Synergistic Dual Spatial-aware Generation of Image-to-Text and Text-to-Image

Yu Zhao[1], **Hao Fei**[2]*, **Xiangtai Li**[3], **Libo Qin**[4], **Jiayi Ji**[2],
Hongyuan Zhu[5], **Meishan Zhang**[6], **Min Zhang**[6], **Jianguo Wei**[1]

[1] Tianjin University [2] National University of Singapore [3] Bytedance [4] Central South University
[5] I^{2}R & CFAR, A*STAR[†] [6] Harbin Institute of Technology (Shenzhen)
zhaoyucs@tju.edu.cn, haofei37@nus.edu.sg

Abstract

In the visual spatial understanding (VSU) area, spatial image-to-text (SI2T) and spatial text-to-image (ST2I) are two fundamental tasks that appear in dual form. Existing methods for standalone SI2T or ST2I perform imperfectly in spatial understanding, due to the difficulty of 3D-wise spatial feature modeling. In this work, we consider modeling the SI2T and ST2I together under a dual learning framework. During the dual framework, we then propose to represent the 3D spatial scene features with a novel 3D scene graph (3DSG) representation that can be shared and beneficial to both tasks. Further, inspired by the intuition that the easier 3D→image and 3D→text processes also exist symmetrically in the ST2I and SI2T, respectively, we propose the Spatial Dual Discrete Diffusion (SD^3) framework, which utilizes the intermediate features of the 3D→X processes to guide the hard X→3D processes, such that the overall ST2I and SI2T will benefit each other. On the visual spatial understanding dataset VSD, our system outperforms the mainstream T2I and I2T methods significantly. Further in-depth analysis reveals how our dual learning strategy advances.

1 Introduction

Within the research topic of Visual Spatial Understanding (VSU) [42, 95, 74, 89, 8], Spatial Image-to-Text (SI2T) [95, 97] and Spatial Text-to-Image (ST2I) [58] are two representative task forms across vision and language. SI2T aims to understand the spatial relationships of objects in the given image, while ST2I focuses on synthesizing a spatial-faithful image based on the input text prompts. Existing efforts mostly formalize SI2T and ST2I tasks as the normal I2T and T2I problems, applying general-purpose I2T and T2I models for task solutions [95, 97, 58, 37, 38]. Technically, researchers widely employ the vision-language generative architecture [79, 10, 83] for I2T tasks, i.e., generally, there is a visual encoder and a text decoder. For the T2I task, recent diffusion-based methods have shown the extraordinary capability of image generation and achieved state-of-the-art (SoTA) performance on a mount of benchmarks [4, 12, 28, 71, 63, 86, 18, 11].

Unfortunately, these strong-performing I2T and T2I methods, without the deliberate spatial semantics modeling in VSU tasks, largely fall short in precisely extracting spatial features from visual or textual inputs. For instance, in the SI2T process, the spatial relationships are often incorrectly recognized due to layout overlap and perspective illusion [97]. This is due to the inherent characteristics of 2D images, which, lacking 3D feature modeling, inevitably fail to understand spatial relations. On the other hand, in ST2I, synthetic images frequently fail to match strictly the spatial constraints specified

---

*Corresponding Author: Hao Fei
[†]Institute for Infocomm Research (I^{2}R) & Centre for Frontier AI Research (CFAR), A*STAR

in the prompts, such as the position, pose, and perspective [58]. This is because language's abstract nature allows only for a general description of the content rather than detailed depictions of spatial scenes. Moreover, unlike general image generation, ST2I should place greater emphasis on 3D spatial modeling, while the input sequential language tokens intrinsically do not portray the kind of specific spatial scene.

Let us revisit the human process in solving SI2T and ST2I tasks. We typically process the input image or text within our minds by constructing a 3D spatial scene, which serves as the center for further spatial textual descriptions or image generation. Specifically for SI2T, we intuitively project a 2D image into a reasonable 3D scene based on common sense before describing that scene in words. In contrast, for ST2I tasks, we start by imaginatively converting the text of user instruction into a 3D conceptual scene which is then rendered into a 2D image. Interestingly, we can actually find that SI2T and ST2I are dual processes, with each task's input and output being the reverse of the other, as illustrated in Figure 1(a). More importantly, there are two important observations in these dual tasks. **First, [*intermediate processing sharing*]**, they can complement and benefit each other. For SI2T, the 'Image→3D' reasoning process is challenging in acquiring necessary 3D features, whereas the descriptive '3D→Text' process is relatively easier. Conversely, for ST2I, the 'Text→3D' process requires complex reasoning of the 3D scene feature, while rendering '3D→Image' is much more straightforward. Ideally, if complementing the information during each learning process, i.e., letting the easy part aid the hard part, it should

(a) Complementary dual tasks of ST2I and SI2T.

(b) Diffusion Based 3DSG Generation.

Figure 1: Demonstration of SI2T and ST2I tasks.

enhance the performance of both tasks. **Second, [*3D scene feature sharing*]**, both dual tasks urgently require modeling of the respective 3D features, where such stereospecific insights in 3D perspective can be essentially shared and also complementary between each other. Intuitively, vision offers very concrete clues describing spatial scenes, e.g., objects and attributes, while the 3D features derived from texts are more likely to define the semantics-oriented relations or constraints, which are often abstract and rough.

Inspired by such intuition, in this paper, we introduce a novel synergistic dual framework for SI2T and ST2I. To start with, we propose a spatial-aware 3D scene graph (namely **3DSG**), where the spatial objects, their relations and the layouts within the 3D scene are formulated with a semantically structured representation. The 3DSG representation effectively depicts the stereospecific attributes of all objects, and meanwhile models the spatial relations between them, from which both the SI2T and ST2I processes can be beneficial. Technically, 3DSG is obtained from a shared graph diffusion model for both SI2T and ST2I processes, as shown in Figure 1(b). Trained with our '2DSG-3DSG' pair data, the graph diffusion is learned to propagate and evolve the initial 2D visual SG (parsed from input image at SI2T side) or the textual SG (parsed from input text at ST2I side) into the final 3DSG, i.e., by adding all necessary and reasonable spatial details.

With *3DSG*, we next implement the whole dual framework of SI2T and ST2I generation. Instead of directly employing the SoTA generative I2T models or diffusion-based T2I methods, we consider a solution fully based on discrete diffusions [3], due to several key rationales. Primarily, for VSU, the most crucial spatial information that determines a scene consists of objects and their relationships, which presents the characteristic of discretization and combination in the spatial layout, while other background and spatial unrelated information would be noisy. Thus, the discrete representation is more appropriate to model pure spatial semantics in our scenario. Besides, in our scenario both the textual token and SG representations possess discrete characteristics [54], which can be perfectly modeled by discrete diffusion. Moreover, the discrete diffusion works on the limited index space
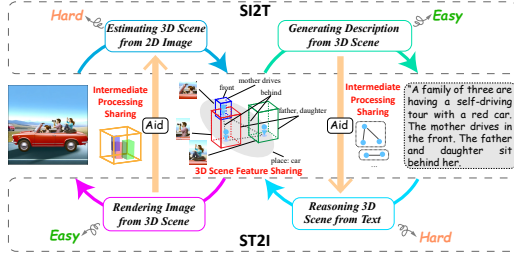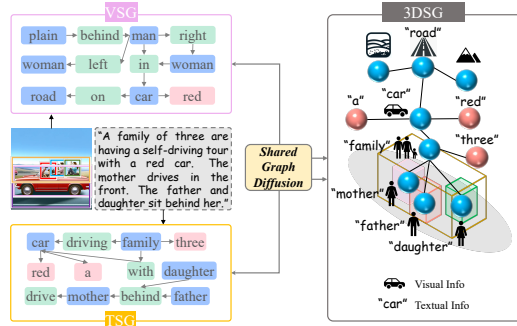
2

[28, 54, 35, 99], which is much more computationally efficient, especially for visual synthesis tasks. As illustrated in Figure 2, our **S**patial **D**ual **D**iscrete **D**iffusion (dubbed as **SD^3**) system consists of three components: 1) *3DSG* generation, 2) SI2T generation and 3) ST2I generation. During dual learning, both SI2T and ST2I first induce the *3DSG* representations from the input image or text respectively via one shared graph diffusion, where the modality-variant features are simultaneously preserved in *3DSG* for unbiased and holistic modeling of 3D spatial scene. Then, the image synthesis (for ST2I) and text generation (SI2T) are carried out via two separate discrete diffusion models, during which the *3DSG* feature is integrated for better generation. At the meanwhile, the intermediate features of the '3D→X' (X means text or image) diffusion steps are also passed to the counterpart hard 'X→3D' processes for further facilitation.

We conduct experiments on the VSD dataset [95], which is a VSU benchmark with paired images and texts of spatial descriptions that allow for both SI2T and ST2I. Extensive results demonstrate that our proposed system significantly outperforms all baselines on both ST2I and SI2T, including the vision-language model based and diffusion-based methods. Further analysis reveals that the dual framework helps align the asymmetric spatial semantics across image and text modalities. To our knowledge, this is the first attempt to resolve the SI2T and ST2I tasks with a novel dual perspective, and further successfully investigate synergistic learning in between by sufficiently modeling the spatial 3D scene representations.

2 Related Work

Visual Spatial Understanding. VSU is an important topic within the research of multimodal learning [51, 17, 19, 21]. VSU aims to extract spatial information from a given scene, developed within the forms of reasoning [48], relation extraction [56], role labeling [42], question answering [44, 61, 52], image-to-text generation [95, 97], image synthesis [58], 3D reconstruction [74, 69], etc. With the VSU capability, many downstream applications can achieve, such as robotics [24, 53], navigation [33, 80, 6], and language grounding [49]. Among various VSU tasks, the SI2T and ST2I generation attract significant attention due to their fundamental positions in vision-language cross-modal tasks. Current efforts mostly study ST2I or SI2T separately. For the text-to-image generation, the diffusion models have emerged as the SoTA approaches [15]. For image-to-text generation, i.e., image captioning, it has been a long-standing task and has achieved great progress. Recent advances can be largely attributed to vision-language pre-training (VLP) [79, 10]. Besides conventional VLP methods, [99] are the first to use diffusion models for image captioning (and more specifically, for text generation). In this work, we consider the two tasks together under a dual learning framework, via which we aim to achieve mutual benefits of the spatial feature modeling from each other.

Discrete Diffusion Models. The diffusion model is first proposed by [32], and has achieved impressive performance for text-to-image generation [63, 12, 71, 65, 86]. Original diffusion models are parameterized Markov chains trained to translate simple distributions to more sophisticated target distributions in a finite set of steps on the continuous data or its latent representations. In this work we adopt the SoTA diffusion-based model as our T2I backbone. Recently, diffusion models on discrete space are introduced to markedly reduce the computing cost [28, 3, 34, 35]. Discrete diffusion methods are also used in text generation [34] and structure generation [54] due to its natural adaptability for the data with discrete nature. This work follows the line and takes the SoTA discrete diffusion model as the backbone for image, text and scene graph generation.

Scene Graph Representations. This work is also closely related to scene graph (SG)-based representation learning. Scene graphs have advanced in depicting the intrinsic semantic structures of scenes in images or texts [43, 81, 93]. In SGs, key object and attribute nodes are connected via pairwise relations to describe semantic contexts, which have been shown to be useful as auxiliary features that carry rich contextual and semantic information for a wide range of downstream applications, such as image retrieval [39], image generation [40, 86], translation [17], image captioning [90, 83] and video modeling [96, 18]. However, research on utilizing 3D scene graph representations to enhance various downstream tasks can be currently very scarce. In this paper, we incorporate both visual and language scene graphs within a 3D scope to enhance cross-modal alignment learning for better spatial semantics understanding.

Dual Learning. The dual learning method is proposed to enhance the coupled tasks that have the same exact input and output but in reverse [88, 91]. With dual bidirectional learning, the model could capture potential mutual information from the primary and dual tasks, improving their performance.

Recently, the idea of dual learning has been applied to various tasks, such as intertranslation [88], speech recognition with text-to-speech [77], question answering with question generation [76, 75], and image classification with image generation [16]. The key point of dual learning is to model the paired and complementary features between the coupled tasks, and then the learning process can reinforce them mutually. In this work, we first connect the dual SI2T and ST2I tasks with their shared 3D spatial feature and use dual learning to enhance 3D feature construction. To our knowledge, we are the first to achieve synergy between two spatial-aware cross-modal dual generations.

3 Preliminaries

3.1 Task Formulation

We process the dual visual spatial understanding tasks, ST2I and SI2T. Given a textual prompt Y, ST2I aims to generate an image $\hat{I}$ that semantically matches the spatial constraints with Y. Its dual task is SI2T, which is also known as the visual spatial description (VSD), aiming to generate a piece of textual description $\hat{Y}$ based on an input image I. In this paper, we process the two tasks parallelly.

3.2 Discrete Diffusion

Diffusion models [32] are generative models characterized by a forward and reverse Markov process. In the forward process, the given data x_0 with distribution $q(x_0)$ is corrupted into a Gaussian distribution variable x_T in T steps, formulated as $q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$. In the reverse process, the model learns to recover the original data x_0 from x_T, denoted as $p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$. In order to optimize the generative model $p_\theta(x_0)$ to fit the data distribution $q(x_0)$, one typically optimizes a variational upper bound on the negative log-likelihood:

$$\mathcal{L}_{vlb} = \mathbb{E}_{q(x_0)} \left[D_{KL} \left[q(x_T|x_0)||p(x_T) \right] + \sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \left[D_{KL} \left[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t) \right] \right] \right]. \quad (1)$$

Vanilla diffusion models are defined on continuous space. Recently, the discrete diffusion model has been introduced where a transition probability matrix Q_t is defined to indicate how x_0 transits to x_t for each step of the forward process, where $x_t \in \mathbb{Z}^N$ is defined in discrete space. The matrices $[Q_t]_{ij} = q(x_t = i|x_{t-1} = j)$ defines the probabilities that x_{t-1} transits to x_t. Then the forward and reverse process could be rewritten as $q(x_t|x_{t-1}) = v^\top(x_t)Q_t v(x_{t-1})$ and $q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)/q(x_t|x_0)$, where $v(x)$ means the one-hot representation of x. Appendix A.1 gives more technical details about the discrete diffusion models.

3.3 Scene Graph Representation

The scene graph presents a scene via the objects with their attributes and relationships in the form of a graph, which can be constructed from either text (TSG) [68] or image (VSG) [92, 46, 45]. We represent a scene graph as $G = \{V, E\}$, where V is the node set, and E is the edge set. There are three types of nodes in the scene graph: object nodes, attribute nodes, and relationship nodes. Each type of node has ts own unique tag vocabulary.

For SI2T and ST2I, we focus on spatial information extraction, especially the object relationships in the 3D space. However, due to the limited information presented by text or image, a 2D TSG or VSG hardly fully models these 3D spatial semantics. Here, we introduce the

Table 1: Comparison of 3DSG and 2DSG.

2D TSG
• **Objects:** Only key objects
• **Attributes:** Any attributes
• **Relationships:** Semantic relationships between objects

2D VSG
• **Objects:** The front objects in one perspective
• **Attributes:** Visual related attributes
• **Relationships:** Simple 2D spatial relationships

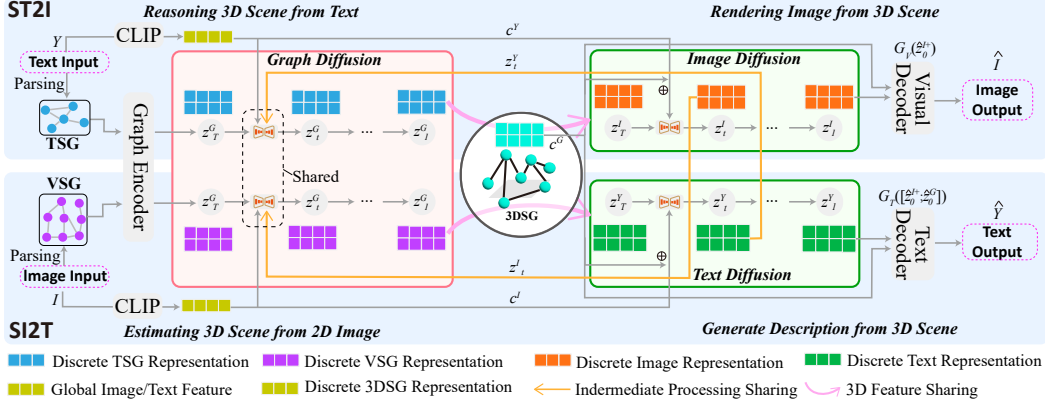| **3DSG** |
| --- |
| • **Objects:** All the objects in the 3D scene; High-level spatial concepts. |
| • **Attributes:** Any attributes including 3D attributes such as pose and texture. |
| • **Relationships:** 3D spatial relationships and subordinate relationships between spatial concepts. |

spatial-aware 3D scene graph (3DSG) [41], which thoroughly depicts 3D scenes. The 3DSG is formally equivalent to the 2D scene graph, i.e. the object, attribute, and relationship, but organized

Figure 2: Overall Framework of the S³D. The figure presents the dual processes of ST2I and SI2T. The RED block represents the hard X→3D processes, and the GREEN block represents the 3D→X processes. There are three diffusion processes in total, i.e., a shared graph diffusion model for VSG/TSG→3DSG generation, and the image diffusion model and text diffusion model.

hierarchically, which contains more abundant elements such as nodes for high-level spatial concepts, texture attributes, and spatial relationships in the 3D perception. Table 1 presents the differences between 2DSG and our 3DSG. Conventionally, the 3DSG should be constructed directly from a 3D scene, i.e., sensor data [64], point clouds [78], 3D mesh [2], and RGB-D sequences [87, 26]. In this work, we uniform the VSG/TSG/3DSG representations, modeling the object nodes v_{obj}, attribute nodes $\boldsymbol{v}_{attr}$, and relationship nodes $\boldsymbol{v}_{rel}$ by the embedding of their textual tags $\boldsymbol{e}_{obj}, \boldsymbol{e}_{attr}, \boldsymbol{e}_{rel} \in \mathbb{R}^d$, where d is the dimension of the tag embedding.

4 Methodology

4.1 Spatial Dual Discrete Diffusion Framework

In Figure 2, we illustrate the overall framework of the proposed Spatial Dual Discrete Diffusion (SD³), consisting of three separate discrete context diffusion models, i.e., the 3DSG diffusion model for the **X→3D** process, the ST2I diffusion model for the **3D→Image** process, and the SI2T diffusion model for the **3D→Text** process.

X→3D: 3DSG Diffusion Model. We set a graph diffusion model to convert the initial TSG (for ST2I) and VSG (for SI2T) to the 3DSG. We first acquire the initial VSG and TSG from the input image and textual description through the off-the-shelf VSG [68] and TSG [92] parsers. Then the initial VSG and TSG nodes (represented as $\{\boldsymbol{e}_{obj}, \boldsymbol{e}_{attr}, \boldsymbol{e}_{rel}\}$) are encoded to the latent quantized representations by a discrete graph auto-encoder (DGAE) [54], on which the discrete diffusion process on discrete graph representations is conducted to generate the 3DSG. Concretely, we denote the quantized gold 3DSG as $\boldsymbol{z}_0^G$. Then, following conventional discrete diffusion, we can calculate $p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G)$. Moreover, on our final dual training, the 3DSG diffusion model leverages the intermediate feature of the following **3D→X** process to aid the **TSG→3DSG** and **VSG→ 3DSG** generation process. For **TSG→3DSG**, the model takes global text features and the intermediate feature of the dual SI2T diffusion, i.e., $p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{c}^Y)$, $p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{z}_t^Y)$, and for **VSG→3DSG** it will be $p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{c}^I)$, $p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{z}_t^I)$, where the $\boldsymbol{z}_t^Y$ and $\boldsymbol{z}_t^I$ are the t step representations of SI2T and ST2I diffusion respectively. Then we align them by:

$$\mathcal{L}_{X-T23D} = D_{KL}(p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{c}^Y) || p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{z}_t^Y)), \qquad (2)$$

$$\mathcal{L}_{X-I23D} = D_{KL}(p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{c}^I) || p_\psi^{Graph}(\boldsymbol{z}_{t-1}^G | \boldsymbol{z}_t^G, \boldsymbol{z}_t^I)). \qquad (3)$$

To enable the training for this 2DSG→3DSG process, we should construct the paired '2DSG-3DSG' data. We follow previous work [72, 78, 27] to construct a gold 3DSG dataset. After that, we adopt a strong captioning model to generate text descriptions from view images. With all the inputs and

outputs ready, we can train the graph diffusion model on aligned 3DSG-Image-Text data, after which the model will acquire the 3D estimation capability.

3D→Image: ST2I Diffusion Model. The ST2I diffusion model is used to generate image $\hat{I}$ with the condition of input textual prompts Y. In this task, we adopt a vector quantized variational autoencoder (VQ-VAE) [14] as the quantized model to encode image data to embedding vectors. We denote the quantized image as z_0^I and we could also calculate $q(z_t^I|z_{t-1}^I)$ and $q(z_{t-1}^I|z_t^I, z_0^I)$. On the other hand, the ST2I diffusion model takes two conditions, i.e., the global text feature c^Y extracted via a CLIP model [60], and the shared 3DSG feature c^G. Then c^Y and c^G are fused and fed to the diffusion model to calculate $p_\theta^{ST2I}(z_{t-1}^I|z_t^I, c^T \oplus c^G)$. At last, the decoder from VQ-VAE is used to generate images from latent codes.

3D→Text: SI2T Diffusion Model. In the SI2T model, the inputs and outputs are reversed. The diffusion is applied on the text representation z_0^Y with a similar process as ST2I. The latent text representation is the word embedding, which is functionally equivalent to the visual codebook. We use another denoising network $p_\phi^{SI2T}(z_{t-1}^Y|z_t^Y, c^I \oplus c^G)$, while the input condition becomes the visual tokens c^I and also the 3DSG feature c^G. Afterward, the language model decoder generates textual results based on the latent representations.

Image and Text Decoding. Finally, through three diffusion models, we acquire the predicted latent representations of image, text, and 3DSG, denoted as $\hat{z}_0^I, \hat{z}_0^Y, \hat{z}_0^G$. For ST2I, we fuse the generated graph vectors $\hat{z}_0^G$ to visual vectors $\hat{z}_0^I$ by adopting an attention mechanism:

$$Attn_{i,m} = \underset{m}{softmax}(\hat{z}_0^G[m] \cdot \hat{z}_0^I[i]), \quad \hat{z}_0^{I+}[i] = \hat{z}_0^I[i]) \oplus \sum_m Attn_{i,m} \cdot \hat{z}_0^G[m], \tag{4}$$

where $\oplus$ means the concatenation operation. Afterwards, the image decoder G_V is used to reconstruct image from the scene graph enhanced codes $\hat{z}^{I+}$:

$$\hat{I} = G_V(\hat{z}^{I+}). \tag{5}$$

For SI2T, we append the flatten $\hat{z}_0^G$ after generated textual codes $\hat{z}_0^Y$ and use a language model decoder G_T to generate the description:

$$\hat{Y} = G_T([\hat{z}_0^Y; \hat{z}_0^G]). \tag{6}$$

4.2 Mutual Spatial Synergistic Dual Learning

To exploit the complementarity of the dual processes, we elaborate a dual training strategy for our framework. To conduct the strategy effectively, we introduce the essential training objectives.

Dual Learning Objective. Our dual learning framework contains two tasks, i.e. ST2I and SI2T, where their inputs and outputs are just reversed. We denote ST2I as $f_\theta : i \rightarrow y, i \in I, y \in Y$ and SI2T as $f_\phi : y \rightarrow i, i \in I, y \in Y$. Their learning objectives should be:

$$\mathcal{L}_{ST2I} = \mathbb{E}_{i,y} \log p_\theta(y|i), \quad \mathcal{L}_{ST2I} = \mathbb{E}_{i,y} \log p_\phi(y|i). \tag{7}$$

Based on the dual supervising learning [88], given the duality of the two tasks, if the learned ST2I and SI2T models are perfect, we should have the probabilistic duality:

$$p(i)p_\theta(y|i) = p(y)p_\phi(i|y) = p(i,y), \forall i, y, \tag{8}$$

where $p(i)$ and $p(y)$ are the marginal distributions. Then we add this constraint to the dual learning target as an equivalent regularization term:

$$\mathcal{L}_{dual} = \mathcal{L}_{ST2I} + \mathcal{L}_{SI2T} + || \log \hat{p}(i) + \log p_\theta(y|i) - \log \hat{p}(y) - \log p_\phi(i|y)||, \tag{9}$$

where $\hat{p}(i)$ and $\log \hat{p}(y)$ are the estimated marginal distribution by a pre-trained vision-language model, which is illustrated detailly in Appendix §A.4.

Loss for Mutual 3D Feature Learning. The 3DSG diffusion model is used to estimating 3D scene information from TSG or VSG, denoted as $f_\psi : i \rightarrow g$ and $f_\psi : y \rightarrow g$, where $i \in I, y \in Y, g \in G$. Based §4.1 and Eq. 1, the graph diffusion loss can be acquired, denoted as $\mathcal{L}_{I23D}$ and $\mathcal{L}_{T23D}$

$$\mathcal{L}_{I23D} = \mathbb{E}_{i,y} \log p_\theta(g|i), \quad \mathcal{L}_{T23D} = \mathbb{E}_{i,y} \log p_\phi(g|y), \tag{10}$$

where z^I and z^Y are the intermediate features of image diffusion and text diffusion. Along with the X→3D alignment loss $\mathcal{L}_{X-T23D}$ and $\mathcal{L}_{X-I23D}$, the final mutual 3D Feature learning target is:

$$\mathcal{L}_{mutual} = \mathcal{L}_{I23D} + \mathcal{L}_{X-I23D} + \mathcal{L}_{T23D} + \mathcal{L}_{X-T23D}. \tag{11}$$

Loss for Spatial Feature Alignment. The paired latent visual and textual representations are initialized in different feature spaces. We thus conduct a spatial feature alignment to bridge them by the shared spatial representation of the 3DSG. Concretely, we consider the image decoder, text decoder and graph encoder. Given the paired image I, text Y and the corresponding 3DSG G_{host}, we can easily get their quantified representations z_0^I, z_0^Y, z_0^G, and then the fused feature z_0^{I+} and $[z_0^Y; z_0^G]$ based on the Eq. 5 and Eq. 6. After that, we adopt the reconstruct loss of VQ-VAE $\mathcal{L}_{v-dec}$, and the next token prediction loss $\mathcal{L}_{t-dec}$ of text decoder for optimization:

$$\mathcal{L}_{v-dec} = ||I - G_V(z_0^{I+})||^2, \quad \mathcal{L}_{t-dec} = -\sum_{i=1}^{|y|} \log p(y_i|y_{<i}, [z_0^Y; z_0^G]). \tag{12}$$

Loss for 3DSG Reconstruction. To initialize the graph encoder and decoder, i.e., the DGAE, we follow [5] to calculate the graph reconstruction loss:

$$\mathcal{L}_{DGAE} = -\mathbb{E}_{\hat{\mathcal{Z}}^G} \ln(p(G_{host}|\hat{\mathcal{Z}}^G)), \tag{13}$$

where $\hat{\mathcal{Z}}^G$ is the latent representations of the predicted 3DSG and the G_{host} is the gold 3DSG.

Training Remarks. To achieve a perfect convergency, we use a four-step strategy for the overall training process, shown in Figure 3. **Step-1 DGAE pre-training.** First, we pre-train the DGAE model with the 3DSG reconstruction target $\mathcal{L}_{DGAE}$ to initialize the DGAE and the graph codebook, with the self-supervised training on gold 3DSG data. For this step, we specially choose the 3DSG datasets [72, 78, 27] for training and the graph codebook will be well initialized. **Step-2 Spatial Alignment.** Then we tune the visual and text decoder with the loss in Eq. 12. During this step, only visual decoder and text decoder are updated with the loss $\mathcal{L}_{v-dec} + \mathcal{L}_{t-dec}$. **Step-3 2DSG→3DSG Diffusion Training.** In this step, we use the constructed aligned 3DSG-Image-Text data (§4.1) to train the 2DSG→3DSG graph diffusion, thus the model will acquire the prior 3D estimation capability. The optimization objectives here is $\mathcal{L}_{I23D} + \mathcal{L}_{T23D}$. **Step-4 Overall Training.** Finally we tune the whole model with dual learning objectives, where we update the three diffusion models and freeze other modules. The overall learning target is $\mathcal{L}_{dual} + \mathcal{L}_{mutual}$.
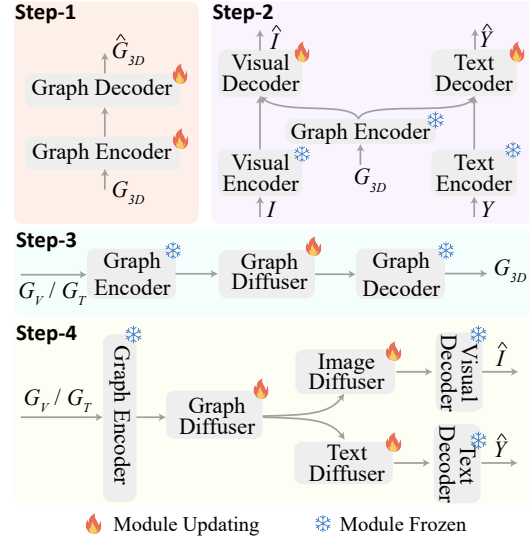

Figure 3: Illustrations of the training steps of SD^3.

After above training steps, the SI2T or ST2I can be launched alone without the intermediate process sharing. At this point, the critical graph diffusion module has learned sufficiently well from the previous training, so without the dual task aiding, the 3DSG generated during inference is also of high quality for the following 3D→X generation.

5 Experiment

5.1 Experiment Settings

Dataset. To demonstrate the capability of our proposed method for both ST2I and SI2T generation, we conduct experiments on the VSD [95, 97] dataset, which is constructed for visual spatial understanding. The VSD dataset contains about 30K images from SpatialSense [89] and Visual Genome

Table 2: Main results on the 256×256-sized VSD dataset with 200 DDIM steps. **Bold** numbers are the best and the Underline numbers denote the best baselines. Our results are averaged on five running with different seeds.

| | VSDv1 | | | | | VSDv2 | | | | |
| | ST2I | | | SI2T | | ST2I | | | SI2T | |
	FID↓	IS↑	CLIP↑	BLEU4↑	SPICE↑	FID↓	IS↑	CLIP↑	BLEU4↑	SPICE↑
● T2I Baselines										
DALLE [62]	32.55	17.01	62.16	-	-	28.52	21.18	64.58	-	-
Cogview [13]	32.30	17.07	61.85	-	-	28.17	21.74	64.76	-	-
LAFITE [98]	30.73	24.39		-	-	25.73	25.47		-	-
VQ-Diffusion [28]	18.34	20.58	63.42	-	-	15.66	24.75	66.30	-	-
Friodo [15]	<u>12.86</u>	<u>25.92</u>	<u>64.65</u>	-	-	<u>11.41</u>	<u>26.02</u>	<u>67.01</u>	-	-
● I2T Baselines										
3DVSD [97]	-	-	-	<u>54.85</u>	<u>68.76</u>	-	-	-	<u>26.40</u>	<u>46.97</u>
MNIC [25]	-	-	-	34.21	66.87	-	-	-	20.01	43.88
FNIC [23]	-	-	-	37.03	66.50	-	-	-	22.62	43.52
DiffCap [29]	-	-	-	34.75	66.39	-	-	-	20.27	43.30
DDCap [99]	-	-	-	37.93	67.10	-	-	-	23.14	44.07
Singleton	18.05	20.42	63.51	48.77	66.59	14.70	24.62	66.41	23.51	43.70
Singleton + 3D	12.56	26.92	65.62	50.05	67.20	10.43	25.62	67.29	25.37	45.13
Vanilla Dual Learning	11.80	27.85	67.18	51.59	67.79	11.67	27.80	68.46	26.10	46.72
SD^3 (Ours)	**11.04**	**29.20**	**68.31**	**56.23**	**68.02**	**10.09**	**29.76**	**71.10**	**27.63**	**48.03**

[43], and each image in this dataset has aligned text for spatial scene description. Now, VSD has two released versions, VSDv1 and VSDv2, where the spatial descriptions in VSDv2 are more meticulous.

Evaluation metrics. For spatial text-to-image synthesis, following prior work, we adopt Fréchet Inception Distance (FID) [66], Inception score (IS) [31], and CLIP score [30] to evaluate authenticity of generated images. For spatial image-to-text, we follow [95] and adopt BLEU4 [55] and SPICE [1] to measure text generation. Furthermore, we perform human evaluation to compare the spatial understanding effectiveness. Appendix B details more experiment settings.

Implementation Details. We use the pre-trained VQ-VAE of VQ-GAN [14], which leverages the GAN loss to get a more realistic image. It converts 256×256 images into 32×32 tokens. We follow [28] and remove the useless codes and obtain a codebook with size K=2886. For text encoder, we adopt the CLIP model and its tokenizer, which has 77 as the max token lengths. We adopt the pre-trained GPT-2 [59] as the text decoder, following its default settings. We freeze the VQ-VAE encoder and CLIP while training. For the user Unet to learn the denoising process for all the three diffusion models. We follow the default settings of DGAE [54] for the graph diffusion model. We optimize the framework using AdamW [50] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate is set to 5e-5 after 10,000 warmup iterations in the final dual tuning.

5.2 Quantitative Results

Main Comparisons. The main results are shown in Table 2 We first compare some strong baselines of I2T and T2I methods. Overall, the ST2I results on VSDv2 are consistently better, and the SI2T results on VSDv1 are better. This is because the VSDv2 contains more complicated and detailed textual descriptions, which is beneficial to ST2I but challenging for ST2I. We see that the proposed SD^3 substantially outperforms the compared baselines on both VSDv1 and VSDv2. For the ST2I task, we outperform Frido by 1.82% FID, 3.28% IS. and 3.66% CLIP score on VSDv1 and 1.32% FID, 3.74% IS and 4.09% CLIP score on VSDv2 For the SI2T task, we surpass 3DVSD by 1.38% BLEU4 and 0.74% SPICE on VSDv1 and 1.23% BLEU4 and 1.06% SPICE on VSDv2. The results directly demonstrate the efficacy of our method.

We further show the key module ablation studies on the last four lines in Table 2. First, removing the 3DSG integration ("Vanilla Dual Learning"), i.e., train the SI2T and ST2I diffusion with loss of Eq. 9, the performance drops much on main metrics, indicating the critical influence of the 3DSG guidance. Also, the drops of "Singleton+3D" reveal the effectiveness of our dual learning strategy.
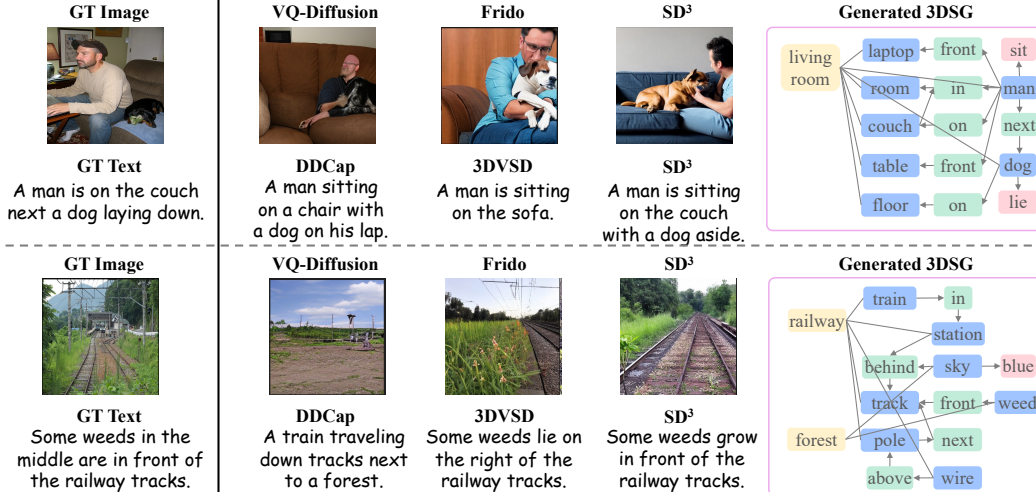
Figure 4: Qualitative results by different models, where the samples are selected from VSDv2.

When removing both 3DSG modeling and dual learning ("Singleton"), the model decay to the simple discrete diffusion model, which is just comparable to the VQ-Diffusion and DDCap model.

Spatial Evaluation. To compare the model's effectiveness, we conduct human evaluations for both ST2I and SI2T. We follow [97] to ask ten volunteers to answer the 5-point Likert scale on 100 samples. For SI2T, we collect scores for *Spatial Description Accuracy* and *Spatial Description Diversity*. For ST2I, we collect scores for *Visual Spatial Accuracy*. The results are shown in Table 3. Overall, our SD^3 method shows the best spatial understanding capability on SI2T and ST2I. More details are put in Appendix §B.

Table 3: Human Evaluation for Spatial Accuracy.

	ST2I	SI2T	
	SpaAcc.	SpaAcc.	Div.
3D^3	3.86	4.03	3.62
Frido	3.04	-	-
3DVSD	-	3.03	2.76

5.3 Qualitative Results

To present the ST2I and SI2T performance of our model in a more intuitive way, we show some qualitative results in Figure 4. We visualize the generated images and texts using different methods. For ST2I, compared with Frido and VQ-Diffusion, our SD^3 generates more spatial-faithful images. For SI2T, compared with 3DVSD and DDCap, our method is able to extract and correctly describe the key spatial information.
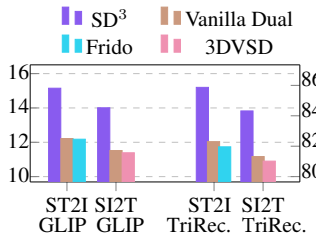

Figure 5: The evaluation of structure matching on VSDv2.

Table 4: Comparing the generated 3DSG with the gold 3GSD with TriRec.

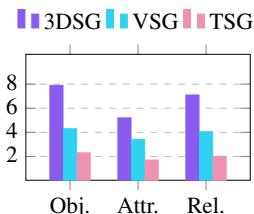| Hydra (GT)	100
Initial TSG	57.61
Initial VSG	24.20
3D^3 w/o DiffPre.	56.82
3D^3 w/o Xfeat.	65.71
Full 3D^3	84.51

Figure 6: The change of average number of SG nodes.

5.4 In-depth Analyses and Discussions

Previously, we have verified the effectiveness of our dual learning by thorough numerical evaluations. Now, we explore how our methods advance via the following research questions.

How does 3DSG guidance aid the generation of spatial-faithful image and text? First, after removing the 3DSG integration, comparing the last two lines in Table 2, we can see that the 3DSG feature contributes great influence. Further, we follow GLIP [47] to assess how well the objects-

attribute matching between input image/text and generated ones, Also, we assess how well the subject-predicate-object triplets of the gold TSG and VSG could be retrieved in the ones from the generated image/text via the triplets recall (TriRec.) between scene graphs. We compare the SoTA SI2T and ST2I baselines and the results are shown in Figure 5. We observe that our SD^3 significantly outperforms the model without 3DSG guidance ("Vanilla Dual") on the two metrics, which is just comparable to the baseline models. This suggests that with the 3DSG guidance, the model could capture correct and sufficient spatial structures from inputs for the generation.

How does the dual intermediate sharing aid the graph diffusion model? On one hand, we report the performance of 3DSG generation on the gold 3DSG dataset [78] to evaluate how well the structures of generated 3DSG could be retrieved in the gold ones. As shown in Table 4, we can see that compared with the initial VSG/TSG the generated 3DSG of our full model highly matches to the ground truth while without the dual intermediate sharing ("w/o Xfeat."), i.e., remove the $\mathcal{L}_{X-I23D}$ and $\mathcal{L}_{X-T23D}$ in the $\mathbf{X}{\rightarrow}\mathbf{3D}$ diffusion, the performance dramatically drops, showing the effectiveness of intermediate process sharing. Besides, if the graph diffusion pre-training is removed ("w/o DiffPre."), it will lose the ability to generate 3DSG (degrade to initial TSG/VSG). On the other hand, we assess how much the graph diffusion model produce the new structures, compared with the initial TSG/VSG. As shown in Figure 6, after graph diffusion, the average numbers of the three types of elements are enriched substantially.

The influence of the training strategy.
To show the effectiveness of our four-step training strategy, here we conduct an objective ablation study. We investigate the model without DGAE pre-training in step 1, the model without spatial alignment in step 2, and the model without 2DSG→3DSG in step 3, respectively. As shown in Figure 7, we can see both ST2I and SI2T performance decreases when removing any training steps. No-

Figure 7: Comparison of different training strategy.

tably, the 2DSG→2DSG step contributes the most to the final results. The essential 3D understanding function lies in the capability of the graph diffusion module, which should be largely enhanced via this step.
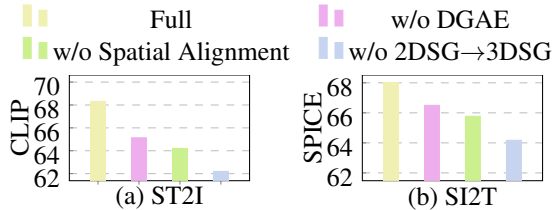
6 Conclusion

In this work, we study the ST2I and SI2T tasks under a dual learning framework. We first propose the spatial-aware 3D scene graph to model the 3D feature which is essentially shared by ST2I and SI2T. The 3DSG is constructed by being initialized with a 2D TSG/VSG and then evolved to the 3DSG by a graph diffusion model. We then leverage the dual learning to enhance the 2DSG→3DSG evolving by sharing clues of the dual 3DSG→Image/Text process, through which the 3DSG diffusion model is adequately guided and then facilitates the whole ST2I and SI2T. On the VSD dataset, our method shows great superiority in both SI2T and ST2I. Further analyses demonstrate how our dual learning method could capture 3D spatial structures and then help generate spatial-faithful images and texts.

Looking forward, there can be quite rich explorations in future research. First, the graph diffusion model highly relies on the quality of 3DSG training data, which has significant impact to the final performance of our method. Correspondingly, we in the future will explore the data augmentation methods to optimize the training. Second, the evaluation for spatial understanding of ST2I and SI2T has not been fully explored in this work, i.e., with human evaluation instead. Also, one promising direction for VSU is constructing multimodal large language models (MLLMs) [22, 85, 20, 84, 94, 57, 82], especially for 3D spatial understanding [9].

Acknowledgements

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2016. doi: 10.1007/978-3-319-46454-1_24. URL https://doi.org/10.1007/978-3-319-46454-1_24.

[2] Iro Armeni, Zhi-Yang He, Amir Zamir, JunYoung Gwak, Jitendra Malik, Martin Fischer, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5663–5672. IEEE, 2019. doi: 10.1109/ICCV.2019.00576. URL https://doi.org/10.1109/ICCV.2019.00576.

[3] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

[4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 1737–1752. PMLR, 2023. URL https://proceedings.mlr.press/v202/bar-tal23a.html.

[5] Yoann Boget, Magda Gregorova, and Alexandros Kalousis. Discrete graph auto-encoder. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/pdf?id=bZ80b0wb9d.

[6] Allan M. C. Bretas, Alexandre Mendes, Martin Jackson, Riley Clement, Claudio Sanhueza, and Stephan K. Chalup. A decentralised multi-agent system for rail freight traffic management. *Ann. Oper. Res.*, 320(2):631–661, 2023. doi: 10.1007/s10479-021-04178-x. URL https://doi.org/10.1007/s10479-021-04178-x.

[7] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00081. URL https://doi.org/10.1109/3DV.2017.00081.

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *CoRR*, abs/2401.12168, 2024. doi: 10.48550/ARXIV.2401.12168.

[9] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024.

[10] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 2021. URL http://proceedings.mlr.press/v139/cho21a.html.

[11] Fatemeh Daneshfar, Ako Bartani, and Pardis Lotfi. Image captioning by diffusion models: A survey. *Engineering Applications of Artificial Intelligence*, 138:109288, 2024.

[12] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

[13] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 19822–19835, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a4d92e2cd541fca87e4620aba658316d-Abstract.html.

[14] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01268. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

[15] Wan-Cyuan Fan, Yen-Chun Chen, Dongdong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 579–587. AAAI Press, 2023. doi: 10.1609/AAAI.V37I1.25133. URL https://doi.org/10.1609/aaai.v37i1.25133.

[16] Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning*, pages 6373–6391, 2022.

[17] Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5980–5994, 2023.

[18] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.

[19] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.

[20] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024.

[21] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8, 2024.

[23] Zheng-cong Fei. Fast image caption generation with position alignment. *CoRR*, abs/1912.06365, 2019. URL http://arxiv.org/abs/1912.06365.

[24] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *J. Artif. Intell. Res.*, 74:459–515, 2022. doi: 10.1613/jair.1.13646. URL https://doi.org/10.1613/jair.1.13646.

[25] Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. Masked non-autoregressive image captioning. *CoRR*, abs/1906.00717, 2019. URL http://arxiv.org/abs/1906.00717.

[26] Nishad Gothoskar, Marco F. Cusumano-Towner, Ben Zinberg, Matin Ghavamizadeh, Falk Pollok, Austin Garrett, Josh Tenenbaum, Dan Gutfreund, and Vikash K. Mansinghka. 3dp3: 3d scene perception via probabilistic programming. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9600–9612, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4fc66104f8ada6257fa55f29a2a567c7-Abstract.html.

[27] Elias Greve, Martin Büchner, Niclas Vödisch, Wolfram Burgard, and Abhinav Valada. Collaborative dynamic 3d scene graphs for automated driving. *CoRR*, abs/2309.06635, 2023. doi: 10.48550/ARXIV.2309.06635. URL https://doi.org/10.48550/arXiv.2309.06635.

[28] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10686–10696. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01043. URL https://doi.org/10.1109/CVPR52688.2022.01043.

[29] Yufeng He, Zefan Cai, Xu Gan, and Baobao Chang. Diffcap: Exploring continuous diffusion on image captioning. *CoRR*, abs/2305.12144, 2023. doi: 10.48550/ARXIV.2305.12144. URL https://doi.org/10.48550/arXiv.2305.12144.

[30] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595. URL https://doi.org/10.18653/v1/2021.emnlp-main.595.

[31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[33] Yicong Hong, Cristian Rodriguez Opazo, Qi Wu, and Stephen Gould. Sub-instruction aware vision-and-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3360–3376. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.271. URL https://doi.org/10.18653/v1/2020.emnlp-main.271.

[34] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *CoRR*, abs/2102.05379, 2021. URL https://arxiv.org/abs/2102.05379.

[35] Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and Ponnuthurai N. Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11492–11501. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01121. URL https://doi.org/10.1109/CVPR52688.2022.01121.

[36] Nathan Hughes, Yun Chang, Siyi Hu, Rajat Talak, Rumaisa Abdulhai, Jared Strader, and Luca Carlone. Foundations of spatial perception for robotics: Hierarchical representations and real-time systems. *CoRR*, abs/2305.07154, 2023. doi: 10.48550/ARXIV.2305.07154. URL https://doi.org/10.48550/arXiv.2305.07154.

[37] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1655–1663, 2021.

[38] Jiayi Ji, Yiwei Ma, Xiaoshuai Sun, Yiyi Zhou, Yongjian Wu, and Rongrong Ji. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31: 4321–4335, 2022.

[39] Justin Johnson et al. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL https://arxiv.org/abs/1504.00325.

[40] Justin Johnson et al. Image generation from scene graphs. In *European Conference on Computer Vision (ECCV)*, 2018. URL https://arxiv.org/abs/1703.07370.

[41] Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE Trans. Cybern.*, 50(12): 4921–4933, 2020. doi: 10.1109/TCYB.2019.2931042. URL https://doi.org/10.1109/TCYB.2019.2931042.

[42] Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. CLEF 2017: Multimodal spatial role labeling task working notes. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017. URL https://ceur-ws.org/Vol-1866/invited_paper_19.pdf.

[43] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. doi: 10.1007/S11263-016-0981-7. URL `https://doi.org/10.1007/s11263-016-0981-7`.

[44] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. Toward 3d spatial reasoning for human-like text-based visual question answering. *CoRR*, abs/2209.10326, 2022. doi: 10.48550/ARXIV.2209.10326. URL `https://doi.org/10.48550/arXiv.2209.10326`.

[45] Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation with semantics-prototype learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3145–3153, Mar. 2024. doi: 10.1609/aaai.v38i4.28098. URL `https://ojs.aaai.org/index.php/AAAI/article/view/28098`.

[46] Li Li, You Qin, Wei Ji, Yuxiao Zhou, and Roger Zimmermann. Domain-wise invariant learning for panoptic scene graph generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3165–3169, 2024. doi: 10.1109/ICASSP48485.2024.10447193.

[47] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01069. URL `https://doi.org/10.1109/CVPR52688.2022.01069`.

[48] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11, 2023.

[49] Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2365–2376. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.168. URL `https://doi.org/10.18653/v1/2022.acl-long.168`.

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[51] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[52] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4582–4598. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.364. URL `https://doi.org/10.18653/v1/2021.naacl-main.364`.

[53] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Intell. Res.*, 71:1183–1317, 2021. doi: 10.1613/jair.1.11688. URL `https://doi.org/10.1613/jair.1.11688`.

[54] Van Khoa Nguyen, Yoann Boget, Frantzeska Lavda, and Alexandros Kalousis. Discrete latent graph generative modeling with diffusion bridges. *CoRR*, abs/2403.16883, 2024. doi: 10.48550/ARXIV.2403.16883. URL `https://doi.org/10.48550/arXiv.2403.16883`.

[55] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040/`.

[56] Tao Pu. Video scene graph generation with spatial-temporal knowledge. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 9340–9344. ACM, 2023. doi: 10.1145/3581783.3613433. URL `https://doi.org/10.1145/3581783.3613433`.

[57] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.

[58] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023.

[59] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[61] Navid Rajabi and Jana Kosecka. Towards grounded visual spatial reasoning in multi-modal vision language models. *CoRR*, abs/2308.09778, 2023. doi: 10.48550/ARXIV.2308.09778. URL https://doi.org/10.48550/arXiv.2308.09778.

[62] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. URL http://proceedings.mlr.press/v139/ramesh21a.html.

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

[64] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From SLAM to spatial perception with 3d dynamic scene graphs. *Int. J. Robotics Res.*, 40(12-14):1510–1546, 2021. doi: 10.1177/02783649211056674. URL https://doi.org/10.1177/02783649211056674.

[65] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

[66] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

[67] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJrFC6ceg.

[68] Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language, VL@EMNLP 2015, Lisbon, Portugal, September 18, 2015*, pages 70–80. Association for Computational Linguistics, 2015. doi: 10.18653/V1/W15-2812. URL https://doi.org/10.18653/v1/W15-2812.

[69] Jiansong Sha, Haoyu Zhang, Yuchen Pan, Guang Kou, and Xiaodong Yi. Nerf-is: Explicit neural radiance fields in semantic space. In *ACM Multimedia Asia 2023, MMAsia 2023, Tainan, Taiwan, December 6-8, 2023*, pages 10:1–10:7. ACM, 2023. doi: 10.1145/3595916.3626379. URL https://doi.org/10.1145/3595916.3626379.

[70] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/sohl-dickstein15.html.

[71] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

[72] Jared Strader, Nathan Hughes, William Chen, Alberto Speranzon, and Luca Carlone. Indoor and outdoor 3d scene graph generation via language-enabled spatial ontologies. *CoRR*, abs/2312.11713, 2023. doi: 10.48550/ARXIV.2312.11713. URL https://doi.org/10.48550/arXiv.2312.11713.

[73] Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. Dual supervised learning for natural language understanding and generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5472–5477. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1545. URL https://doi.org/10.18653/v1/p19-1545.

[74] Haifeng Sun, Xiaozheng Zheng, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. SMR: spatial-guided model-based regression for 3d hand pose and mesh reconstruction. *IEEE Trans. Circuits Syst. Video Technol.*, 34(1):299–314, 2024. doi: 10.1109/TCSVT.2023.3285153. URL https://doi.org/10.1109/TCSVT.2023.3285153.

[75] Yibo Sun, Duyu Tang, Nan Duan, Tao Qin, Shujie Liu, Zhao Yan, Ming Zhou, Yuanhua Lv, Wenpeng Yin, Xiaocheng Feng, Bing Qin, and Ting Liu. Joint learning of question answering and question generation. *IEEE Trans. Knowl. Data Eng.*, 32(5):971–982, 2020. doi: 10.1109/TKDE.2019.2897773. URL https://doi.org/10.1109/TKDE.2019.2897773.

[76] Duyu Tang, Nan Duan, Tao Qin, and Ming Zhou. Question answering and question generation as dual tasks. *CoRR*, abs/1706.02027, 2017. URL http://arxiv.org/abs/1706.02027.

[77] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, pages 301–308. IEEE, 2017. doi: 10.1109/ASRU.2017.8268950. URL https://doi.org/10.1109/ASRU.2017.8268950.

[78] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3960–3969. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00402. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wald_Learning_3D_Semantic_Scene_Graphs_From_3D_Indoor_Reconstructions_CVPR_2020_paper.html.

[79] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22al.html.

[80] Shanpeng Wang, Zhenbing Qiu, Panpan Huang, Xiang Yu, Jian Yang, and Lei Guo. A bioinspired navigation system for multirotor UAV by integrating polarization compass/magnetometer/ins/gnss. *IEEE Trans. Ind. Electron.*, 70(8):8526–8536, 2023. doi: 10.1109/TIE.2022.3212421. URL https://doi.org/10.1109/TIE.2022.3212421.

[81] Yang Wang et al. Scene graph generation by iterative message passing. In *European Conference on Computer Vision (ECCV)*, 2018. URL https://arxiv.org/abs/1711.05846.

[82] Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *arXiv preprint arXiv:2407.21534*, 2024.

[83] Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2stra: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. *arXiv preprint arXiv:2305.12260*, 2023.

[84] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.

[85] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*, 2024.

[86] Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

[87] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from RGB-D sequences. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7515–7525. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00743. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wu_SceneGraphFusion_Incremental_3D_Scene_Graph_Prediction_From_RGB-D_Sequences_CVPR_2021_paper.html.

[88] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3789–3798. PMLR, 2017. URL http://proceedings.mlr.press/v70/xia17a.html.

[89] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2051–2060. IEEE, 2019. doi: 10.1109/ICCV.2019.00214. URL https://doi.org/10.1109/ICCV.2019.00214.

[90] Xinlei Yang et al. X-lan: Cross-modal scene graph alignment network for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1906.11097.

[91] Hai Ye, Wenjie Li, and Lu Wang. Jointly learning semantic parser and natural language generator via dual information maximization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2090–2101. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1201. URL https://doi.org/10.18653/v1/p19-1201.

[92] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5831–5840. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00611. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zellers_Neural_Motifs_Scene_CVPR_2018_paper.html.

[93] Meishan Zhang, Huiyao Chen, Xin Zhang, Jing Chen, and Min Zhang. Faking a teacher works! dependency scoring learning and corpus boosting for translation-based cross-lingual dependency parsing. *Dependency Scoring Learning and Corpus Boosting for Translation-Based Cross-Lingual Dependency Parsing*, 2023. URL https://ssrn.com/abstract=4626681.

[94] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024.

[95] Yu Zhao, Jianguo Wei, Zhichao Lin, Yueheng Sun, Meishan Zhang, and Min Zhang. Visual spatial description: Controlled spatial-oriented image-to-text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1437–1449. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.93. URL https://doi.org/10.18653/v1/2022.emnlp-main.93.

[96] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5281–5291, 2023.

[97] Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3d scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7960–7977. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.442. URL https://doi.org/10.18653/v1/2023.acl-long.442.

[98] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. LAFITE: towards language-free training for text-to-image generation. *CoRR*, abs/2111.13792, 2021. URL `https://arxiv.org/abs/2111.13792`.

[99] Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. Exploring discrete diffusion models for image captioning. *CoRR*, abs/2211.11694, 2022. doi: 10.48550/ARXIV.2211.11694. URL `https://doi.org/10.48550/arXiv.2211.11694`.

A Extension of Technical Details

In this section, we introduce the specific details of our method which we cannot present in the main article due to the space limit.

A.1 Discrete Diffusion

We first introduce the discrete representation of continuous data. Given data x in a continuous space, a vector quantized model (VQM) is employed to encode x to embedding vectors. A VQM contains an encoder E, a decoder G and a pre-trained codebook $\mathcal{Z} = \{c_k\}_{k=1}^K \in \mathbb{R}^{K \times d}$, where $\mathcal{Z}$ has a finite number of embedding vectors, K is the size of the codebook and d is the code dimension. Given input x, we obtain a sequence of tokens z_q with the encoder $e = E(x)$ and a quantizer that maps e to its closet codebook entry c_k:

$$z_q = Q(e) = \underset{k \in Z}{argmin} \, \|e - c_k\| \tag{14}$$

Correspondingly, given a quantified tokens z_q, the decoder could faithfully reconstruct the data $\tilde{x} = G(z_q)$.

With this discrete representation, the discrete diffusion model can be applied to it. Given the quantized data z_0, the forward diffusion process gradually corrupts z_0 through the Markov chain $q(z_t|z_{t-1})$, randomly replacing some tokens in z_{t-1}. After a fixed number of T steps, the model outputs a sequence of increasingly noisy latent variables $z_1, ..., z_T$, where z_T is the pure noise tokens. In the reverse process, the model gradually denoise from z_T and reconstruct z_0, by sampling from the distribution $q(z_{t-1}|z_t, z_0)$ sequentially. Specifically, for a token z_0^i of z_0, z_0^i takes the index of one entry of codebook, i.e., $z_0^i \in \{1, ..., K\}$. The probabilities of the transition from z_{t-1} to z_t can be represented by a matrix $Q_t(mn) = q(z_t = m|z_{t-1} = n) \in \mathcal{R}^{K \times K}$. Then the forward diffusion process for z_t is:

$$q(z_t|z_{t-1}) = v^\top(z_t)Q_t v(z_{t-1}), \tag{15}$$

where $v(z)$ means the one-hot representation of x in K categories. $Q_t v(z_{t-1})$ means the categorical distribution over z_t. Due to its Markov property, the $q(z_t|z_0)$ could be written as:

$$q(z_t|z_0) = v^\top(z_t) \left(\prod_1^t Q_{t'} \right) v(z_0)$$
$$= v^\top(z_t) \overline{Q}_t v(z_0). \tag{16}$$

By applying Bayes' rule, we can compute the posterior $q(z_{t-1}|z_t, z_0)$ as:

$$q(z_{t-1}|z_t, z_0) = \frac{q(z_t|z_{t-1}, z_0)q(z_{t-1}|z_0)}{q(z_t|z_0)}$$
$$= \frac{\left(v^\top(z_t)Q_{t-1}v(z_{t-1}) \right) \left(v^\top(z_{t-1})\overline{Q}_{t-1}v(z_0) \right)}{v^\top(z_t)\overline{Q}_t v(z_0)}. \tag{17}$$

Q_t is usually defined as the a small amount of uniform noises and it can be formulated as:

$$Q_t = \begin{bmatrix} \alpha_t + \beta_t & \beta_t & \cdots & \beta_t \\ \beta_t & \alpha_t + \beta_t & \cdots & \beta_t \\ \vdots & \cdots & \ddots & \vdots \\ \beta_t & \beta_t & \cdots & \alpha_t + \beta_t \end{bmatrix} \tag{18}$$

where $\alpha_t \in [0, 1]$ and $\beta_t = (1 - \alpha_t)/K$, which means each token has a probability of $\alpha_t + \beta_t$ to remain the previous value at one step and has a probability of $K\beta_t$ to be sampled uniformly over all the K categories.

Then a noise estimating network $p_\theta(z_{t-1}|z_t, c)$ is trained to approximate the conditional transit distribution $q(z_{t-1}|z_t, x_0)$ with condition c. The network is trained to minimize the variational lower bound (VLB) [70].

$$\mathcal{L} = D_{KL}(q(z_{t-1}|z_t, x_0) \| p_\theta(z_{t-1}|z_t, c)) \tag{19}$$

A.2 Discrete Representation for 3DSG

Following [54], given the 3DSG $\mathcal{G}$, we use a a GNN encoder to encode G to a set of node embeddings $\mathcal{Z} = \boldsymbol{z}_i^G$. We subsequently apply a quantization operator, $\mathcal{F}$, on the continuous node embeddings and map them to fixed points within the same space. The quantization operates independently on each dimension of the latent space, and the quantization of the jth dimension of z_i embedding is given by:

$$z_{ij}^q = \mathcal{F}(z_{ij}^G, L_j) = \mathcal{R}(\frac{L_i}{2} \tanh z_{ij}), 1 \leq j \leq d_Z \tag{20}$$

where $\mathcal{R}$ is the rounding operator and L_j is the number of quantization levels used for the jth dimension of the node embeddings. The quantisation operator maps any point in the original continuous latent space to a point from the set $\overline{Z} = \overline{z}_i$. The discrete latent representation of graph $\mathcal{G}$ is the set $Z^G = \{z_i^q\}, z_i^q \in \overline{Z}$.

The quantization operator is permutation equivariant and so is the mapping from the initial graph G to its discrete representation in the latent space as long as the graph encoder $\mathcal{D}$ is permutation equivariant. Thus for any P permutation matrix the following equivariance relation holds:

$$P^\top Z^G = \mathcal{F}(P^\top Z) = \mathcal{F}(\mathcal{D}(P^\top EP, P^\top X)) \tag{21}$$

Using an equivariant decoder $\mathcal{D}(Z^G)$, which will operate on the assumption of a fully-connected graph over the node embeddings Z^G, results in reconstructed graphs that are node permutation equivariant with the original input graphs.

A.3 Aligned Image-Text-3DSG Data Construction.

We follow [72] to take the 3D datasets Matterport3D (MP3D) [7] 3DSSG [78] and CURB-SG [27], using the Hydra [36] parser to generate the 3D scene graphs and then leverage large language models to refine it as the ground-truth. Then we use ChatGPT to generate alignede spatial descriptions for the RGB images.

A.4 Marginal Distribution Estimation for the Dual Learning Target

Based on Eq. 9, the marginal distributions $p(x)$ and $p(y)$ can not be aqcired directly. Thus we estimate these marginal distribution $p(x)$ and $p(y)$ with a surrogate distribution $\hat{p}(x)$ and $\hat{p}(y)$, by observing the target in the scope of the whole data. For the text, we use a Transformer-based language model that is trained over the specific data to calculate the $\hat{p}(x)$ [73]. For the image, we follow [16] to calculate $\hat{p}(y) = \prod_{t=1}^m px_i | x_{<i}$. We serialize the image pixels as x_i and use PixelCNN++ [67] to model this distribution.

B Detailed Experiment Settings

B.1 Evaluation Metric Implication

We employ Fréchet Inception Distance (FID) [66], Inception score (IS) [31], CLIP score [30], and GLIP [47] used in [15] to quantitatively evaluate the quality of the generated images. Additionally, we introduce Triplet Recall (TriRec.) to measure the percentage of the correct relation triplet among all the relations. Technically, given a set of ground truth triplets (subject-predicate-object), denoted GT, and the TriRec. is computed as TriRec.$=|PT \cap GT|/|GT|$, where PT are the relation triplets extracted from the generated images by a visual or textual SG parser.

B.2 Human Evaluation Criterion

We conduct a human evaluation to mainly focus on spatial understanding quality.

Then we design a 5-point Likert scale is designed as follows:

- SI2T Spatial Accuracy: The sentences correctly describe the spatial relationship of the objects in the given image.
- SI2T Spatial Diversity: The generated sentences describe diversified spatial semantics.

- ST2I Spatial Accuracy; The image present the spatial description correctly.

Each question will be answered by a number from 1 to 5, denoting "Strongly Disagree", "Disagree", "Neither disagree nor agree", "Agree" and "Strongly agree". We generated 5 sentences for SI2T and 3 images for ST2I from 100 samples.

C   Extended Experiment Results and Analyses

Table 5: Comparison with latest T2I methods.

	ST2I		
	FID$\downarrow$	IS$\uparrow$	CLIP$\uparrow$
SD-1.5	20.12	26.35	64.79
SDXL	12.60	27.95	66.41
CogView 3	13.28	28.63	67.29
Ours	**11.04**	**29.20**	**68.31**

C.1   Comparison with More Methods

We add more experiments of the latest T2I methods in table 5. Comparing with the recent SD and CogView, our method exhibit leading performance with the similar scale.

C.2   The Superiority of the Discrete Modeling

To verify the superiority of the discrete modeling, we compare with the continuous diffusion backbone in table 6. We find that the results of continuous backbone drops slightly on ST2I and SI2T while drops more on 3DSG generation, revealing that the discrete model show superiority for structure data modeling and further benefit the final performance.

Table 6: Comparison between discrete diffusion and continuous diffusion.

	3DSG	ST2I			SI2T	
	TriRec.	FID$\downarrow$	IS$\uparrow$	CLIP$\uparrow$	BLEU4$\uparrow$	SPICE$\uparrow$
w/ Descrete Diff (SD^3)	84.51	10.09	29.76	71.10	27.63	48.03
w/ Continuous Diff	83.14	10.25	29.68	70.02	27.51	47.82

C.3   Can Non-diffusion SI2T models be used in our Dual Learning Framework?

We use the diffusion based SI2T model to maintain the symmetry for diffusion based ST2I and 3DSG generation, so that their intermediate feature could be mutually enhanced. Here we explore whether a vision-language model (VLM) based SI2T model could be integrated in our dual framwework. We follow [97] to use a OFA model as SI2T backbone. We replace the z_t^Y in $p_\psi^{Graph}(z_{t-1}^G | z_t^G, c^Y \oplus z_t^Y)$ to the OFA encoder output hidden states z_{enc}^Y. Then we train our model with the same training data and training strategy. The results are shown in Table 7, where we find the ST2I performance drops while the SI2T performance keep comparable. This reveal the necessity of choosing diffusion based SI2T.

Table 7: Comparison between diffusion and non-diffusion SI2T on VSDv2.

	ST2I			SI2T	
	FID$\downarrow$	IS$\uparrow$	CLIP$\uparrow$	BLEU4$\uparrow$	SPICE$\uparrow$
w/ Diff-SI2T (SD^3)	10.09	29.76	71.10	27.63	48.03
w/ OFA-SI2T	11.20	28.56	69.62	27.40	47.91

Table 8: Efficiency analysis for each step.

	Training Parameters	Training Time
Step-1 DGAE	110M	1 hours
Step-2 Spatial Alignment	127M	1.5 hours
Step-3 2DSG→3DSG Diffusion Training	350M	12 hours
Step-4 Overall Training	1.1B	20 hours

C.4 How the quality of 3DSG dataset influences the performance

Instinctive, the quality of 3DSG training data has strong impact to the graph diffusion. To verify this, we corrupt the gold 3DSG by random replace the node and edges with a probability and compare the model in Figure 8. We can see that with the noise increase, both the 3DSG and the final performance decrease sharply. This means the graph diffusion model is sensitive to the 3DSG data quality.
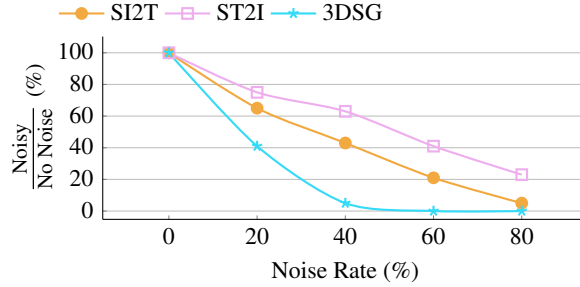

Figure 8: The evaluation of structure matching on the gold 3DSSG dataset with noise.

C.5 Visualization of 3DSG Generation Process

In Figure 9, we visualize the 3DSG generation process with two examples, where we sample several time steps and plot the generated 3DSG can be seen, compared with the initial VSG/TSG our graph diffusion is capable of enrich certain reasonable structures.

C.6 Efficiency Analysis

In table 8, we give the computational complexity analysis as well as our training settings.
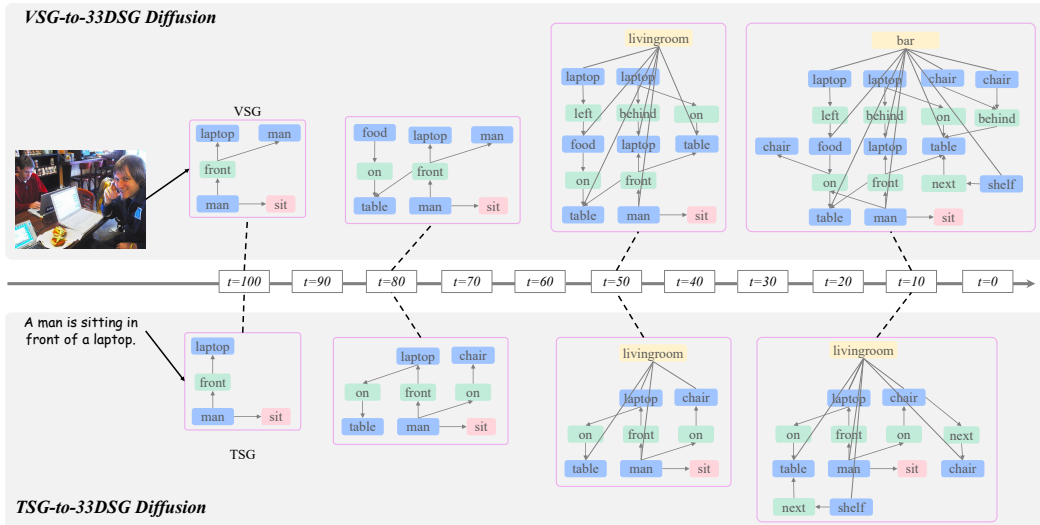

Figure 9: Visualization of 3DSG generation process from initial VSG and TSG.

The man in the yellow striped hoodie holds a colorful umbrella over his head.

Maroon curtain panel in front of window.

We can see a big bridge above the water in the distance.

Figure 10: More cases of ST2I generated by SD^3.

C.7 More Cases

Here we showcase more examples of the ST2I and SIT@ via SD^3 in Figure 10 and Figure 11.

	There are some fruits in the bowl on the table.	Some grapes are put in a white bowl.	A bowl of fruit with a lemon and a banana in it.
	The red chairs are put in front of the wooden table.	Rows of chairs and tables in the room.	A red chair with a counter top in the front.
	Two guys stand on the rock with bikes nearby.	Two guys are riding bicycles on the mountain.	Two guys handing the bikes are standing next to each other.

Figure 11: More cases of SI2T generated by SD^3.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have highlighted the contributions in abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: We discuss limitaitons in the Conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide main proof in the Method and full set of assumptions in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the detailed implementation settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will open source at Github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the detailed implementation settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our results are significantly comparable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We detail in the appendix.

Guidelines:
- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The experiments are conducted with the Code.

Guidelines:
- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: [NA]

Guidelines:
- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.