability to perform arithmetic, call API functions, and implement other Python language features. It can be seen that the GPT-4 model can well process multimodal inputs as instructions, then convert them into high-level perception features programmatically via vision, language, and denoising diffusion APIs, and finally call the low-level robot controller APIs to generate rearrangement actions. Owing to our denoising diffusion model being trained on a self-constructed synthetic dataset, it is usually challenging for traditional work to overcome the sim2real gap to implement our robotic arrangement task. However, by combining the open-vocabulary Grounded SAM model as a vision module, our GPT-based prompt-learning method can generalize to new objects and environments in real experimental scenarios well.

V. Experiments

A. Implementation Details

The token encoder produces 512-dimensional features as the input for the transformer, which has 2 layers with 8 heads of attention. The hidden layers of the transformer have 512 dimensions. We optimize our model on the proposed object rearrangement dataset, which contains 1640 clean scenes for training and 583 clean scenes for testing. We adopt the Adam optimizer with a base learning rate of 10^{-4}. The batch size is selected as 64 and the denoising model is trained on an A800 GPU for $30,000$ iterations, which takes about 3 hours. During inference, we choose to iterate 35 steps after considering both rearrangement efficiency and generative effectiveness, which takes seconds to rearrange a messy scene.

B. Evaluation Metrics & Baselines

To thoroughly estimate the performance of our proposed model in the simulated object rearrangement task, we utilize the following quantitative metrics:

Discrepancy between Results and the Ground Truth ($Dist2GT$): To measure the rearrangement quality, we compare the difference between the rearrangement result and the pre-perturbed scene. We compute the Earth Mover's Distance (EMD) between our rearranged configuration and the ground truth. We further calculate the cosine distance between the orientation of instances in the scene before and after rearranging consulting the new assignment from EMD. We report the average difference in position and orientation of scenes in the test dataset separately.

Distance Moved ($Movement$): To measure the rearrangement efficiency, we compute the average movement distance required for the scene to clean up. More specifically, we calculate the average Euclidean distance between the paired instances in the messy and rearranged layout for each scene. Then we report the mean value of all scenes in the test dataset. It is essential to consider the initial messy configuration and provide a solution that moves instances as little as possible to save time and energy for the robot.

Intersection over Union Threshold ($IOU_{threshold}$): In our work, we take the 2D oriented bounding box to represent the geometric attribute of the object. To quantitatively evaluate the arrangement accuracy, we compute the Intersection over Union (IoU) values between the predicted-target bounding box pairs for each instance. If the IoU value of arbitrary bounding box pairs is larger than a threshold δ (e.g. $\delta = 0.25, 0.5$), it is regarded as a success. Then we report the mean success rate of all rearranged objects in the test dataset.

In summary, $Dist2GT$ represents quality and alignment, $Movement$ shows the efficiency, and $IOU_{threshold}$ indicates success. For $Dist2GT$ and $Movement$, a lower value denotes a better performance of the generated results while a higher $IOU_{threshold}$ indicates a higher success rate. The metrics for real-world experiments will be introduced later.

Baselines: We reproduce 2 state-of-the-art baselines about object rearrangement on our dataset for comparison: 1) Struct-Diffusion [7], an object-centric and language-based iterative method that utilizes point clouds and instructions to learn global structures of object rearrangement. Unlike our approach, it introduces an extra time embedding in its diffusion framework to iterate from pure Gaussian noise without considering the initial messy configuration that the robot encounters. 2) LEGO-Net [8], a transformer-based data-driven method that learns to rearrange objects in messy rooms, where the concept of the moving distance of each object is first introduced. However, it lacks specialized designs for rich language conditions and semantic structures, and it cannot be directly applied to robot manipulation tasks. We reproduce these methods on our 2D rearrangement dataset, where training and evaluation splits remain the same as our method.

TABLE II
QUANTITATIVE COMPARISONS ON THE TASK OF REARRANGING INTO THREE KINDS OF GLOBAL STRUCTURES IN THE SIMULATION EXPERIMENTS.

Method	Horizonal					Vertical					Circle				
	$Dist2GT_T \downarrow$	$Dist2GT_R \downarrow$	$Movement \downarrow$	$IoU_{0.25} \uparrow$	$IoU_{0.5} \uparrow$	$Dist2GT_T \downarrow$	$Dist2GT_R \downarrow$	$Movement \downarrow$	$IoU_{0.25} \uparrow$	$IoU_{0.5} \uparrow$	$Dist2GT_T \downarrow$	$Dist2GT_R \downarrow$	$Movement \downarrow$	$IoU_{0.25} \uparrow$	$IoU_{0.5} \uparrow$
StructDiffusion [7]	0.308±0.004	0.071±0.010	0.481±0.004	13.5±0.7	3.3±0.5	0.249±0.004	0.016±0.003	0.448±0.006	21.0±1.0	6.2±0.6	0.248±0.003	0.011±0.002	0.429±0.013	23.8±0.8	10.3±0.4
LEGO-Net [8]	0.192±0.014	0.077±0.014	0.404±0.016	42.2±2.0	23.0±0.9	0.190±0.016	0.087±0.020	**0.394±0.009**	39.9±2.4	23.4±1.5	0.205±0.003	0.079±0.006	**0.375±0.008**	33.1±1.2	17.9±0.9
Ours	**0.103±0.002**	**0.005±0.002**	**0.397±0.012**	**51.8±0.7**	**28.5±1.2**	**0.109±0.002**	**0.001±0.000**	0.415±0.006	**53.0±0.9**	**27.6±1.8**	**0.153±0.002**	**0.001±0.000**	0.383±0.007	**43.4±0.7**	**21.6±0.5**

TABLE III
QUANTITATIVE COMPARISONS ON THE TASK OF PLACING INTO CONTAINERS IN THE SIMULATION EXPERIMENTS.

Method	$Dist2GT_T \downarrow$	$Dist2GT_R \downarrow$	$Movement \downarrow$	$IoU_{0.25} \uparrow$	$IoU_{0.5} \uparrow$
StructDiffusion [7]	0.106±0.001	0.002±0.000	0.479±0.011	19.7±1.0	7.4±0.4
LEGO-Net [8]	0.095±0.001	0.001±0.000	0.482±0.005	26.2±1.0	11.4±1.0
Ours	**0.085±0.001**	**0.001±0.000**	**0.458±0.024**	**37.1±2.4**	**17.8±1.4**

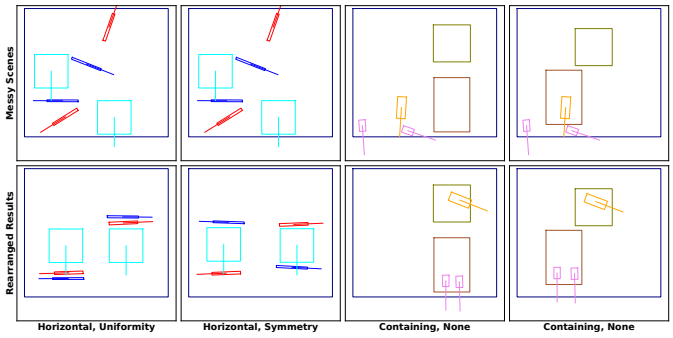

Fig. 8. Qualitative results from different rearrangement tasks: Food preparation (column 1 and column 2): knives, forks, and plates; Object containing (column 3 and column 4): cans and a banana. Our model can rearrange the same messy scene of both tasks following different instructions.

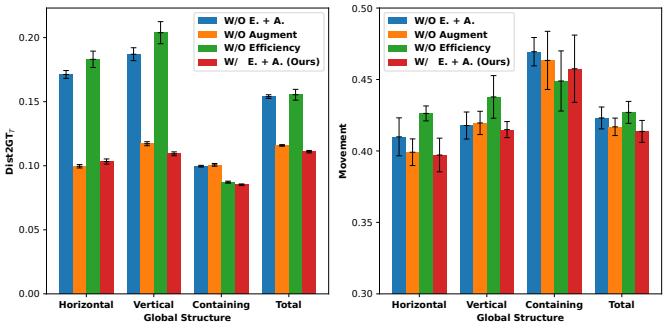

Fig. 9. Rearrangement results on different global structures. We compare our method with several variants under 2 metrics.

C. Simulation Experiments

Quantitative Results: We compare our method against the baselines mentioned above. We perturb clean scenes in the test set and rearrange these messy scenes with various methods. We conduct 5 replication experiments for each algorithm and report the average and confidence interval values on several metrics. The main results are presented in Tab. II and Tab. III, with the best results shown in **bold** and the inferior results within the confidence interval underlined. Our method is shown to outperform previous methods in most aspects. Due to StructDiffusion [7] starting denoising from pure noise, it ignores the initial configuration. Therefore, the rearrangement results obtained require a longer movement distance. As for LEGO-Net [8], it does not consider language conditions, thus causing uncertainty about achieving which kinds of structure. Our multi-modal transformer network increases the controllability of the rearrangement process, allowing for more precise implementation of various regularities. Moreover, the "container" object serves a distinct function compared to the regular objects being arranged. By introducing an extra "mask" object class and applying dedicated models to handle it, we achieve a better performance in the *containing* task, as evidenced in Table III.

Qualitative Results: Considering different human preferences from the same messy scene, we visualize some rearranged results of our model in Fig. 8. We use oriented bounding boxes to represent instances on the table, with different colors conveying different classes, whereas containers are depicted by directionless bounding boxes. For the same messy scene of each task, our method can rearrange it into different layouts according to different conditions. For instance, the final placements of forks (red) and knives (blue) for food preparation conform to the local regularities of *symmetry* and *uniformity* while the horizontal structure is also achieved. In addition, for the object containing task, our model can rearrange the objects from different categories into corresponding containers regardless of their location variations. One noteworthy point is that the misalignment of containers in column 4 has not appeared in the training split. To sum up, our model can learn how to leverage multi-modal conditional constraints for rearrangement, which makes our method applicable and generalizable to practical scenarios.

Ablation Study: As one of our contributions is to propose an efficiency-oriented rearrangement method, we further compare our method with "W/O Efficiency" that does not employ the operations in Eq. 10, "W/O Augment" that does not adopt data augmentation, and "W/O E. + A." that utilizes neither. These variants are evaluated on the test split across various global structures. As shown in the results of "Total" in Fig. 9, our method achieves a better overall performance than other variants in terms of both quality and efficiency of rearrangement. Especially for the *horizontal* and *vertical* structures, a significant improvement in the $Dist2GT_T$ metric can be seen owing to mitigating the ambiguity of the projection target. Besides, due to obtaining the optimal target on the clean scene manifold, we achieve a smaller motion distance in the $Movement$ metric. Moreover, the data augmentation operation can slightly improve the performance on various structures in both metrics, especially for the *containing* structure.

Analysis: Since we can perturb the clean scene with differ-

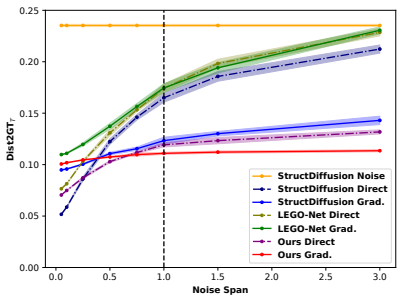

Fig. 10. Rearrangement results under different noise spans. Various baselines and inference strategies are compared. The black dashed line represents the noise span we adopted for training.

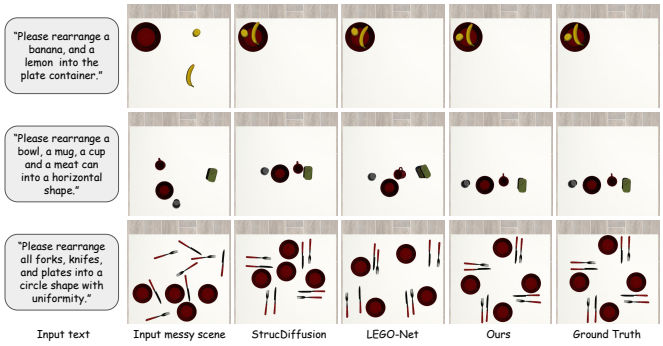

Fig. 11. Visualization results in simulation. We compare our method with state-of-the-art methods StructDiffusion [7] and LEGO-Net [8].

ent noise levels, we further investigate the model's denoising ability in dealing with different perturbing noises. Following [8], we use a noise span hyper-parameter σ to characterize the spectrum of σ_t, which originates from the positive half of $\mathcal{N}(0, \sigma^2)$. When we disturb scenes with σ_t derived from σ, the larger the σ value, the more likely the mess becomes severe. Meanwhile, as the trained denoising network ϵ_{θ^*} approximates the added noise $S_0 - S_t$, we can set $\alpha = 1$ and $\beta = 0$ in Eq. 4 to directly obtain S_0 and denote it as the *Direct* denoising strategy, distinguished from the standard *Gradual* denoising strategy. Based on the $Dist2GT_T$ metric, we evaluate StructDiffusion [7], LEGO-Net [8], and our method combined with these inference strategies on the test split. For StructDiffusion, we further adopt its original inference process starting from pure noises and name it *Noise*. As shown in Fig. 10, our *Gradual* recipe demonstrates the best denoising performance as the increment of noise spans, indicating that the proposed multi-modal transformer architecture can stably reconstruct a regular scene, even though the perturbation added to the scene is quite significant. Moreover, it can be seen that the *Direct* strategy exhibits a worse denoising ability than the *Gradual* strategy in handling high-noise scenes among all methods, possibly due to inaccurately estimated scores in low-density regions. The comparison results prove that iterative denoising is crucial for rearrangement, as it can gradually update data to high-density regions that possess more accurate estimates.

Visualization Comparison: In Fig. 11, we further visualize several comparison results of rearranged scenes in the physical simulator, where the dynamics of object collisions

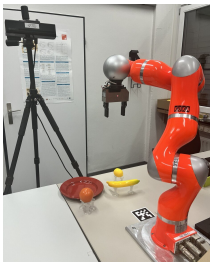

Fig. 12. Experimental setup for tabletop robotic rearrangement in the real world, consisting of the robotic arm, gripper, vision system, and messy scene.

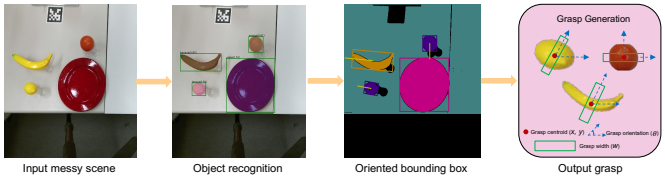

Fig. 13. The process of grasp generation for each object in the real robot experiments.

are accounted for. Given the special language command and the messy scene, it can be seen that among all scenes, our method can achieve the most precise arrangement of objects that conforms to human intentions while meeting the demand for efficiency. For example, in the second scene, the rearranged result from StrcutDiffusion [7] appears satisfactory, yet a more substantial movement for each object is needed compared to our rearranged result. In the case of LEGO-Net [8], a physical collision occurred between the mug and the meat can, leading to their dispersion across the table's surface.

D. Real-world Experiments

Experimental setup: To verify the proposed scheme in the real world, we also deploy a robotic experimental system shown in Fig. 12. The robotic manipulator selected for our setup consists of a 7-DOF KUKA LWR arm paired with a Schunk WSG50 gripper, which is mounted on the side of a table. The fixed point where the robot arm connects to the table is considered the base, with its centre position in the real-world coordinate system formulated as $[x, y, z] = [0.0, 0.0, 0.8]$. Our vision system incorporates the Kinect V2 in *qhd* mode to capture raw images. The *qhd* mode, while offering a wide field of view (FOV), also introduces the challenge of potentially detecting extraneous objects, such as camera fixtures and robotic equipment, as noise for the open-vocabulary Grounded SAM model. To mitigate this, we trim the raw image data to a uniform size of 448 pixels for both height and width. An AprilTag located on the table is used to calibrate the vision system, which will further facilitate the transformation of object pose from pixel coordinates to world coordinates. Finally, the process of generating grasps for the real robot experiment is illustrated in Fig. 13. To calculate the grasp on each object, we employ the antipodal method on its oriented bounding box. This involves determining the grasp point (x, y) and orientation θ based on the bounding box's average position and rotation value. The whole experimental system is operated

Fig. 14. All testing objects for robotic table rearrangement in the real world.

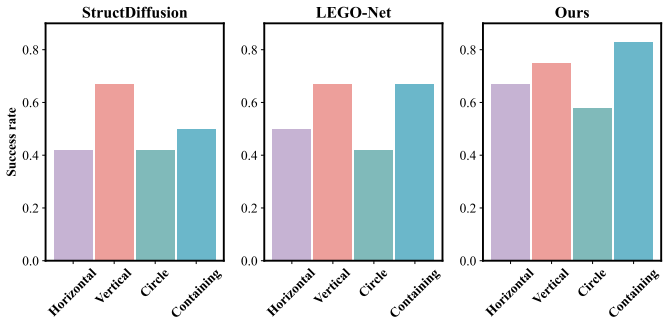

Fig. 15. Comparisons of the average success rate on various rearrangement structures for different methods in the real robot experiments.

TABLE IV
QUANTITATIVE COMPARISONS ON THE TASK OF REARRANGING INTO DIFFERENT STRUCTURES IN THE REAL ROBOT EXPERIMENTS.

Object structure	Horizontal	Vertical	Circle	Containing
Duration (s)	75.5	87.0	69.0	79.5
Collision-free rate (%) avg (std)	75.0 (16.7)	83.3 (8.33)	75.0 (8.3)	91.7 (8.33)
Success rate (%) avg (std)	66.7 (16.7)	75.0 (8.3)	58.3 (8.3)	83.3 (8.33)

by a ROS interface and the *pick_and_place* API of the robot in our VLM planner is achieved based on MoveIt!.

Robotic Rearrangement results: We conduct 12 evaluations for each task by altering the position and orientation of objects within a messy scene and arranging them according to a language-conditioned structure. For the structures defined as *horizontal*, *vertical*, and *circle*, the categories of objects in the messy scene include small boxes, toothpaste boxes, knives, forks, and spoons. As for the *containing* structure, the scene's objects consist of small boxes, box containers, plate containers, and various fruits. All objects utilized in our robotic experiments are displayed in Fig. 14.

In Tab. IV, we first compare the average rearrangement duration, collision-free rate, and final successful rate for different structures in the real robot experiment. The collision-free rate is calculated by observing whether all objects in the rearrangement scene predicted by our denoising diffusion model are collision-free. Additionally, the success rate is assessed after the robot finishes each rearrangement task. Unlike the abstract estimation metrics in simulation, a real-world rearrangement is considered successful only if the objects are positioned without any collisions and the overall structure adheres to the semantic constraints set forth by the provided language instructions. It can be seen that the *horizontal*, *vertical*, and *circle* configurations present significant challenges due to the entirely novel nature of various boxes in our dataset and the limited tolerant positions for sequential placement. The requirement to rearrange the same categories of objects into these three distinct structures simultaneously further complicates the task for our inference model. Additionally, we encounter failures when the vision system struggles to accurately perceive the depth of particular objects, such as spoons, knives, forks, and bananas, due to their reflective surfaces. This incorrect depth data prevents the robot's gripper from achieving a stable grasp.

Furthermore, we compare the average success rate with baselines on four kinds of language-conditioned structures, shown in Fig. 15. Our method outperforms other baselines for all global structures, with an average improvement of 15% on the success rate compared to LEGO-Net [8]. We also find that the real-world rearrangement becomes more challenging when more objects are added to the initial messy scene. This may be attributed to the struggle of diffusion models to learn a more complicated inter-object relationship, along with the increase in robot planning and execution horizons.

VI. CONCLUSION AND DISCUSSION

We present a solution for the language-conditioned robotic rearrangement task with different global structures and local regularities. Firstly, we collect and process a 2D synthetic arrangement dataset based on the physical simulator. To capture long-range dependencies between visual and textural inputs, we build our conditional diffusion model based on the multi-modal transformer architecture, which endows the robot with the ability to imagine the target pose information of different objects from the observation scene. In particular, we introduce an efficient-oriented rearrangement learning strategy to reduce object motion distance and create a more appropriate layout. Inspired by the recent prompt-based learning, we further integrate the generative model into the most advanced VLM module (GPT-4) to generate robot planning and action policy. Finally, we carefully design three kinds of quantitative metrics to evaluate our model in the simulation experiments, showing that our generative model outperforms related state-of-the-art methods. Extensive experiments on the real robot further demonstrate that our proposed scheme can satisfy the human language-based requirements and finish different rearrangement tasks successfully on diverse unseen objects.

Concerning limitations, this work simplifies the inter-object relationship in a clean rearrangement scene based on human preference, with the global structure and the local regularity limited to 4 and 2, respectively. In addition, when dealing with more complex object interactions, our approach tends to rearrange objects into a simpler layout composed of fewer objects by overlapping some instances. Therefore, extending our research to include and comprehend the inter-object relationship on a larger scale is important. Moreover, since we use prompt-based learning to achieve language instruction parsing and robot action policy generation, we still need to

pre-define a series of examples to instruct the VLM module to interpret the prompt. As for future work, the introduction of explicit collision avoidance mechanisms in the denoising process can be explored, which may make the generated layout more plausible. Besides, we currently use the same number of inference steps, and we wonder whether it is possible to determine a more accurate number of denoising steps by assessing the level of mess in the present scene, which may enhance efficiency. Finally, designing an end-to-end VLM model to estimate robot actions directly rather than separating the dreaming and planning process may further improve the effectiveness of robot manipulation.

REFERENCES

[1] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi *et al.*, "Rearrangement: A challenge for embodied ai," *arXiv preprint arXiv:2011.01975*, 2020.

[2] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[3] A. Irpan, A. Herzog, A. T. Toshev, A. Zeng, A. Brohan, B. A. Ichter, B. David, C. Parada, C. Finn, C. Tan *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*, no. 2022, 2022.

[4] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[5] G. Zhai, X. Cai, D. Huang, Y. Di, F. Manhardt, F. Tombari, N. Navab, and B. Busam, "Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs," *arXiv preprint arXiv:2309.12188*, 2023.

[6] I. Kapelyukh, V. Vosylius, and E. Johns, "Dall-e-bot: Introducing web-scale diffusion models to robotics," *IEEE Robotics and Automation Letters*, 2023.

[7] W. Liu, T. Hermans, S. Chernova, and C. Paxton, "Structdiffusion: Object-centric diffusion for semantic rearrangement of novel objects," *arXiv preprint arXiv:2211.04604*, 2022.

[8] Q. A. Wei, S. Ding, J. J. Park, R. Sajnani, A. Poulenard, S. Sridhar, and L. Guibas, "Lego-net: Learning regular rearrangements of objects in rooms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 037–19 047.

[9] J. Tang, Y. Nie, L. Markhasin, A. Dai, J. Thies, and M. Nießner, "Diffuscene: Scene graph denoising diffusion probabilistic model for generative indoor scene synthesis," *arXiv preprint arXiv:2303.14207*, 2023.

[10] W. Liu, C. Paxton, T. Hermans, and D. Fox, "Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6322–6329.

[11] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[13] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.

[14] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 576–11 582.

[15] Y. Jiang, S. S. Gu, K. P. Murphy, and C. Finn, "Language as an abstraction for hierarchical deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] D. Misra, J. Langford, and Y. Artzi, "Mapping instructions and visual observations to actions with reinforcement learning," *arXiv preprint arXiv:1704.08795*, 2017.

[17] C. Lynch and P. Sermanet, "Language conditioned imitation learning over unstructured data," *arXiv preprint arXiv:2005.07648*, 2020.

[18] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor, "Language-conditioned imitation learning for robot manipulation tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 139–13 150, 2020.

[19] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9118–9147.

[20] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[21] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[22] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[23] W. Chen, C. Zeng, H. Liang, F. Sun, and J. Zhang, "Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly," *IEEE Transactions on Cybernetics*, 2023.

[24] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[25] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

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

[28] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[29] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[30] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," *arXiv preprint arXiv:2210.11427*, 2022.

[31] C. Kong, D. Jeon, O. Kwon, and N. Kwak, "Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 848–857.

[32] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218.

[33] L. Tang, H. Liu, H. Huang, X. Xie, N. Liu, and M. Li, "A reinforcement learning method for rearranging scattered irregular objects inside a crate," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[34] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," *arXiv*, 2022.

[35] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems, NeurIPS*, 2019, pp. 11 895–11 907.

[36] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Comput.*, vol. 23, no. 7, pp. 1661–1674, 2011.

[37] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems, NeurIPS*, vol. 30, 2017.

[39] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[40] Y. Jin, D. Li, A. Yong, J. Shi, P. Hao, F. Sun, J. Zhang, and B. Fang, "Robotgpt: Robot manipulation learning from chatgpt," *IEEE Robotics and Automation Letters*, 2024.