Towards Precise Embodied Dialogue Localization via Causality Guided Diffusion

Haoyu Wang[1] Le Wang[1*] Sanping Zhou[1] Jingyi Tian[1]
Zheng Qin[1] Yabing Wang[1] Gang Hua[2] Wei Tang[3]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Dolby Laboratories at Bellevue [3]University of Illinois at Chicago

Abstract

Embodied localization based on vision and natural language dialogues presents a persistent challenge in embodied intelligence. Existing methods often approach this task as an image translation problem, leveraging encoder-decoder architectures to predict heatmaps. However, these methods frequently experience a deficiency in accuracy, largely due to their heavy reliance on resolution. To address this issue, we introduce CGD, a novel framework that utilizes causality guided diffusion model to directly model coordinate distributions. Specifically, CGD employs a denoising network to regress coordinates, while integrating causal learning modules, namely back-door adjustment (BDA) and front-door adjustment (FDA) to mitigate confounders during the diffusion process. This approach reduces the dependency on high resolution for improving accuracy, while effectively minimizing spurious correlations, thereby promoting unbiased learning. By guiding the denoising process with causal adjustments, CGD offers flexible control over intensity, ensuring seamless integration with diffusion models. Experimental results demonstrate that CGD outperforms state-of-the-art methods across all metrics. Additionally, we also evaluate CGD in a multi-shot setting, achieving consistently high accuracy.

1. Introduction

With the rapid development of embodied intelligence in the field of artificial intelligence, precise localization has become a key capability for agents to perform downstream tasks such as emergency rescue [33, 45] and navigation [7, 22, 65]. However, achieving precise localization is still challenging for agents, especially in many human-robot interaction application scenarios [43]. To address this problem, Embodied Dialogue Localization (EDL) [18] is emerging as an important area of research that aims to provide agents with human-like map localization capabilities.

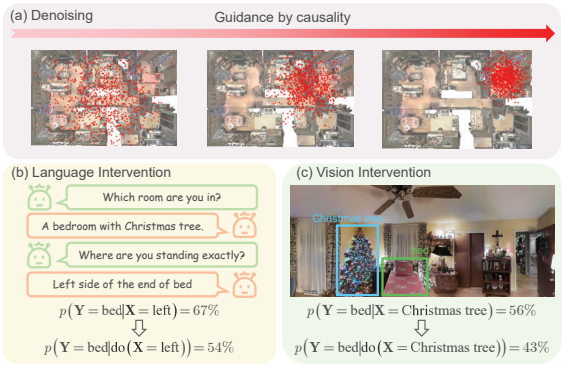

Figure 1. **Thumbnail of CGD.** De-confounded features are used to guide stable and unbiased denoising process. Interventions are implemented in both language and vision, enhancing the model's ability to generalize to unfamiliar environments.

The main challenge of EDL is to achieve cross-modal alignment and accurate localization on top-down maps. Previous approaches [17, 18, 71] modeled this challenge as an image-to-image transformation problem, using UNet or encoder-decoder architectures that convert raw images into heatmaps to predict the probability of coordinates. However, while these methods achieve good accuracy in a coarse range, they struggle with precise localization. DiaLoc [71] attempts to alleviate this problem by dividing dialogs into multiple rounds and updating the heatmap iteratively. However, heatmap-based methods are highly dependent on resolution, and resolution increase leads to exponential growth in computational complexity. On the other hand, these methods experience significant accuracy drops when generalized to unseen environments. Although using data augmentation or generating additional dialogs using LLMs

*Corresponding author.

has made progress in improving generalization capabilities, such approaches may be insufficient due to the inherent dataset biases [42, 68].

To address the above challenges, we propose a novel solution: improving the accuracy of EDL using causality guided diffusion models. By directly modeling the probability distribution of continuous locations, we aim to reduce reliance on resolution while improving localization accuracy. Due to the ambiguity of natural language, the probability distribution of continuous locations is often complex and variable. We observe that existing diffusion models [20, 40] offer an efficiency advantage in modeling such complex distributions, and therefore we utilize the denoising process in diffusion models to improve the prediction accuracy. However, as discussed before, objects in images and dialogs may introduce biases while guiding localization. For instance, as illustrated in Fig. 1, the conditional probability $p(\texttt{bed} \mid \texttt{Christmas tree})$ is high. However, there are no such strong causal relationships between these objects. Additionally, when certain room types are overrepresented in the training set, the model may inappropriately associate these features, leading to incorrect predictions when the text instruction changes. Due to such biases in the dataset, the effectiveness of diffusion models is inherently limited: irrelevant information in the input is independent of the predicted labeling, and thus the gradient can produce arbitrary or even antagonistic directions in the input space. Therefore, inspired by Classifier-free guidance [19], we creatively introduce causality guidance in the diffusion model to make the denoising process more robust and improve the generalization of the model in unseen environments.

Specifically, we propose a novel framework called Causality Guided Diffusion (CGD) to address the two main challenges mentioned above: 1) improving localization accuracy, and 2) reducing the interference of confounders. As shown in Fig. 2, the CGD framework comprises a diffusion network and a unified causal inference module, which includes both back-door adjustment (BDA) and front-door adjustment (FDA). First, we use the denoising process of the diffusion network to optimize the final coordinate prediction gradually. Second, a unified causal inference mechanism is introduced to deal with confounders in EDL. This unified approach addresses two categories of confounders: observable and unobservable. Observable confounders are related to content and are relatively straightforward to identify, such as keywords in dialogue and room types in the environment. On the other hand, unobservable confounders involve more subtle stylistic factors that are harder to detect but can still influence the system as a whole, such as decoration styles and lighting in visual inputs, or sentence structures in language. Using the de-confounded features as guidance, we seamlessly integrate the causal learning module with the diffusion network, leading to a robust and unbiased denoising process and enhancing the model's generalization ability in unseen environments.

In summary, our contributions can be summarized as follows: 1) We propose a novel CGD framework, leveraging diffusion models for directly modeling coordinate distributions to address embodied dialogue localization. This approach reduces the dependence on resolution for improving accuracy and enables precise localization within fine range. 2) We unify causal learning frameworks with diffusion models, integrating front-door and back-door adjustments to address observable and unobservable confounders, respectively. The de-obfuscated features serve as guidance for the denoising process, thereby enhancing the model's generalization ability in unseen environments. 3) Extensive evaluation of CGD shows that it achieves state-of-the-art performance, especially in unseen environments.

2. Related Work

2.1. Embodied Dialogue Localization

Recently, multimodal embodied AI tasks have attracted much attention, including embodied QA [11, 70], visual language navigation (VLN) [8, 16, 22, 63], and visual rearrangement [53, 64], etc. One essential capability embodied agents need for practical applications is localizing under language guidance. Pate et al. [37] propose a dataset that uses language descriptions instead of dialogues to guide localization. Text2Loc [66] explores localization in outdoor scenes based on point cloud. LingUNet [18] is the first to propose the EDL task. They propose the WAY dataset containing about 10,000 natural language dialogues, all of which are manually annotated by human operators from the first-person perspective of the agent. LingUNet models this complex task as an image-to-image conversion problem, using the upsampling and downsampling process of a UNet network to predict the target location on the top-down map. In contrast, LED-BERT [17] uses the navigation graph extracted from Matterport3D [6] instead of the top-down map and is the first to introduce Transformer [55] for multimodal alignment. DiaLoc [71] explores a new setting for embodied dialogue localization under multi-shot, divides conversations into multiple rounds, and inputs them into the model separately, which is more in line with human positioning thinking mode. Despite their advantages, heatmap-based methods still struggle with precise, fine-range localization due to limitations tied to resolution. To address this, we introduce a novel approach that directly models the original coordinate distribution using a diffusion model, enhancing model generalization and robustness under the guidance of causality.

2.2. Diffusion Models

Diffusion models [51] represent a class of generative models that operate through a two-step process: first, they progressively add noise to observed data, effectively corrupting it, and then they reverse this process to recover the original data. This mechanism of gradual noise addition and subsequent denoising makes diffusion models highly effective for generative tasks. The introduction of Denoising Diffusion Probabilistic Models (DDPMs) [20] has been particularly influential, as they establish a connection between diffusion models and denoising score matching, which has generated significant interest in the field. The DDPM framework has led to substantial advances in a variety of applications, including image generation [5, 21, 34, 47–49], cross-modal generation [2, 14, 23, 25, 26], graph generation [24, 35, 56, 67], semantic segmentation [1, 4], and object detection [9], among others. Notably, SR3 [49] adapts the principles of DDPMs to the domain of image super-resolution. Similarly, DiffusionDet [9] extends diffusion models to the task of object detection, utilizing them to recover bounding boxes from noisy data. Additionally, Pix2Seq-D [10] employs diffusion models in conjunction with analog bits to address the challenge of panoptic segmentation, demonstrating the versatility of DDPMs in different generative tasks. Embodied dialogue localization suffers from inherent ambiguity, making probabilistic generative methods suitable for this task. Since diffusion models are well-known for their ability to fit complex distributions, we apply DDPM to embodied dialogue localization.

2.3. Causal Inference

Causal inference [38] has emerged as a significant research area within machine learning, distinguishing itself from traditional methods by focusing on uncovering high-level causal relationships from low-level data, rather than merely identifying correlations between variables. In recent years, various studies have applied causal learning techniques to vision-and-language tasks [36, 44, 57, 58, 61, 73], including image recognition [60, 62, 72], image captioning [29, 68], and visual question answering [28, 36]. Notably, Lopez-Paz et al. [31] introduced the observational causal discovery method, which targets the causal relationships of objects in images. Wang et al. utilize the normalized weighted geometric mean (NWGM) [3] to approximate the softmax function and incorporate causal interventions to enhance visual region classification. Niu et al. [36] proposed a counterfactual inference framework designed to capture language biases as direct causal effects of questions on answers. In our approach, we innovatively integrate causal inference with diffusion, leveraging causal inference to eliminate ambiguities in the diffusion condition and enhance the robustness of the denoising process.

3. Preliminary

Diffusion models are a class of generative models for modeling high-dimensional data distributions $p(x)$. Unlike the direct estimation of $p(x)$, the diffusion model achieves modeling by estimating a score function. The diffusion model consists of a forward process and a backward process. In the forward process, the model iteratively adds a small amount of Gaussian noise at each time step, based on a specific variance scheme $\beta_1, \ldots, \beta_T$, until the image is gradually transformed into an isotropic Gaussian noise image. The process can be viewed as a Markov chain starting from $q(\mathbf{x}_0)$, where the noise is gradually accumulated at each time step to progressively perturb the input data:

$$\begin{aligned} q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) &:= \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \\ q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) &:= \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \end{aligned} \tag{1}$$

where $\mathbf{x}_0$ denotes the initial data distribution $x_0 \sim q(x)$, and $\mathbf{x}_1, \ldots, \mathbf{x}_T$ denotes a series of noise samples generated during T times of stepwise noise addition. To compute $\mathbf{x}_T$ directly from $\mathbf{x}_0$ and the fixed-value sequence $\{\beta_T \in (0,1)\}_{t=1}^{T}$, define $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$, then according to the parameter reorganization trick:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right). \tag{2}$$

In the backward process, the model recovers the original image by gradually removing the Gaussian noise added in the forward process, which can also be represented as a Markov chain:

$$\begin{aligned} p(\mathbf{x}_{T:0} \mid \mathbf{x}_T) &= \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{t+1}), \\ p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) &:= \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \textstyle\sum_\theta(\mathbf{x}_t, t)\right). \end{aligned} \tag{3}$$

The model represents the denoising process through a parameterized mean function $\mu_\theta(\mathbf{x}_t, t)$, which enables the prediction of noise from $\mathbf{x}_T$ at each time step t. The model is trained by minimizing the mean square error (MSE) loss between predicted noise and true noise:

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|_2^2\right]. \tag{4}$$

According to Song et al. [52], the above loss is equivalent to a score matching method:

$$\mathbb{E}_{p_{data}(\mathbf{x}), p_{\alpha_t}(\mathbf{x}_t|\mathbf{x}_0)}\left[\|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{\alpha_t}(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2\right]. \tag{5}$$

4. Methods

Given a top-down map $\mathbf{V} \in \mathbb{R}^{H \times W \times 3}$ and dialogs $\mathbf{I} \in \mathbb{R}^{L}$, where H and W are the height and width of the image, and

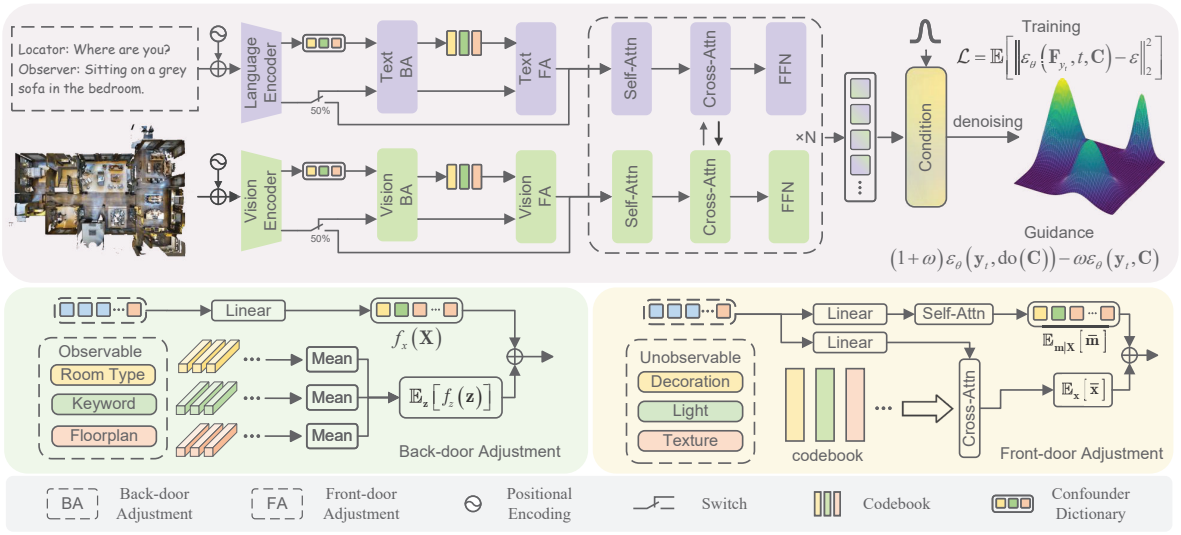

Figure 2. **Overview of the proposed CGD model.** We regress coordinates from randomly sampled Gaussian noise via diffusion, and in order to reduce the bias caused by irrelevant information, we handle observable and unobservable confounders through back-door adjustment and front-door adjustment, respectively, and use the unbiased features as guidance for the denoising process.

L is the length of tokens of the texts, the goal of embodied dialog localization is to accurately predict the coordinates of the observer's final location $\mathbf{Y}$ through multiple rounds of dialogs between the locator and the observer.

Since the diffusion model is capable of generating high-quality samples, we apply it to embodied dialog localization and propose the CGD method, as shown in Fig. 2. In this section, we first describe the proposed diffusion-based localization in detail in Sec. 4.1, then introduce the specific unified causal inference network in Sec. 4.2, and finally show how to use the de-confounded features as guidance in Sec. 4.3.

4.1. Diffusion Based Localization

Given the top-down map $\mathbf{V}$ and dialogs $\mathbf{I}$, we first encode them as tokens by pre-trained visual and textual encoders, respectively, and then project them to the same dimension, denoted as $\mathbf{F}_V \in \mathbb{R}^{N\times D}$ and $\mathbf{F}_I \in \mathbb{R}^{L\times D}$, where N is the image token length, i.e. $H \times W$. Next, the coordinates are progressively regressed through the forward and backward processes of the diffusion model, as outlined in Sec. 3.

We first sample the time step t uniformly from $\{0, \ldots, T-1\}$, where T denotes the maximum range of time steps. We treat the ground truth two-dimensional coordinates $\mathbf{y}$ as $\mathbf{y}_0$ in the diffusion model. At the sampled time step t we add independent Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to $\mathbf{y}_0$ to obtain the perturbed noise coordinates $\mathbf{y}_t$:

$$\mathbf{y}_t = \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}. \tag{6}$$

After obtaining the noise coordinates $\mathbf{y}_t$, we encode them into tokens using an MLP, denoted as $\mathbf{F}_{y_t} \in \mathbb{R}^{1\times D}$. The visual token $\mathbf{F}_V$ and the textual token $\mathbf{F}_I$ are then concatenated to form the condition $\mathbf{C} \in \mathbb{R}^{(N+L)\times D}$:

$$\mathbf{C} = \texttt{concat}\left(\mathbf{F}_V, \mathbf{F}_I\right). \tag{7}$$

We feed the condition $\mathbf{C}$ and the noise token $\mathbf{F}_{y_t}$ together into a standard Transformer. The noise token $\mathbf{F}_{y_t}$ is then extracted from the output of the Transformer and converted into the predicted noise ϵ_θ through a regression head. The model is gradually optimized during training by minimizing the mean squared error (MSE) loss between the predicted noise and the ground truth noise:

$$\mathcal{L}_{diff} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{F}_{y_t}, t, \mathbf{C}\right)\|_2^2\right]. \tag{8}$$

It should be noted that we employ Adaptive Layer Norm [41] to incorporate the time step t as a conditioning input to the model.

The diffusion model can enhance localization accuracy through multiple iterations, but its performance is heavily influenced by the condition $\mathbf{C}$. However, due to the presence of confounders in condition $\mathbf{C}$ [15, 50], the model may learn spurious correlations, which may produce wrong gradients and seriously affect the generalization performance. We will discuss the elimination of these confounders in the following subsection.

4.2. Unified Causal Inference Network

Causal inference seeks to uncover high-level causal relationships from low-level data. As illustrated in Fig. 3, the model receives inputs including top-down maps $\mathbf{V}$ and dialogue sequences $\mathbf{I}$, with the objective of predicting the coordinate $\mathbf{Y}$. Denoting the learned features as $\mathbf{F}$, conventional

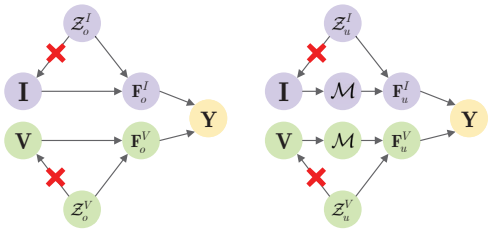

Figure 3. **Illustration of the causal graph.** $\mathbf{I}$, $\mathbf{V}$ and $\mathbf{Y}$ denote the visual inputs, language inputs, and coordinate prediction.

EDL methods only learn $\{\mathbf{I}, \mathbf{V}\} \rightarrow \mathbf{F} \rightarrow \mathbf{Y}$, which learn the ambiguous statistics-based association $p(\mathbf{Y} \mid \mathbf{I}, \mathbf{V})$, but ignore the spurious association brought by a series of confounders $\mathcal{Z}$. To strengthen the robustness of the denoising process and the generalization ability of the model, our goal is to remove both visual confounders $\mathcal{Z}^V$ and textual confounders $\mathcal{Z}^I$ in the denoising condition $\mathbf{C}$ using causal inference. Specifically, we introduce a unified causal inference network consisting of both front-door adjustment and back-door adjustment to handle observable and unobservable confounders, respectively.

Observable Confounders. For simplicity, we denote the inputs uniformly as $\mathbf{X}$. According to Bayes' theorem, the prediction without considering confounders can be expressed as:

$$p(\mathbf{Y} \mid \mathbf{X}) = \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathbf{X}, \mathbf{z})\, p(\mathbf{z} \mid \mathbf{X}) \tag{9}$$

where $p(\mathbf{z} \mid \mathbf{X})$ can cause the model to learn spurious correlations. For example, since most sofas are gray and are placed in living rooms, the model will easily learn a spurious association between "gray sofa" and "living room". Do-operator [38] provides a way to eliminate observable confounders by cutting off the backdoor-link between $\mathcal{Z}$ and $\mathbf{X}$. This process can be modeled using a neural network $f(\mathbf{X}, \mathbf{z})$, which can be expressed as follows:

$$\begin{aligned} p(\mathbf{Y} \mid \mathrm{do}(\mathbf{X})) &= \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathrm{do}(\mathbf{X}), \mathbf{z})\, p(\mathbf{z} \mid \mathrm{do}(\mathbf{X})), \\ &\approx \sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathbf{X}, \mathbf{z})\, p(\mathbf{z}), \\ &= \mathbb{E}_{\mathbf{z}}\left[f(\mathbf{X}, \mathbf{z})\right]. \end{aligned} \tag{10}$$

According to the additive property of expectations, $f(\mathbf{X}, \mathbf{z})$ can be expressed as $f_x(\mathbf{X}) + f_z(\mathbf{z})$, which allows us to represent the causal relationship as $f_x(\mathbf{X}) + \mathbb{E}_{\mathbf{z}}\left[f_z(\mathbf{z})\right]$. Next, following previous methods [30, 59], $\mathbb{E}_{\mathbf{z}}\left[f_z(\mathbf{z})\right]$ can be calculated using statistical techniques:

$$\mathbb{E}_{\mathbf{z}}\left[f_z(\mathbf{z})\right] = \sum_i \frac{|\mathbf{z}_i|}{\sum_j |\mathbf{z}_j|} f_z(\mathbf{z}_i), \tag{11}$$

where $|\mathbf{z}_i|$ denotes the count of instances of $\mathbf{z}$ belonging to the i-th category in the confounder dictionary.

We handle text and visual features independently to create the confounder dictionary. For text features, we extract spatial directions and key landmark words from conversations, then compute the average feature based on the likelihood of each word's occurrence. For visual features, we use the pre-trained VQA model BLIP [27] to obtain each room type by asking "What kind of room is this?" and then calculating the average feature for each room type. The deconfounded features $\mathbf{F}_o^V$ and $\mathbf{F}_o^I$ are derived as follows:

$$\begin{aligned} \mathbf{F}_o^V &= \mathrm{LN}\left[\phi_v\left[\mathbb{E}_{\mathbf{z}}\left[f\left(\mathbf{F}_V, \mathbf{z}\right)\right]\right]\right], \\ \mathbf{F}_o^I &= \mathrm{LN}\left[\phi_i\left[\mathbb{E}_{\mathbf{z}}\left[f\left(\mathbf{F}_I, \mathbf{z}\right)\right]\right]\right], \end{aligned} \tag{12}$$

where ϕ_v and ϕ_i denote learnable full-connection layers.

Unobservable Confounders. In the previous section, we used back-door adjustment to handle bias. However, this method requires us to identify confounders beforehand. However, there are unobservable confounders that can't be directly modeled, which also contribute to bias. To address this, front-door adjustment [39] introduces an extra mediator $\mathcal{M}$ between $\mathbf{X}$ and $\mathbf{Y}$, which creates a front-door path $\mathbf{X} \rightarrow \mathcal{M} \rightarrow \mathbf{Y}$ to transmit knowledge:

$$\begin{aligned} p(\mathbf{Y} \mid \mathrm{do}(\mathbf{X})) &= \sum_{\mathbf{m} \in \mathcal{M}} p(\mathbf{m} \mid \mathrm{do}(\mathbf{X}))\, p(\mathbf{Y} \mid \mathrm{do}(\mathbf{m})), \\ &= \sum_{\mathbf{m} \in \mathcal{M}} p(\mathbf{m} \mid \mathbf{X}) \sum_{\mathbf{x} \in \mathbf{X}} p(\mathbf{x})\, p(\mathbf{Y} \mid \mathbf{m}, \mathbf{x}), \\ &= \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{m}|\mathbf{X}}\left[p(\mathbf{Y} \mid \mathbf{m}, \mathbf{x})\right], \end{aligned} \tag{13}$$

where $\mathbf{m}$ denotes the selected knowledge from mediator $\mathcal{M}$.

Given the sensitivity of the EDL task to regions, we design $\mathcal{M}$ as a feature selector based on VQ-VAE [54]. Specifically, as both the image and text are pre-encoded into tokens, we utilize a VQ-VAE to project these tokens into the latent space individually, effectively performing an implicit clustering of features [69, 74]. Through the learned VQ-VAE, $\mathbf{V}$ and $\mathbf{I}$ are represented by corresponding discrete encoding sequences. The prior distribution for the discrete code follows a categorical distribution, which depends on the other codes within the feature map. Subsequently, the VQ-VAE model can be employed to extract key features from $\mathbf{X}$, which are then used to predict the coordinates $\mathbf{Y}$.

We use $\bar{\mathbf{m}}$ to denote the in-sampling features obtained by VQ-VAE acting on the current input, and $\bar{\mathbf{x}}$ to mean the cross-sampling features randomly sampled from the codebook of VQ-VAE. Based on the linear mapping model, Eq. (13) becomes $\mathbb{E}_{\mathbf{m}|\mathbf{X}}\left[\bar{\mathbf{m}}\right] + \mathbb{E}_{\mathbf{x}}\left[\bar{\mathbf{x}}\right]$. Following previous works [30, 68], two embedding functions are used to transmit input $\mathbf{X}$ into two query sets $\mathcal{Q}_1 = q_1(\mathbf{X})$ and

$\mathcal{Q}_2 = q_2(\mathbf{X})$. The front-door adjustment is then approximated as follows:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}[\bar{\mathbf{x}}] &\approx \sum_i \frac{\exp\left(\mathcal{Q}_1 \bar{\mathbf{x}}_i^T\right)}{\sum_j \exp\left(\mathcal{Q}_1 \bar{\mathbf{x}}_j^T\right)} \bar{\mathbf{x}}_i, \\ \mathbb{E}_{\mathbf{m}|\mathbf{X}}[\bar{\mathbf{m}}] &\approx \sum_i \frac{\exp\left(\mathcal{Q}_2 \bar{\mathbf{m}}_i^T\right)}{\sum_j \exp\left(\mathcal{Q}_2 \bar{\mathbf{m}}_j^T\right)} \bar{\mathbf{m}}_i, \\ \mathcal{F}(\mathbf{X}, \bar{\mathbf{x}}) &= \mathbb{E}_{\mathbf{x}}[\bar{\mathbf{x}}] + \mathbb{E}_{\mathbf{m}|\mathbf{X}}[\bar{\mathbf{m}}].\end{aligned} \tag{14}$$

Denoting the quantized features from VQ-VAE as $\mathbf{F}_V^q$ and $\mathbf{F}_I^q$. The final de-confounded features through back-door adjustment and front-door adjustment can be expressed as:

$$\begin{aligned}\mathbf{F}_V' &= \mathcal{F}\left(\mathbf{F}_o^V, \mathbf{F}_V^q\right), \\ \mathbf{F}_I' &= \mathcal{F}\left(\mathbf{F}_o^I, \mathbf{F}_I^q\right).\end{aligned} \tag{15}$$

4.3. Causality Guidance

To simplify notation, we represent the causal inference process from the previous section as $\text{do}(\mathbf{C})$. Classifier-free guidance [19] leverages implicit classifier gradients to adjust the gradient direction in diffusion model propagation. Inspired by this approach, implicit causal intervention can be formulated as $p^i(\text{do}(\mathbf{C}) \mid \mathbf{y}_t) \propto p(\mathbf{y}_t \mid \text{do}(\mathbf{C}))/p(\mathbf{y}_t)$ to mitigate potential adversarial gradient behavior during causal inference. As discussed in Sec. 3, the diffusion process conceptionally aligns with score matching. By obtaining accurate scores $\epsilon_\theta(\mathbf{y}_t \mid \text{do}(\mathbf{C}))$ and $\epsilon_\theta(\mathbf{y}_t)$, the gradient of this implicit intervention becomes:

$$\begin{aligned}&\nabla_{\mathbf{y}_t} \log p^i(\text{do}(\mathbf{C}) \mid \mathbf{y}_t) \\ &\qquad \propto -\frac{1}{\sigma_\lambda}\left[\epsilon_\theta(\mathbf{y}_t \mid \text{do}(\mathbf{C})) - \epsilon_\theta(\mathbf{y}_t, \mathbf{C})\right].\end{aligned} \tag{16}$$

Guidance using this implicit intervention updates the score estimate as:

$$\tilde{\epsilon}_\theta(\mathbf{y}_t, \text{do}(\mathbf{C})) = (1+\omega)\,\epsilon_\theta(\mathbf{y}_t, \text{do}(\mathbf{C})) - \omega \epsilon_\theta(\mathbf{y}_t, \mathbf{C}). \tag{17}$$

We train both the causal and non-causal models jointly by randomly setting $\mathbf{C}$ to the disambiguation condition $\text{do}(\mathbf{C})$.

4.4. Training & Inference

Training. The loss calculation for VQ-VAE is given by the following equation:

$$\mathcal{L}_{vqvae} = \left\|\text{sg}\left[\mathbf{F}_{V,I}\right] - \mathbf{F}_{V,I}^q\right\|_2^2 + \beta \left\|\mathbf{F}_{V,I} - \text{sg}\left[\mathbf{F}_{V,I}^q\right]\right\|_2^2, \tag{18}$$

where sg represents the stop-gradient operator, defined as an identity operation during the forward computation pass, but with zero partial derivatives, effectively blocking gradient flow during backpropagation. β denotes a weight coefficient. Subsequently, the total loss can be calculated as:

$$\mathcal{L}_{total} = \gamma_1 \mathcal{L}_{diff} + \gamma_2 \mathcal{L}_{vqvae}, \tag{19}$$

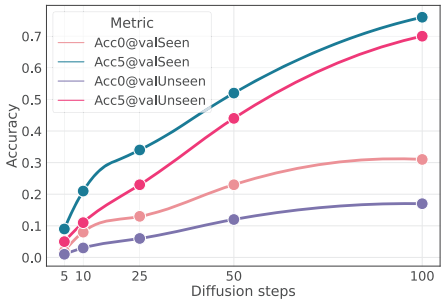

Figure 4. Ablation experiments on the number of diffusion steps.

where γ_1 and γ_2 are weight coefficients.

Inference. During inference, Since the data can be effectively approximated as a Gaussian distribution after the forward pass, we can obtain random initial coordinates $\mathbf{y}_T$ by sampling noise from a unit Gaussian. In the backward pass, we obtain the noise-free coordinates $\mathbf{y}_0$ by removing the noise predicted by the model from $\mathbf{y}_T$:

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(y_t - \frac{1-\alpha_t}{1-\bar{\alpha}_t}\tilde{\epsilon}_\theta(y_t \mid t, c)\right) + \sigma_t \delta, \tag{20}$$

where $\delta \sim \mathcal{N}(0, \mathbf{I})$.

It is worth noting that in our experiments, we found that generating the initial coordinates $\mathbf{y}_T$ solely through random sampling from the unit Gaussian can result in performance fluctuations. Since the entire denoising process aims to fit the original distribution, random sampling is akin to sampling from the original distribution, while our objective is to maximize the likelihood of the predicted coordinates. Therefore, we adopt a sampling method based on kernel density estimation (KDE): sample multiple coordinates simultaneously, fit the kernel density distribution, and select the coordinate with the highest probability likelihood as the final predicted coordinate. For details on this ablation experiment, see Sec. 5.3.

5. Experiments

5.1. Setting

Datasets. To evaluate our method, we conducted experiments on the WAY dataset. Consistent with prior approaches [18, 71], we employ a floor-level evaluation, which assumes knowledge of the ground-truth floor in the environment. We use the training split of the dataset for model training and the valSeen split for validation, while performance on the valUnseen split demonstrates the model's generalization to novel scenarios.

Metrics. Following the methodologies of DiaLoc [71] and LingUNet [18], we use geodesic distance as our evaluation metric. Specifically, we project the predicted coordinate $\mathbf{y}_0$ to the nearest node G_0 and measure its distance to the

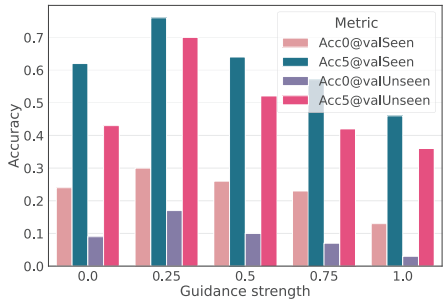

Figure 5. Ablation experiments on the guidance strength.

ground-truth node G_T, calculated as:

$$d = \|G_0 - G_T\|_2 . \tag{21}$$

We report the accuracy of d in the valSeen and valUnSeen splits at thresholds of 0 meters and 5 meters. This metric provides a clear and quantitative assessment of our model's localization performance in both familiar and unfamiliar environments.

Multi-shot setting. It is worth noting that there are many different positioning applications in reality. Therefore, similar to previous SOTA DiaLoc [71], we also apply our model to the multi-shot setting.

In the multi-shot setting, N rounds of dialogue are input to the model individually. Each time, the model processes language input from only a single round, rather than the entire dialogue sequence. This approach not only simulates real-world application scenarios but also aligns more closely with human cognitive processing habits. Specifically, the text feature $\mathbf{F}_I$ and vision feature $\mathbf{F}_V$ are expanded by an additional dimension to represent the number of dialogue rounds. During both training and inference, only the features from one dialogue round are fed into CGD at a time. For loss calculation, we apply the single-shot loss to each prediction independently.

Implementation Details. During training, the top-down map $\mathbf{V}$ is resized to 224×224 pixels, and we use CLIP-VIT [46], pretrained on ImageNet-21k [12] with a patch size of 16, as the visual backbone. To mitigate overfitting, we employ data augmentation techniques, including color jittering, random rotations up to $180°$, and random cropping of 5%. The text encoder, BERT [13], remains frozen during training to preserve its pretrained language representations. All experiments were conducted on eight NVIDIA GeForce GTX 2080 Ti GPUs with a batch size of 30. We used the AdamW [32] optimizer with a learning rate of 3×10^{-5} and a weight decay rate of 1×10^{-4}.

5.2. Benchmark Evaluation

We compare our method with several existing approaches. As shown in Tab. 1, in the single-shot setting, our method achieves significant improvements of 9.7% on Acc0@valUnseen and 29.15% on Acc5@valUnseen over the previous state-of-the-art method DiaLoc [71]. Additionally, in the multi-shot setting, our method demonstrates similar advancements, surpassing the state-of-the-art by 3.32% on Acc0@valUnseen and 15.73% on Acc5@valUnseen. It should be noted that compared with DiaLoc, our method does not generate additional data using BLIP or Chatgpt. These substantial gains underscore the effectiveness of our method in both familiar and novel environments, highlighting its robustness and ability to generalize across diverse evaluation settings.

	valSeen		valUnseen	
	Acc0	Acc5	Acc0	Acc5
Single-shot				
LingUNet	19.87	59.29	6.16	33.33
DiaLoc†	25.64	66.02	7.02	40.41
CGD(ours)	30.21	76.34	16.72	69.56
Multi-shot				
LingUNet-i	4.32	24.16	3.25	20.21
LingUNet-e	14.47	46.15	5.31	36.30
DiaLoc-i	18.43	57.18	6.42	37.89
DiaLoc-e	18.36	60.00	8.44	47.15
CGD(ours)	26.25	71.23	11.76	62.88

Table 1. Evaluations on WAY dataset under single-shot and multi-shot scenarios. DiaLoc† uses BLIP [27] to generate auxiliary location captions, DiaLoc-i and DiaLoc-e use Chatgpt to expand text data.

Id	Diff	BDA	FDA	valSeen		valUnseen	
				Acc0	Acc5	Acc0	Acc5
1	✗	✗	✗	16.51	48.64	4.17	26.32
2	✓	✗	✗	24.12	62.24	8.94	42.62
3	✗	✓	✗	17.46	50.33	4.31	29.63
4	✗	✗	✓	19.09	54.45	5.02	32.66
5	✓	✓	✗	25.66	63.21	10.11	49.23
6	✓	✗	✓	26.33	64.27	12.98	58.59
7	✗	✓	✓	21.15	57.98	8.67	42.17
8	✓	✓	✓	30.21	76.34	16.72	69.56

Table 2. Ablation experiments of key components in CGD. A simple MLP is used as regression head to ablate the effect of diffusion.

5.3. Ablation Studies

To evaluate the design of each component, we conduct extensive ablation experiments, all of which use 100 diffusion steps and 0.25 guidance strength as the baseline.

Effect of Diffusion & Causal Inference. Fig. 5 illustrates the impact of the proposed modules: back-door adjustment (BDA), front-door adjustment (FDA), and diffusion models on CGD. To isolate the specific effect of the diffusion model, we use a simple MLP for direct coordinate regression. Compared to the baseline (#1), using diffusion alone or applying either BDA or FDA individually yields performance improvements. The results indicate that

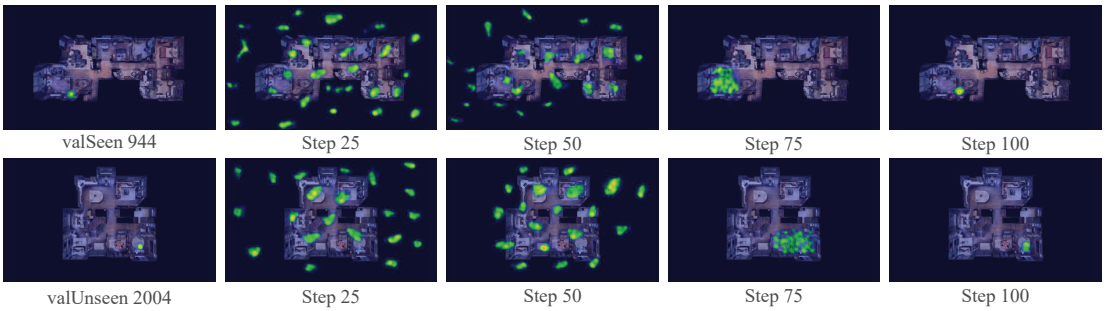

Figure 6. **Visualization of KDE as heatmap.** Illustrating how the model progressively denoises from random Gaussian noise to accurately predict coordinates as the steps increase.

Num of Kernals	valSeen		valUnseen	
	Acc0	Acc5	Acc0	Acc5
0	25.12	61.47	11.10	51.89
20	26.88	63.42	12.62	54.13
50	27.16	66.43	13.47	58.10
80	28.32	69.41	14.87	62.43
100	30.21	76.34	16.72	69.56

Table 3. Ablation experiments on different number of kernels.

regressing coordinates through a diffusion-based approach significantly enhances acc0m performance, likely due to diffusion's ability to effectively model complex data distributions. The combined application of all three modules further improves performance, particularly when BDA and FDA are used together. This finding supports our assumption about the presence of both observable and unobservable confounders, affirming that integrating back-door and front-door adjustments is essential for comprehensively addressing dataset bias and strengthening the model's robustness and generalization. Additionally, in Fig. 4, we show the effect of varying diffusion steps on overall performance.

Effect of Causality Guidance. Fig. 5 illustrates the effect of varying causality guidance strengths on performance. The results indicate that adjusting the guidance strength coefficient ω allows for flexible control over the influence of causal reasoning on the diffusion model. Our analysis suggests that a coefficient ω that is too large may hinder the model's ability to learn the original features, while a coefficient ω that is too small reduces the impact of causal reasoning.

5.4. Visualization

Since we use kernel density estimation, we can visualize the results as heatmap by calculating the distance between Gaussian kernels and each original image grid, as shown in Fig. 6. We ablated the impact of different numbers of kernels on the performance, as shown in Tab. 3. And we show the Cumulative Matching Characteristic (CMC) curves of CGD and DiaLoc in Fig. 7. CGD outperforms the baseline in both single-shot and multi-shot configurations. Especially the outstanding performance in valUnseen shows that our method achieves better performance in novel environments.

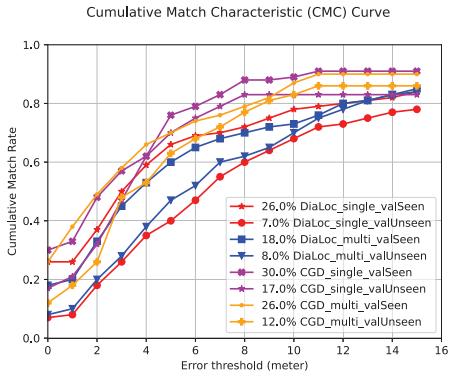

Figure 7. **Visualization of CMC curves.** We depict the CMC curves for both CGD and DiaLoc for single-shot and multi-shot settings. CGD consistently outperforms the baseline.

6. Conclusion

In this paper, we propose CGD, a novel method leveraging causality guided diffusion to enhance localization accuracy while mitigating dataset bias. Unlike prior heatmap-based methods that rely on resolution, CGD introduces controlled noise to original coordinates and reverses this process to reconstruct the original signal. Through causal learning, CGD effectively addresses both observable and unobservable confounding factors via front-door and back-door interventions, respectively. Comprehensive experiments on the WAY dataset demonstrate that our approach consistently outperforms state-of-the-art methods, underscoring its effectiveness and robustness. While CGD demonstrates excellent localization accuracy and robustness, its inference speed remains relatively slow. Future efforts will aim to enhance the speed by leveraging advanced diffusion models.

Acknowledgement

This work was supported in part by the National Key Research and Development Project under Grant 2024YFB4708100, National Natural Science Foundation of China under Grants 62088102, U24A20325 and 12326608, and Key Research and Development Plan of Shaanxi Province under Grant 2024PT-ZCK-80.

References

[1] Emmanuel Brempong Asiedu, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Decoder denoising pretraining for semantic segmentation. *arXiv preprint arXiv:2205.11423*, 2022. 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 3

[3] Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014. 3

[4] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[5] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 3

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2

[7] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 1

[8] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16537–16547, 2022. 2

[9] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023. 3

[10] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 909–919, 2023. 3

[11] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[13] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[14] 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 *Proceedings of the AAAI conference on artificial intelligence*, pages 579–587, 2023. 3

[15] Nima Fathi, Amar Kumar, Brennan Nichyporuk, Mohammad Havaei, and Tal Arbel. Decodex: Confounder detector guidance for improved diffusion-based counterfactual explanations. *arXiv preprint arXiv:2405.09288*, 2024. 4

[16] Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643, 2021. 2

[17] Meera Hahn and James M Rehg. Transformer-based localization from embodied dialog with large-scale pre-training. *arXiv preprint arXiv:2210.04864*, 2022. 1, 2

[18] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James M Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. *arXiv preprint arXiv:2011.08277*, 2020. 1, 2, 6

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 6

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 3

[22] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 1, 2

[23] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605, 2022. 3

[24] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International conference on machine learning*, pages 10362–10383. PMLR, 2022. 3

[25] Sungwon Kim, Heeseung Kim, and Sungroh Yoon. Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *arXiv preprint arXiv:2205.15370*, 2022. 3

[26] Alon Levkovitch, Eliya Nachmani, and Lior Wolf. Zero-shot voice conditioning for denoising diffusion tts models. *arXiv preprint arXiv:2206.02246*, 2022. 3

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5, 7

[28] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022. 3

[29] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022. 3

[30] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023. 5

[31] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6979–6987, 2017. 3

[32] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[33] Melvin P Manuel, Mariam Faied, Mohan Krishnan, and Mark Paulik. Robot platooning strategy for search and rescue operations. *Intelligent Service Robotics*, 15(1):57–68, 2022. 1

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[35] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4484. PMLR, 2020. 3

[36] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710, 2021. 3

[37] Seth Pate and Lawson LS Wong. "the wallpaper is ugly": Indoor localization using vision and language. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1558–1564. IEEE, 2023. 2

[38] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 3, 5

[39] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 5

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[41] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[42] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 2

[43] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023. 1

[44] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869, 2020. 3

[45] Jorge Pena Queralta, Jussi Taipalmaa, Bilge Can Pullinen, Victor Kathan Sarker, Tuan Nguyen Gia, Hannu Tenhunen, Moncef Gabbouj, Jenni Raitoharju, and Tomi Westerlund. Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision. *Ieee Access*, 8: 191617–191643, 2020. 1

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[48] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[49] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 3

[50] Tatsuhiro Shimizu. Diffusion model in causal inference with unmeasured confounders. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 683–688. IEEE, 2023. 4

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[53] Brandon Trabucco, Gunnar Sigurdsson, Robinson Piramuthu, Gaurav S Sukhatme, and Ruslan Salakhutdinov. A simple approach for visual rearrangement: 3d mapping and semantic search. *arXiv preprint arXiv:2206.13396*, 2022. 2

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 5

[55] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[56] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022. 3

[57] Liuyi Wang, Zongtao He, Ronghao Dang, Huiyi Chen, Chengju Liu, and Qijun Chen. Causality-based cross-modal representation learning for vision-and-language navigation. *arXiv preprint arXiv:2403.03405*, 2024. 3

[58] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13139–13150, 2024. 3

[59] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10760–10770, 2020. 5

[60] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 3

[61] Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3933–3948, 2022. 3

[62] Yuqing Wang, Xiangxian Li, Zhuang Qi, Jingyu Li, Xuelong Li, Xiangxu Meng, and Lei Meng. Meta-causal feature learning for out-of-distribution generalization. In *European Conference on Computer Vision*, pages 530–545. Springer, 2022. 3

[63] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762, 2024. 2

[64] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5922–5931, 2021. 2

[65] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019. 1

[66] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14958–14967, 2024. 2

[67] Qi Yan, Zhengyang Liang, Yang Song, Renjie Liao, and Lele Wang. Swingnn: Rethinking permutation invariance in diffusion models for graph generation. *arXiv preprint arXiv:2307.01646*, 2023. 3

[68] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12996–13010, 2021. 2, 3, 5

[69] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 5

[70] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019. 2

[71] Chao Zhang, Mohan Li, Ignas Budvytis, and Stephan Liwicki. Dialoc: An iterative approach to embodied dialog localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12585–12593, 2024. 1, 2, 6, 7

[72] Hua Zhang, Liqiang Xiao, Xiaochun Cao, and Hassan Foroosh. Multiple adverse weather conditions adaptation for object detection via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1742–1756, 2022. 3

[73] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020. 3

[74] Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22798–22807, 2023. 5