Customized Multi-Subject Text-to-Image Generation with Causal Tuning

Chaoyang Li, Xin Wang, *Member, IEEE*, Wenwu Zhu, *Fellow, IEEE*, Lingzhi Wang, Ning Hu, and Qing Liao, *Member, IEEE*

***Abstract*—Subject-driven text-to-image generation aims to generate customized high-fidelity images based on text descriptions for specific subjects, which has gained increasing attention. Despite recent advancements in single-subject customization, existing methods often struggle with multi-subject scenarios, leading to distortions in subject identity. This challenge arises because entangled identity-relevant and irrelevant information can obscure subject identities, and inter-subject interference can cause confusion or loss of individual identities. To address these issues, we propose CausalT2I, a customized multi-subject text-to-image generation framework with causal tuning. First, we propose a *subject-aware causal disentanglement* method, which can self-adaptively distinguish causally relevant and irrelevant information for subjects through causal intervention and a causal disentangled objective. Then, we design a *soft cross-attention guidance* strategy to mitigate interference among different subjects by aligning the textual attributes of each subject with its identity-relevant visual attributes. Last, we introduce a *causal denoising objective* to optimize the denoising process using identity-preserved textual embeddings and identity-irrelevant visual embeddings. Extensive experiments show that CausalT2I has superior generation ability in subject-driven text-to-image generation over existing baseline methods and brings more flexibility and controllability for generating customized multi-subject images.**

***Index Terms*—Multi-Subject Text-to-Image, Causal Tuning, Diffusion Models**

I. Introduction

CURRENT diffusion-based text-to-image (T2I) methods [6]–[9] can produce high-quality, detailed images from text descriptions, demonstrating significant potential for diverse multimedia and multimodal applications [10]–[12]. However, these models often struggle with personalizing imagery for unique subjects like individual pets or portraits due to the absence of such personal subjects in their large-scale pre-training data. To address this, subject-driven T2I generation techniques have gained attention [1]–[3], [5], [13], [14]. These methods use a limited set of user-provided images to generate high-fidelity customized images, such as altering styles or scenes [15]–[17].

However, existing subject-driven T2I generation methods excel with single subjects but struggle with multiple unseen subjects. On the one hand, they often face issues with entangling identity-relevant and identity-irrelevant information, leading to distorted subject identities and inadequate preservation of details [16]. As shown in the first row and first column of Fig. 1, DreamBooth [1] is tasked with generating an image based on the text description “a q* wooden pot and a s* cat play in a garden”. However, the “q* wooden pot” is over-coupled with the background information, distorting its identity and neglecting the description of “in a garden”. On the other hand, during the customized generation process, different subjects can interfere with each other, leading to the loss of certain subjects or confusion among subject attributes. As illustrated in the third row and third column of Fig. 1, Custom Diffusion [3] generates the image that loses the identity of “v* dog”. Additionally, ELITE [2], MasaCtrl [4], and Cones 2 [5] struggle with the confusion between the “s* cat” and the “v* dog” as depicted in the third row, spanning the second, fourth, and fifth columns of Fig. 1.

In the process of generating customized multi-subject images, the target subject may focus on information irrelevant to its identity [16], [18] (such as background or other subjects), leading to identity distortion in the generated images. To address this issue, we propose CausalT2I, a customized multi-subject T2I generation framework through causal tuning. It aims to disentangle causally relevant information associated with subject identities while effectively reducing interference among multiple subjects, ensuring more accurate and coherent image generation. Specifically, we first propose the *subject-aware causal disentanglement* to distinguish subjects’ identity-relevant and identity-irrelevant information from a unique causal perspective. Second, to alleviate the interference among different subjects, we design a *soft cross-attention guidance* strategy to align textual embeddings and identity-relevant visual embeddings for each subject. Third, we simultaneously use identity-preserved textual embedding and identity-irrelevant visual embedding as conditions to optimize the denoising process. Finally, extensive experiments demonstrate that CausalT2I outperforms existing state-of-the-art methods in subject-driven T2I generation, achieving exceptional results without a significant increase in fine-tuning time costs. The key contributions of this paper are summarized as follows:

- We propose CausalT2I, a customized multi-subject T2I

This work was supported in part by the National Natural Science Foundation of China under Grant 62227808, in part by the Tencent Foundation/Xplorer Prize, and in part by the Major Key Project of Peng Cheng Laboratory (PCL) under Grant PCL2024A05 (Corresponding author: Qing Liao).

Chaoyang Li and Qing Liao are with the Department of Computer Science and Technology, Harbin Institute of Technology(Shenzhen), Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: lichy@pcl.ac.cn; liaoqing@hit.edu.cn).

Xin Wang and Wenwu Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with Beijing National Research Center for Information Science and Technology, Beijing 100084, China (e-mail: xin_wang@tsinghua.edu.cn; wwzhu@tsinghua.edu.cn).

Lingzhi Wang is with the Department of Computer Science and Technology, Harbin Institute of Technology(Shenzhen), Shenzhen 518055, China (e-mail: wanglingzhi@hit.edu.cn).

Ning Hu is with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: hun@pcl.ac.cn).

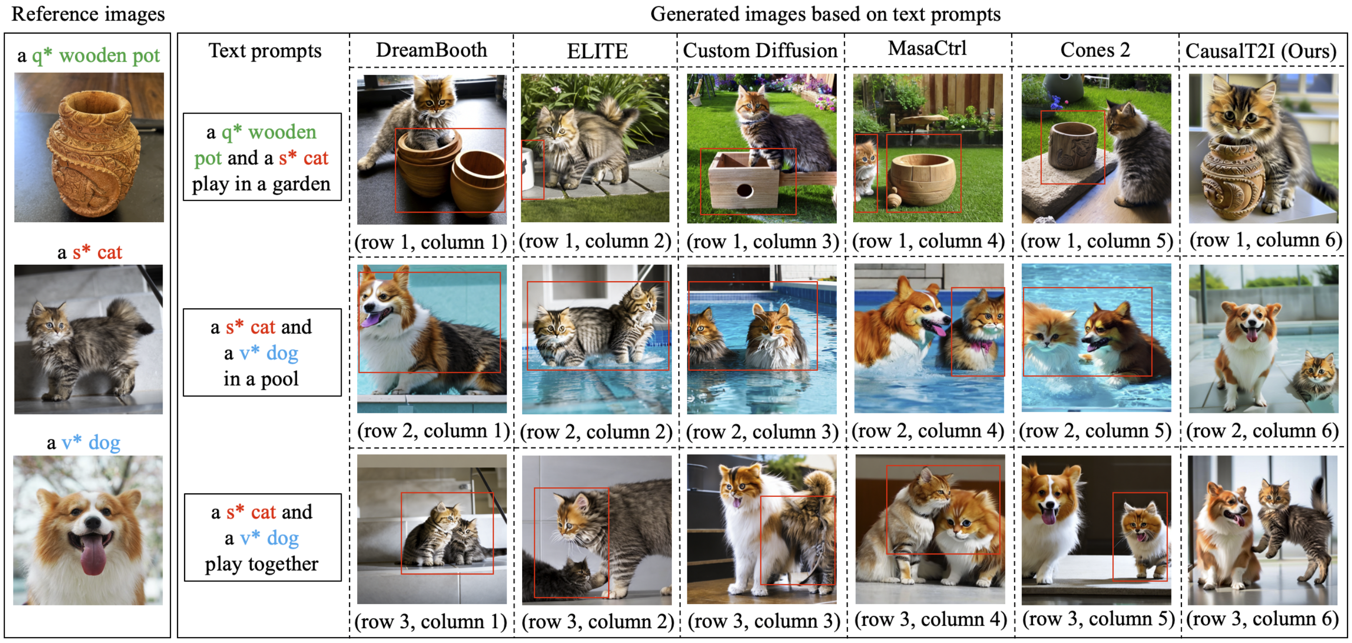


Fig. 1. Comparing between our proposed CausalT2I and the representative methods, including DreamBooth [1], ELITE [2], Custom Diffusion [3], MasaCtrl [4], and Cones 2 [5]. The special tokens "q* wooden pot", "s* cat", and "v* dog" are placeholders for three subject identities. It exhibits limitations of such baseline methods in generating customized multi-subject images, struggling with the distortion of subject identity information.

generation framework, which can preserve subject identity information and mitigate inter-subject interference from a unique causal perspective.

- We propose *subject-aware causal disentanglement*, which self-adaptively distinguishes causally relevant and irrelevant information related to subject identities through causal intervention and the causal disentangled objective.
- We design a *soft cross-attention guidance objective* to mitigate interference among multiple subjects by aligning the textual attributes of each subject with its identity-relevant visual attributes.
- Extensive experiments show that CausalT2I has superior generation ability in subject-driven T2I generation over existing baseline methods, creating more flexibility and controllability for multi-subject image generation.

The rest of this paper is organized as follows: In Section II, we review the literature on T2I generation. In Section III, we present the details of CausalT2I. In Section IV, we conduct experiments to evaluate the effectiveness of CausalT2I. Section V discusses the limitations of CausalT2I and provides the resolutions. Finally, Section VI concludes this paper.

II. Related Work

A. Text-to-Image Generation

Text-to-image generation has emerged as an interesting technique, leveraging textual prompts to create realistic images [19], [20]. Recent advancements, particularly with Generative Adversarial Networks [21]–[28], have revolutionized semantic image synthesis. For instance, GALIP [24] integrates the pre-trained CLIP [29] model for controllable T2I synthesis. Diffusion models, pioneered by [30]–[34], have further propelled the field, with notable works like DALLE-2 [35], Imagen [36], Stable Diffusion [37], Re-Imagen [38], Uni-ControlNet [39], MultiDiffusion [40], Isolated Diffusion [41], and Attend-and-Excite [42] generating more realistic and diverse images from text. For example, Re-Imagen [38] excels with its retrieval-augmented approach, accurately reflecting textual descriptions. However, such models struggle to customize unseen subjects from users' personal lives. Subject-driven T2I generation, which focuses on creating personalized visuals for unseen subjects, has attracted considerable research interest [43], [44].

B. Subject-Driven Text-to-Image Generation

Subject-driven T2I generation aims to create customized images of unseen subjects using a limited set of reference images, which mainly includes two categories: single-subject and multi-subject T2I generation methods.

1) Customized Single-subject T2I generation: Early subject-driven T2I methods mainly focused on single-subject generation [45]–[48]. Techniques vary widely, from mapping image features or reference images into the textual embedding space [2], [49], to fine-tuning models with identifier tokens [1], [46], and leveraging disentangled or concatenated embeddings to preserve identity while handling irrelevant variations [14], [16], [50], [51]. Tuning-free approaches such as MasaCtrl [4], ConsiStory [52], and Customization Assistant [53] improve efficiency and user interaction through attention mechanisms, shared activations, or pre-trained backbones. More recent efforts aim to enhance identity fidelity and generalization, including HyperDreamBooth's hypernetwork architecture [54], CustomContrast's contrastive disentanglement of subject attributes [55], and JeDi's joint modeling of multiple subject-related prompts [56]. Additionally, MetaCloak [57] addresses privacy concerns by introducing subject-irrelevant perturbations via meta-learning, although its protection mechanism often comes at the cost of generation quality.

2) Customized Multi-subject T2I generation: As customized generation techniques advance, increasing attention has shifted toward the more complex task of multi-subject generation [3], [5], [58], [59]. Many approaches enhance subject representation through various supervision signals. For instance, FastComposer [60], Subject-Diffusion [43], and FreeCustom [58] rely on pre-constructed datasets with subject masks or bounding boxes, using attention control or weighted mask strategies to handle multiple subjects. Similarly, SpaText [61] and MuDI [62] leverage segmentation maps or semantic masks for spatial control, though their reliance on external segmentation models can limit end-to-end efficiency.

Some methods integrate advanced model architectures or optimization strategies. Custom Diffusion [3] fine-tunes cross-attention layers, while SVDiff [13] introduces a Cut-Mix-Unmix augmentation strategy. Mix-of-Show [63] adopts decomposed LoRA modules for attribute disentanglement and controllable sampling. Cones [17] and Cones 2 [5] identify subject-relevant neurons or residual embeddings with layout guidance to preserve subject identity. OMG [64] fuses latent noise across concepts, and Modular Customization [18] achieves disentanglement via orthogonal decomposition. Concept Weaver [59] introduces a two-step process of template generation followed by personalized concept fusion.

In broader applications, DreamStory [65] enables training-free, story-consistent generation using large language models, while ConceptGuard [66] addresses continual customization by mitigating catastrophic forgetting through shift embeddings and regularization techniques.

However, most existing subject-driven T2I generation methods still face challenges in multi-subject scenarios. They struggle with over-coupling of identity-relevant and identity-irrelevant information, as well as interference among different subjects, as illustrated in Fig. 1. In contrast to these methods, we propose CausalT2I to disentangle causally relevant information for subject identities while mitigating interference among multiple subjects.

III. Method

In this section, we present the motivations behind and the details of our proposed CausalT2I framework. The important variables of the CausalT2I framework are summarized as shown in Table I.

A. Framework

Given a few reference images of subjects from different viewpoints, we aim to generate customized images for multiple subjects with vivid and precise details. To achieve this, we propose CausalT2I, a causal tuning framework for customized multi-subject T2I generation. As illustrated in Fig. 2, we first introduce *subject-aware causal disentanglement* (SaCD) to distinguish between causally relevant and irrelevant information within subjects, mitigating the interference of subject-irrelevant information. Second, to reduce interference among different subjects, we design *soft cross-attention guidance* (SCaG), which aligns the cross-attention maps of textual attributes with relevant visual embeddings. This ensures that the textual attributes of each subject focus on identity-relevant visual information. Furthermore, since reference images may contain identity-irrelevant visual details (e.g., backgrounds), we introduce the *causal denoising objective* (CdO) to optimize the denoising process by leveraging identity-preserved textual embeddings and identity-irrelevant visual embeddings. This objective maintains identity-relevant textual information while improving the model's comprehension of global visual context.

TABLE I
IMPORTANT VARIABLES.

Variable	**Description**
$\boldsymbol{\varepsilon}$	Gaussian noise
$\boldsymbol{z}_t$	Latent representation with noise at the time step t
$\boldsymbol{f}_i^s(i=1,2,...,n)$	Textual condition embedding of the i-th subject
$\boldsymbol{f}_{n+1}^s$	Textual condition embedding of merged subjects
$\boldsymbol{f}^m$	Visual embedding of merged image
$\mathcal{M}=\{\boldsymbol{M}_i\}_{i=1}^{n+1}$	Set of learnable masks
$\mathcal{P}=\{\mathcal{P}_i\}_{i=1}^{n+1}$	Set of text prompts
$\boldsymbol{f}_i/\bar{\boldsymbol{f}}_i(i=1,2,...,n)$	Retained/filtered-out portions of visual embeddings for the i-th subject
$\boldsymbol{f}_{n+1}/\bar{\boldsymbol{f}}_{n+1}$	Retained/filtered-out portions of visual embeddings for merged subjects
$\boldsymbol{Atten}_i$	Cross-attention embedding of the i-th subject

B. Preliminaries

1) Stable Diffusion Models: In this paper, we adopt Stable Diffusion (SD) as the foundational model, built upon the Latent Diffusion Model (LDM) [37]. SD is a large text-to-image model pre-trained on large-scale text-image pairs $\{(P,\boldsymbol{x})\}$, where P is the text prompt for the image $\boldsymbol{x}$. For a given image $\boldsymbol{x}$, SD initially uses the encoder $\mathcal{E}$ of VAE [67] to project $\boldsymbol{x}$ into a latent representation $\boldsymbol{z}=\mathcal{E}(\boldsymbol{x})$. The diffusion forward process is subsequently executed on the latent representation by introducing noise $\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\mathbf{I})$ into $\boldsymbol{z}$, forming a fixed-length Markov Chain denoted as $\{\boldsymbol{z}_1,...,\boldsymbol{z}_T\}$, where T signifies the length of the chain, $\boldsymbol{z}_t=\alpha_t\boldsymbol{z}+\sigma_t\boldsymbol{\varepsilon}, t\in[1,T]$, α_t and σ_t are the coefficients that control the noise schedule at the time step t. The LDM is trained with the following objective for denoising [30], [31]:

$$\mathcal{L}_{ldm}=\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon},t,P}[\|\,\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t,t,\ E_T(P))\,\|_2^2], \tag{1}$$

where $\boldsymbol{\varepsilon}_\theta$ denotes the U-Net model that predicts the noise by taking the noisy latent $\boldsymbol{z}_t$, the text conditional embedding $E_T(P)$ obtained by the text encoder E_T [29], and the time step t as input.

The intermediate representation $E_T(P)$ is linked to the intermediate layers of the U-Net through cross-attention layers using the following mapping:

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\cdot\boldsymbol{V}, \tag{2}$$

$$\boldsymbol{Q}=\boldsymbol{W}_{\boldsymbol{Q}}^{(i)}\cdot\varphi_i\left(\boldsymbol{z}_t\right),\boldsymbol{K}=\boldsymbol{W}_{\boldsymbol{K}}^{(i)}\cdot E_T\left(P\right),\boldsymbol{V}=\boldsymbol{W}_{\boldsymbol{V}}^{(i)}\cdot E_T\left(P\right), \tag{3}$$

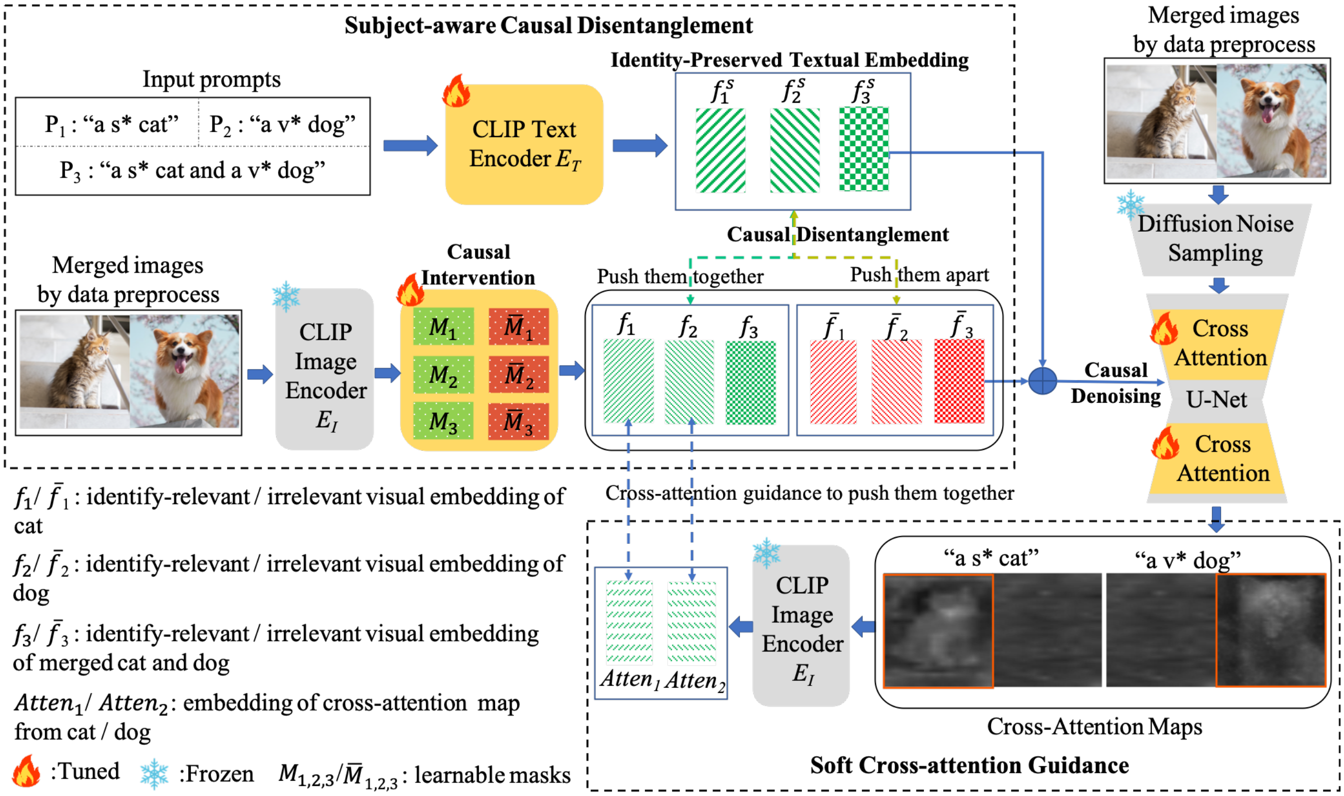

Subject-aware Causal Disentanglement
Input prompts
P$_1$: "a s* cat"
P$_2$: "a v* dog"
P$_3$: "a s* cat and a v* dog"
CLIP Text Encoder E_T
Identity-Preserved Textual Embedding
f_1^s f_2^s f_3^s
Merged images by data preprocess
CLIP Image Encoder E_I
Causal Intervention
M_1 $\bar{M}_1$ M_2 $\bar{M}_2$ M_3 $\bar{M}_3$
Causal Disentanglement
Push them together
Push them apart
f_1 f_2 f_3 $\bar{f}_1$ $\bar{f}_2$ $\bar{f}_3$
Causal Denoising
Merged images by data preprocess
Diffusion Noise Sampling
Cross Attention
U-Net
Cross Attention
Cross-attention guidance to push them together
$Atten_1$ $Atten_2$
CLIP Image Encoder E_I
"a s* cat"
"a v* dog"
Cross-Attention Maps
Soft Cross-attention Guidance
f_1/ $\bar{f}_1$: identify-relevant / irrelevant visual embedding of cat
f_2/ $\bar{f}_2$: identify-relevant / irrelevant visual embedding of dog
f_3/ $\bar{f}_3$: identify-relevant / irrelevant visual embedding of merged cat and dog
$Atten_1$/ $Atten_2$: embedding of cross-attention map from cat / dog
:Tuned
:Frozen
$M_{1,2,3}/\bar{M}_{1,2,3}$: learnable masks

Fig. 2. The framework of CausalT2I. CausalT2I first merges images of different subjects into an image and combines their prompts. Each merged image and combined prompt is used to conduct the denoising process. Second, the *subject-aware causal disentanglement* gets the textual embedding $f_{1,2,3}^s$ to preserve subject identities. $f_{1,2,3}^s$ are used as supervision signals to distinguish the identity-relevant and identity-irrelevant visual embeddings. Third, the *soft cross-attention guidance* weakly aligns each subject cross-attention map and its identity-relevant visual embedding. Additionally, the *causal denoising objective* is to optimize the denoising process on each merged image with the identity-preserved textual embedding f_3^s and identity-irrelevant visual embedding $\bar{f}_3$.

where d is the output dimension of the query $(\boldsymbol{Q})$ and key $(\boldsymbol{K})$ features. $\boldsymbol{W}_{\boldsymbol{Q}}^{(i)}$, $\boldsymbol{W}_{\boldsymbol{K}}^{(i)}$, and $\boldsymbol{W}_{\boldsymbol{V}}^{(i)}$ are learnable projection matrices in the i-th cross-attention layer. $\varphi_i(\boldsymbol{z}_t)$ is a flattened intermediate representation of the noisy latent $\boldsymbol{z}_t$. The cross-attention map at the i-th layer is given by:

$$\boldsymbol{Atten}^{(i)} = \text{softmax}\left(\frac{\boldsymbol{Q}^{(i)}\boldsymbol{K}^{(i)^T}}{\sqrt{d}}\right). \tag{4}$$

2) Granger-causal objective: In causal inference, causal effects are used to quantify the causal relationships between variables [68]. To calculate the causal effect of an observed variable on a model's prediction, most works typically perform a causal intervention on the variable and measure the resulting change in the model's prediction [69], [70].

Granger-causal objective [71], [72] is a classical method for measuring causal relationships, which quantifies the effect of an input feature on the performance of model f. Given an input image $\boldsymbol{I} = \{\boldsymbol{I}_i\}_{i=1}^m$ with m patch features, the model prediction is $\hat{\boldsymbol{y}} = f(\boldsymbol{I})$. If the i-th patch feature is removed, the input becomes $\boldsymbol{I}_{\neg i}$ and the corresponding prediction is $\hat{\boldsymbol{y}}_{\neg i} = f(\boldsymbol{I}_{\neg i})$. The causal effect of the i-th feature is measured by the loss difference between the model predictions with and without that feature, formulated by:

$$\Delta(\boldsymbol{I}_i) = \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}}_{\neg i}) - \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}}), \tag{5}$$

where $\mathcal{L}$ is the loss function and $\boldsymbol{y}$ is the ground-truth label. A positive $\Delta(\boldsymbol{I}_i)$ indicates a causal relationship between the feature and the prediction, while a negative $\Delta(\boldsymbol{I}_i)$ implies no causal relationship.

C. Subject-aware Causal Disentanglement

The identity-irrelevant visual information in reference images may be entangled with the subject's identity. This can cause text prompts to focus on these irrelevant details, distorting subject identities. To address this issue, we propose *subject-aware causal disentanglement*, which comprises three components: identity-preserved textual embedding, causal intervention for visual embedding, and a causal disentangled objective. Identity-preserved textual embedding learns textual information that captures the subject's identity. Causal intervention for visual embedding separates visual information using learnable masks. Finally, the causal disentangled objective applies a contrastive cosine loss between the identity-preserved textual embedding and the disentangled visual embeddings, capturing the visual information causally relevant to the subject's identity and alleviating interference from irrelevant visual information.

1) Identity-Preserved Textual Embedding: Text prompts containing subject identifier tokens can be used as supervisory signals to guide the image generation process and enhance the subject's identity information [1], [3], [16], [46]. Motivated by these insights, to obtain the identity-preserved textual embedding, we use the CLIP text encoder E_T to encode the subject identity with the identifier token. Specifically, given n subjects, the text prompts incorporate multiple identifier tokens to distinguish the identities of various subjects, such as "s*" and "v*". We can get the text prompts $\mathcal{P} = \{\mathcal{P}_1, ..., \mathcal{P}_n, \mathcal{P}_{n+1}\}$, where $\mathcal{P}_i (i = 1, 2, ..., n)$ represents the text prompt of the i-th subject and $\mathcal{P}_{n+1}$ denotes the merged text prompt of all subjects. Using Fig. 2 as an example, for the two subjects of "cat" and "dog", there are three text prompts: "a s* cat", "a v* dog", and "a s* cat and a v* dog". Finally, the identity-preserved textual embedding $\boldsymbol{f}_i^s$ can be formulated by,

$$\boldsymbol{f}_i^s = E_T\left(\mathcal{P}_i\right), \quad i = 1, 2, ..., n+1, \tag{6}$$

where $\boldsymbol{f}_i^s (i = 1, 2, ..., n)$ represents the textual condition embedding of the i-th subject identity and $\boldsymbol{f}_{n+1}^s$ denotes the textual condition embedding of merged subject identities.

2) Causal Intervention for Visual Embedding: To alleviate the interference of identity-irrelevant visual information in the reference images, we design a simple and effective causal intervention method to separate identity-relevant and identity-irrelevant visual embeddings. Given n subjects, to better learn the spatial layout information of multiple subjects within the same image, we first randomly merge the reference images of different subjects to obtain $\boldsymbol{x}^m = \text{merge}\left\{\boldsymbol{x}^{s_1}, ..., \boldsymbol{x}^{s_n}\right\}$. It is noticed that the aspect ratios of the subjects remain unchanged before and after merging. The merged image is then passed through the pre-trained CLIP image encoder E_I to obtain the embedding $\boldsymbol{f}^m = E_I\left(\boldsymbol{x}^m\right)$.

Furthermore, to filter out the identity-relevant and identity-irrelevant visual embeddings from $\boldsymbol{f}^m$, we do causal intervention by designing subject-specific learnable masks $\mathcal{M} = \{\boldsymbol{M}_i\}_{i=1}^{n+1}$ with the same dimension as the embedding $\boldsymbol{f}^m$, whose element values belong to $(0, 1)$. Therefore, we obtain the counterfactual embedding (i.e., $\boldsymbol{M}_i * \boldsymbol{f}^m$) of the i-th subject and the merged counterfactual embedding (i.e., $\boldsymbol{M}_{n+1} * \boldsymbol{f}^m$) by the element-wise product between the mask and the pre-trained embedding. In addition, during the SD pre-training stage, the text encoder is pre-trained with the U-Net, while the image encoder is not jointly trained [16], [37]. There is often a distribution shift between the visual embedding obtained by the image encoder and the latent representation in SD. So, we use the MLP with skip connection to transform counterfactual embedding into the same space as the textual embedding $\boldsymbol{f}_s$, formulated as follows,

$$\boldsymbol{f}_i = \text{MLP}(\boldsymbol{M}_i * \boldsymbol{f}^m + \text{MLP}(\boldsymbol{M}_i * \boldsymbol{f}^m)), i = 1, 2, ..., n+1, \tag{7}$$

$$\bar{\boldsymbol{f}}_i = \text{MLP}(\bar{\boldsymbol{M}}_i * \boldsymbol{f}^m + \text{MLP}(\bar{\boldsymbol{M}}_i * \boldsymbol{f}^m)), i = 1, 2, ..., n+1, \tag{8}$$

where $\bar{\boldsymbol{M}}_i = \mathbf{1} - \boldsymbol{M}_i$ and $\mathbf{1}$ is a matrix of all ones. Through this causal intervention, $\boldsymbol{f}_i$ and $\bar{\boldsymbol{f}}_i$ represent the retained and filtered-out portions of the visual embeddings for the i-th subject, respectively. Similarly, $\boldsymbol{f}_{n+1}$ and $\bar{\boldsymbol{f}}_{n+1}$ represent the retained and filtered-out portions of the visual embeddings for the merged subjects, respectively.

3) Causal Disentangled Objective: Drawing inspiration from the Granger-causal objective [71], [72], which measures causal relationships by computing the loss difference between model predictions with and without a specific feature, we propose a causal disentangled objective. This objective contrasts cosine similarity to capture the causal relationship between the identity-preserved textual embedding $\boldsymbol{f}_i^s$ and the disentangled visual embeddings ($\boldsymbol{f}_i$ and $\bar{\boldsymbol{f}}_i$), thereby distinguishing the causally relevant visual embeddings for subjects. Specifically, the causal disentangled objective uses the cosine similarity between $\bar{\boldsymbol{f}}_i$ and $\boldsymbol{f}_i^s$ as the numerator, while the cosine similarity between $\boldsymbol{f}_i$ and $\boldsymbol{f}_i^s$ serves as the denominator, as formulated below,

$$\mathcal{L}_{cst} = \alpha \sum_{i=1}^{n+1} \frac{\text{cosine}\left(\boldsymbol{f}_i^s, \bar{\boldsymbol{f}}_i\right)}{\text{cosine}\left(\boldsymbol{f}_i^s, \boldsymbol{f}_i\right)}, \tag{9}$$

where α is a hyper-parameter. This ratio of contrastive objective reflects the causal relationship between the disentangled visual embeddings and the identity-preserved textual embedding. During training, as the causal disentangled objective decreases (i.e., as the numerator shrinks and the denominator increases), the mask $\mathcal{M}$ increasingly focuses on the visual embeddings that are relevant to the subject's identity, strengthening the causal relationship between the textual embedding $\boldsymbol{f}_i^s$ and the remaining visual embedding $\boldsymbol{f}_i$. This process helps to disentangle identity-relevant from identity-irrelevant information in $\boldsymbol{x}$, facilitating a more accurate causal relationship.

D. Soft Cross-attention Guidance

To mitigate interference among different subjects, we propose a *soft cross-attention guidance* strategy that aligns each subject's cross-attention maps of textual attributes with its relevant visual information. This alignment enhances the model's understanding of positional information while reducing the target subject's focus on irrelevant details from other subjects.

Based on the data preprocessing operations in the subsection III-C, we can obtain a merged image $\boldsymbol{x}$ and the merged text prompt $\mathcal{P}_{n+1} = \{\mathcal{V}_j\}_{j=1}^{j=l}$ containing each subject's preserved prompt $\mathcal{P}_i (i = 1, 2, ..., n)$, where $\mathcal{V}_j$ is the j-th token and l is the number of tokens in the merged prompt. Taking Fig. 2 as an example, the merged text prompt $\mathcal{P}_3$ (i.e., "a s* cat and a v* dog") contains subject preserved prompts $\mathcal{P}_1$ (i.e., "a s* cat") and $\mathcal{P}_2$ (i.e., "a v* dog"). SD uses the pre-trained U-Net with $\boldsymbol{z}_t$ and $\mathcal{P}_{n+1}$ for the diffusion forward process. We can obtain the resulting cross-attention map after averaging all attention layers and heads in the U-Net. The resulting aggregated map $\boldsymbol{Atten}_{\mathcal{V}_j} (j = 1, 2.., l)$ contains K spatial attention maps [15], [73], one for each token of $\mathcal{V}_j$. We aim to extract a spatial attention map for each token $\mathcal{V}_j \in \mathcal{P}_i$, indicating the influence of $\mathcal{V}_j$ on each patch of the i-th subject in the merged image. A single patch with a high attention value could stem from partial information passed from the token $\mathcal{V}_j$. This may occur when the model does not generate the full subject but rather a patch that resembles some part of the subject. In the case of guidance, for the i-th subject, we aggregate all the cross-attention maps $\boldsymbol{Atten}_{\mathcal{V}_j} (\mathcal{V}_j \in \mathcal{P}_i)$ to get its cross-attention embedding $\boldsymbol{Atten}_i$ with the CLIP image encoder E_I,

$$\boldsymbol{Atten}_i = E_I\left(\frac{1}{|\mathcal{P}_i|} \sum_{\mathcal{V}_j \in \mathcal{P}_i} \boldsymbol{Atten}_{\mathcal{V}_j}\right). \tag{10}$$

An intuitive way for aligning textual and visual attributes of each subject is to directly calculate the mean square error (MSE) between $\boldsymbol{f}_i$ and $\boldsymbol{Atten}_i$. However, this may be suboptimal since the cross-attention embedding $\boldsymbol{Atten}_i$ does not contain the fine-grained information of the subject region. To address this, we design a soft cross-attention guidance loss $\mathcal{L}_{attn}$ by introducing a learnable parameter set $\mathcal{B} = \{\boldsymbol{\beta}_i\}_{i=1}^{n}$ to optimize the alignment process between each pair of embeddings, guiding each subject to focus on its visual region. The $\mathcal{L}_{attn}$ can be formulated as follows,

$$\mathcal{L}_{attn} = \sum_{i=1}^{n} \frac{1}{hw} \sum_{j=1}^{h} \sum_{k=1}^{w} \left(\boldsymbol{f}_i\left(j,k\right) - \boldsymbol{Atten}_i\left(j,k\right)\right)^2 \cdot \boldsymbol{\gamma}_i\left(j,k\right), \tag{11}$$

where $\boldsymbol{\gamma}_i\left(j,k\right) = \sigma\left(\boldsymbol{\beta}_i\left(j,k\right)\right)$, σ is the sigmoid function, $\boldsymbol{\beta}_i \in \mathbb{R}^{h \times w}$, $\boldsymbol{\gamma}_i \in \mathbb{R}^{h \times w}$, $h \times w$ is the dimension of embeddings. By optimizing learnable parameters $\mathcal{B}$, the cross-attention map can be adaptively optimized by the causally disentangled embeddings, which can achieve weak alignment between each pair of text and visual attributes.

E. Causal Denoising Objective

Since the Stable Diffusion model denoises all information within reference images, including both identity-relevant and identity-irrelevant embeddings (e.g., background), we propose using the sum of the identity-preserving textual embedding and the identity-irrelevant visual embedding as the condition to denoise each merged image. This allows the model to better capture the global visual context [16]. Specifically, by utilizing the extracted identity-preserved textual embedding $\boldsymbol{f}_{n+1}^{s}$ and the disentangled identity-irrelevant visual embeddings $\bar{\boldsymbol{f}}_{n+1}$, we can optimize the fine-tuning process with the causal denoising objective, as follows:

$$\mathcal{L}_{cdm} = \mathbb{E}_{\boldsymbol{z}=\mathcal{E}(\boldsymbol{x}),\boldsymbol{x}\sim C_s,\boldsymbol{\varepsilon},t}\left[\| \, \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{f}_{n+1}^{s} + \bar{\boldsymbol{f}}_{n+1}) \, \|_2^2\right], \tag{12}$$

where C_s is the set of reference images. The identity-preserved textual embedding $\boldsymbol{f}_{n+1}^{s}$ is shared when denoising all images, capturing the common identity information. Since each image also has its image-specific identity-irrelevant embedding, using $\boldsymbol{f}_{n+1}^{s} + \bar{\boldsymbol{f}}_{n+1}$ as the condition helps preserve certain characteristics of the input images. By combining the textual identity-preserved embedding with the identity-irrelevant visual embedding, this way can provide a more flexible and controllable image generation process [16].

F. Total Objective and Inference

With the previous causal disentangle, soft cross-attention guidance, and the causal denoising objectives, the total tuning objective is as follows:

$$\mathcal{L} = \mathcal{L}_{cdm} + \mathcal{L}_{attn} + \mathcal{L}_{cst}. \tag{13}$$

As shown in Fig. 2, during the training process of CausalT2I, only the parameters of the causal intervention network, the cross-attention layers of the U-Net, and the text encoder of the CLIP model are trainable, while the other parameters remain frozen. The total objective is minimized to optimize the cross-attention layers of the U-Net and the text encoder of the CLIP model, specifically for the subject identifier's textual embedding. The causal intervention network aids in optimizing these parameters; it is not involved during inference. In the inference, CausalT2I loads the Stable Diffusion model and updates the cross-attention layer parameters of the U-Net and the text encoder parameters of the CLIP model based on the subject identifier trained.

IV. Experiment

In this section, we conduct experiments to answer the following questions:

RQ1: How does CausalT2I perform in customized single-subject T2I generation compared to previous methods?

RQ2: How does CausalT2I perform in customized multi-subject T2I generation compared to previous methods?

RQ3: What is the human preference study?

RQ4: Do the key components in CausalT2I contribute to improving the model performance (Ablation studies)?

RQ5: How much is the fine-tuning time cost of CausalT2I?

A. Experimental setup

1) Dataset: Most existing subject-driven T2I methods are evaluated on limited datasets. Some studies use as few as 15 subjects [5], [59], [64]. While Custom Diffusion [3] involves more subjects, it focuses only on single- and two-subject tasks. In contrast, Cones 2 [5] explores more complex three- and four-subject scenarios with richer prompt templates. To enrich subjects and enable more complex multi-subject experiments, we extend the dataset of Cones 2 by adding 15 subjects, resulting in 30 subjects encompassing three scene categories, six live subjects/pets, eighteen objects, and three human subjects. These additional subjects are selected from previous works [1], [3], [46] and downloaded from Unsplash [1]. For the single-subject generation, each subject is paired with 20 prompts, generating 50 images per prompt, resulting in 30,000 evaluation images. For the two-subject generation, 18 subject groups are selected, each assigned 10 prompts with 50 images per prompt, totaling 9,000 evaluation images. For the three-subject generation, 7 groups are chosen, each with 10 prompts and 50 images per prompt, yielding 3,500 evaluation images. For the four-subject generation, 6 groups are used, each with 5 prompts and 50 images per prompt, producing 1,500 evaluation images.

2) Baselines: We compare our method with six baselines, i.e., DreamBooth [1] (third-party implementation [74]), ELITE [2], DisenDreamer [50] , Custom Diffusion [3], MasaCtrl [4] and Cones 2 [5]. Custom Diffusion, MasaCtrl, and Cones 2 can be directly applied to the multi-subject task. MasaCtrl performs inference on a pre-trained T2I-Adapter [75] model. DreamBooth, ELITE, and DisenDreamer are mainly compared with the single-subject task. They are also auxiliary compared to the multi-subject generation. In practice, DreamBooth and DisenDreamer follow the data processing in Custom Diffusion, packaging the reference images and identifier tokens for each subject into a unified data loader. ELITE directly infers the multi-subject target images.

[1] https://unsplash.com/

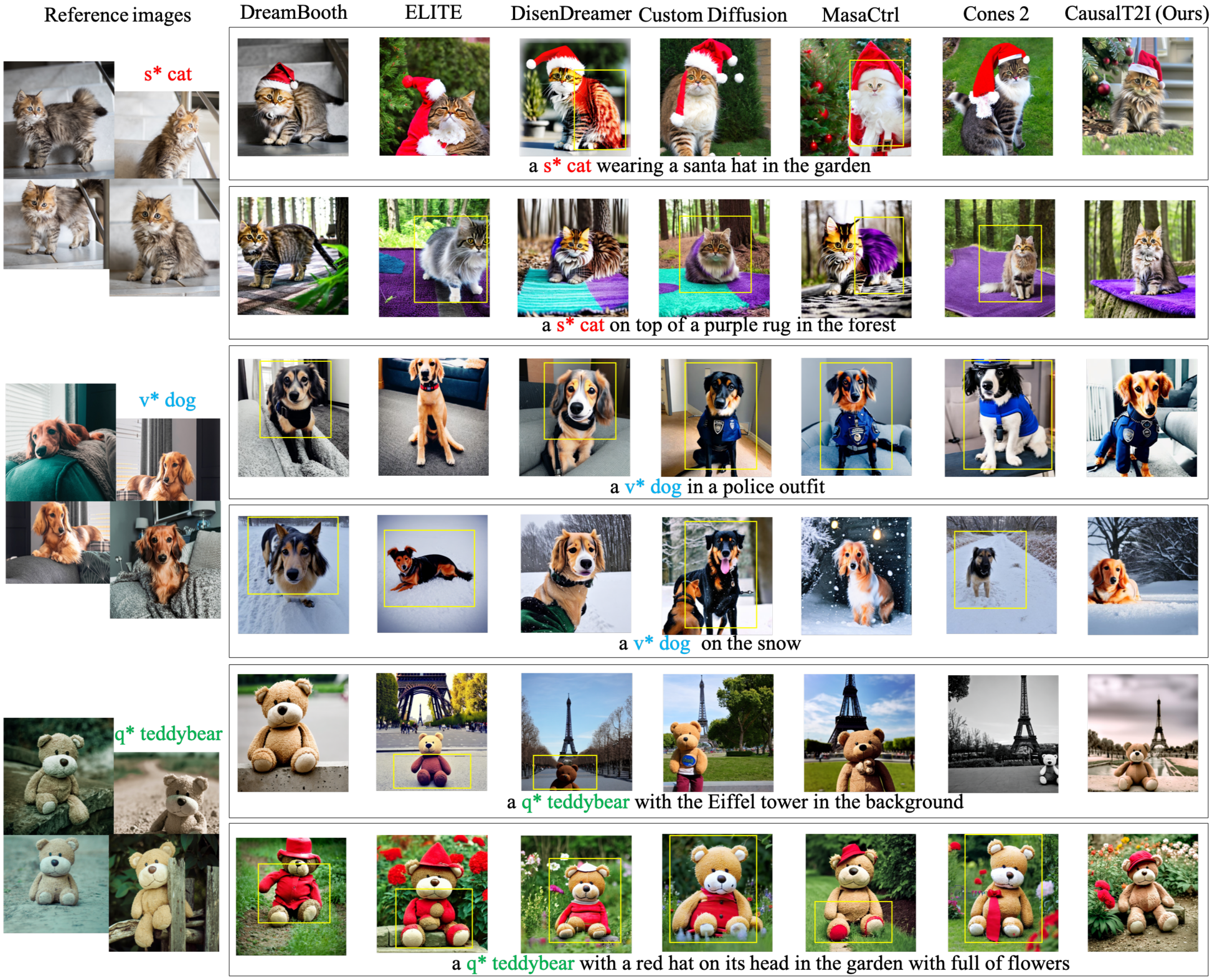

Reference images
DreamBooth
ELITE
DisenDreamer
Custom Diffusion
MasaCtrl
Cones 2
CausalT2I (Ours)
s* cat
a s* cat wearing a santa hat in the garden
a s* cat on top of a purple rug in the forest
v* dog
a v* dog in a police outfit
a v* dog on the snow
q* teddybear
a q* teddybear with the Eiffel tower in the background
a q* teddybear with a red hat on its head in the garden with full of flowers

Fig. 3. Generated images of a single subject given different text prompts with various methods.

3) Evaluation metrics: Let $\hat{\mathcal{X}} = \{\hat{x}_i\}_{i=1}^N$ be the set of generated images and $\mathcal{X} = \{x_j\}_{j=1}^M$ be the set of reference images, where each x_j corresponds to a specific subject (in single-subject generation) or subject group (in multi-subject generation). We evaluate our method against the baselines following three aspects:

(i) Subject Fidelity: This assesses whether the generated images preserve the identity of the target subjects [1], [3], [5], [17], [50]. To evaluate subject fidelity, we adopt two standard metrics: CLIP-I [29] and KID [76]. CLIP-I measures the average cosine similarity between generated images and their reference images using the CLIP image encoder. KID measures the distributional difference between real and generated images based on deep visual features extracted by a pretrained Inception network. The CLIP-I and KID can be formulated as follows:

$$\text{CLIP-I} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\text{cosine}\left(E_I(\hat{x}_i),\ E_I(x_j)\right), \qquad (14)$$

where E_I is the CLIP image encoder.

$$\begin{aligned}\text{KID} = &\frac{2}{M(M-1)}\sum_{j=1}^{M}\sum_{j'=j+1}^{M} k\left(E_F(\boldsymbol{x}_j), E_F(\boldsymbol{x}_{j'})\right)\\ &+\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{i'=i+1}^{N} k\left(E_F(\hat{\boldsymbol{x}}_i), E_F(\hat{\boldsymbol{x}}_{i'})\right) \qquad (15)\\ &-\frac{2}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} k\left(E_F(\hat{\boldsymbol{x}}_i), E_F(\boldsymbol{x}_j)\right),\end{aligned}$$

where E_F is the Inception feature extractor, and $k(\cdot,\cdot)$ is a polynomial kernel with degree 3. For multi-subject generation, we compute the CLIP-I and KID between each generated image and each subject separately, and then take the average score [3], [5], [17]. A higher CLIP-I and a lower KID score mean that the generated images have higher image fidelity to the reference subjects.

(ii) Image-text Alignment: This evaluates how well the generated images align with the text prompt [1], [3], [5],

TABLE II
QUANTITATIVE RESULTS FOR SINGLE-SUBJECT GENERATION. THE SUBOPTIMAL RESULTS ARE ANNOTATED WITH UNDERLINES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓
DreamBooth	0.794	0.412	0.685	0.149
ELITE	0.76	0.372	0.658	0.191
DisenDreamer	0.835	0.418	0.718	0.089
Custom Diffusion	0.801	0.394	0.686	0.121
MasaCtrl	0.778	0.385	0.676	0.12
Cones 2	0.787	0.383	0.683	0.115
CausalT2I (ours)	**0.851**	**0.437**	**0.748**	**0.04**

TABLE III
QUANTITATIVE RESULTS FOR THE TWO-SUBJECT GENERATION. THE SUBOPTIMAL RESULTS ARE ANNOTATED WITH UNDERLINES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓
DreamBooth	0.716	0.325	0.644	0.269
ELITE	0.656	0.298	0.607	0.285
DisenDreamer	0.697	0.328	0.675	0.258
Custom Diffusion	0.743	0.363	0.693	0.211
MasaCtrl	0.703	0.346	0.685	0.238
Cones 2	0.71	0.355	0.689	0.222
CausalT2I (ours)	**0.775**	**0.387**	**0.729**	**0.184**

[17], [50]. We use CLIP-T [29], which calculates the average cosine similarity between the CLIP image embedding of the generated image and the text embedding of the prompt. The identifier token "s*" is omitted when evaluating CLIP-T, as the CLIP text encoder has not undergone fine-tuning on "s*". The CLIP-T is computed as follows:

$$\text{CLIP-T} = \frac{1}{N}\sum_{i=1}^{N} \text{cosine}\left(E_I(\hat{x}_i),\ E_T(\mathcal{P}_t)\right), \qquad (16)$$

where $\mathcal{P}_t$ denotes the prompt corresponding to $\hat{x}_i$. E_I and E_T are the image and text encoders of CLIP, respectively. A higher CLIP-T score indicates better semantic alignment with the input prompt.

(iii) Diversity and Overfitting Resistance: This evaluates the diversity of the generated images and aids in assessing potential model overfitting [1], [15], [50]. Following prior works [1], [15], [50], we use the average Learned Perceptual Image Patch Similarity (LPIPS) [77] to compute the average perceptual distance between generated images of the same subject under identical text prompts. The LPIPS is formulated as follows:

$$\text{LPIPS} = \frac{2}{N\left(N-1\right)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{l=1}^{L}\frac{\left\|w_l \odot \left(\phi_l(\hat{x}_i) - \phi_l(\hat{x}_j)\right)\right\|_2^2}{H_l W_l}, \qquad (17)$$

where $\phi_l(\cdot)$ is the l-th layer activation from a pretrained network, and w_l is the learned perceptual weight. A higher LPIPS score indicates greater diversity and a stronger resistance to overfitting.

4) Implementation details: We use Stable Diffusion XL [78] as the pre-trained model for experiments. All Methods are run on 3 H800 GPUs. CausalT2I uses four reference images of each subject and selects images from different subjects to merge into a single image. Finally, 4 merged images containing all subjects are used to train CausalT2I. We maintain a batch size of 1, employ a learning rate of 5×10^{-5} throughout the training process, and use 5 random seeds to generate images. Our method involves training for approximately 500 iterations per subject. The fine-tuning baselines are trained with the following number of iterations per subject: DreamBooth (500), DisenDreamer (3000), Custom Diffusion (300), and Cones 2 (3000). For a fair comparison, we use 30 steps of the DPM-Solver [79] sampler with a scale of 7.5. The generated images have resolutions of 1024×1024. We ultimately set the hyper-parameter α (i.e., $\alpha = 0.001$) by calculating ratios among loss objectives. The details of hyper-parameter analysis are presented in the subsection IV-G.

B. Single-subject Results (RQ1)

Qualitative Results. Fig. 3 shows the qualitative results of single-subject generation. CausalT2I effectively generates images that preserve abundant semantic information from diverse text prompts. Conversely, the baseline methods fail to capture the identity information of subjects well, resulting in images that deviate from the reference images. For instance, in the fourth row of Fig. 3, there is a noticeable distortion in the identity of the "dog" generated by the baselines compared to the "dog" in the reference images. Additionally, some images generated by the baselines, such as those prompted with "cat" (e.g, the second row of Fig. 3), "dog" (e.g, third row of Fig. 3), or "teddybear" (e.g, the sixth row of Fig. 3), fail to align with the corresponding text descriptions effectively. In contrast, CausalT2I consistently delivers visually accurate images for all subjects that closely adhere to the text prompts, as evident in the last column of Fig. 3.

Quantitative Results. We also comprehensively evaluate seven methods for single-subject generation. Table II summarizes the quantitative results across all target subjects. CausalT2I emerged as the superior choice in four key metrics, effectively outperforming the state-of-the-art baseline methods. Specifically, compared with the suboptimal method DisenDreamer, CausalT2I achieves notable improvements, including a 1.6% increase in CLIP-I score, a 1.9% boost in CLIP-T score, and a 3% enhancement in LPIPS score. Furthermore, CausalT2I demonstrates a 4.9% decrease in KID score, further underscoring its superiority. These results indicate that CausalT2I effectively improves image fidelity, enhances the understanding of semantic information from text prompts, and reduces overfitting, thereby delivering superior performance in customized single-subject T2I generation.

C. Multi-Subject Results

Qualitative Results. The qualitative results of multi-subject generation are presented in Fig. 4, which contains two pets, three objects, a person, and a scene. Fig. 4 demonstrates that the baseline methods fail to generate images that deviate from the reference images and conform to text prompts. As the number of subjects increases, the baseline methods experience more severe issues of forgetting the identity information within the subjects and confusion among the subjects. For instance,

TABLE IV
QUANTITATIVE RESULTS OF THE THREE-SUBJECT AND FOUR-SUBJECT. THE SUBOPTIMAL RESULTS ARE ANNOTATED WITH UNDERLINES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Mothods	Three-subject				Four-subject			
	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓
DreamBooth	0.657	0.279	0.608	0.299	0.621	0.27	0.591	0.281
ELITE	0.631	0.268	0.578	0.315	0.611	0.255	0.56	0.311
DisenDreamer	0.662	0.294	0.604	0.298	0.633	0.291	0.598	0.285
Custom Diffusion	0.692	0.319	0.646	0.276	0.653	0.298	0.613	0.271
MasaCtrl	0.688	0.313	0.642	0.281	0.65	0.292	0.601	0.29
Cones 2	0.696	0.333	0.648	0.275	0.655	0.3	0.622	0.264
CausalT2I (ours)	**0.74**	**0.361**	**0.687**	**0.192**	**0.695**	**0.329**	**0.664**	**0.234**

the images generated by the baseline methods lose the identity information of the “dog” (e.g., the second, sixth, and eight rows of Fig. 4), the “man” (e.g., the fifth and seventh rows in Fig. 4), and the “sunglasses” (e.g., the third and fourth rows in Fig. 4), resulting in low fidelity to the subjects in the reference images. Additionally, the first row in Fig. 4 shows that the customized images generated by the baseline methods, featuring the “cat and dog” as the subjects, suffer from severe confusion, leading to inconsistency with the given subjects and text descriptions. In contrast, our method appropriately presents various subjects (e.g., pets, objects, the human, and the scene, in the last column in Fig. 4) with different text prompts while accurately reflecting semantic information from the text prompts.

Quantitative Results. Table III shows the results of the two-subject generation. Compared to the state-of-the-art baseline (i.e., Custom Diffusion), CausalT2I achieves higher CLIP-T, CLIP-I, and LPIPS scores, with increases of 3.2%, 2.4%, and 3.6%, respectively. Furthermore, CausalT2I exhibits a 2.7% lower KID score. Table IV shows the quantitative metrics for three- and four-subject. CasualT2I demonstrates an overall improvement compared to the state-of-the-art baseline method Cones 2. Specifically, for the three-subject generation, CausalT2I achieves notable increases in CLIP-I score (4.4%), CLIP-T score (2.8%), and LPIPS score (3.9%). Additionally, there is a remarkable 8.3% decrease in the KID score. For the four-subject generation, compared with the suboptimal method, CasualT2I effectively improves the performance, with a 4% increase in CLIP-I score, a 2.9% increase in CLIP-T score, a 4.2% increase in LPIPS score, and a 3% decrease in KID score. These results indicate that CasualT2I effectively preserves the visual information of multiple subjects, enhances the understanding of semantic information from the text prompts, and reduces overfitting, thereby delivering superior performance in customized multi-subject T2I generation.

D. Human Preference Study (RQ3)

This study compares user evaluations of CausalT2I with state-of-the-art baselines. Given the large number of images generated from various text prompt combinations, most methods evaluate only a subset. For instance, Cones [17] and Cones 2 [5] select 3–4 subject combinations. Following this practice, we design four distinct subject combinations for each task involving one to four subjects. For each combination, we randomly select four text prompts. For each prompt, we generate 50 samples using 50 random seeds, and apply two sampling strategies: (1) Top-5: selecting the 5 best samples; (2) Random-5: randomly selecting 5 samples. In total, 2560 images are selected for user evaluation. DreamBooth [1], DisenDreamer [50], and Custom Diffusion [3] are selected as the baselines for single-subject generation, while Custom Diffusion [3], MasaCtrl [4], and Cones 2 [5] are chosen as the baselines for multi-subject generation. We recruit 100 annotators to rate the generated images on a 5-point scale (from 1: worst to 5: best) across the following three aspects: (1) Subject Fidelity (how well the generated image preserves the subject identity); (2) Image-text Alignment (how well the image aligns with the given text prompt); and (3) Image Realism (the overall visual quality and photorealism of the image). For each evaluation task, the outputs of all methods are anonymized and presented in random order, separately for both the Top-5 and Random-5 settings. Table V presents the average user scores across different evaluation tasks. These results demonstrate that our CausalT2I framework is consistently the most preferred by users, particularly in multi-subject generation.

TABLE V
HUMAN PREFERENCE STUDY. EACH VALUE REPRESENTS THE AVERAGE USER RATING (1=WORST, 5=BEST). BEST RESULTS ARE IN **BOLD**.

Subject Type	Method	**Top-5/Random-5**		
		Subject Fidelity	Image-text Alignment	Image Realism
Single -subject	DreamBooth	3.3/3.2	3.2/3.1	3.1/3.0
	DisenDreamer	3.7/3.6	3.8/3.6	3.4/3.2
	Custom Diffusion	3.6/3.4	3.4/3.2	3.3/3.1
	CausalT2I (ours)	**4.5/4.3**	**4.4/4.2**	**4.3/4.1**
Two -subject	Custom Diffusion	3.3/3.1	3.2/3.1	3.1/3.0
	MasaCtrl	2.9/2.7	2.8/2.7	2.7/2.7
	Cones 2	3.4/3.2	3.1/2.9	3.1/3.1
	CausalT2I (ours)	**4.4/4.3**	**4.3/4.3**	**4.2/4.1**
Three -subject	Custom Diffusion	3.0/2.9	3.1/3.0	3.0/2.9
	MasaCtrl	2.7/2.5	2.7/2.6	2.7/2.6
	Cones 2	3.3/2.7	2.9/2.8	2.8/2.8
	CausalT2I (ours)	**4.3/4.1**	**4.4/4.3**	**4.2/4.1**
Four -subject	Custom Diffusion	2.9/2.8	3.0/2.9	2.9/2.8
	MasaCtrl	2.6/2.5	2.7/2.5	2.6/2.5
	Cones 2	3.0/2.8	2.9/2.8	3.0/2.8
	CausalT2I (ours)	**4.2/4.0**	**4.2/4.1**	**4.1/4.0**

E. Ablation Study (RQ4)

1) Ablation Results of Main Components: We ablate three main components (i.e., *subject-aware causal disentanglement*

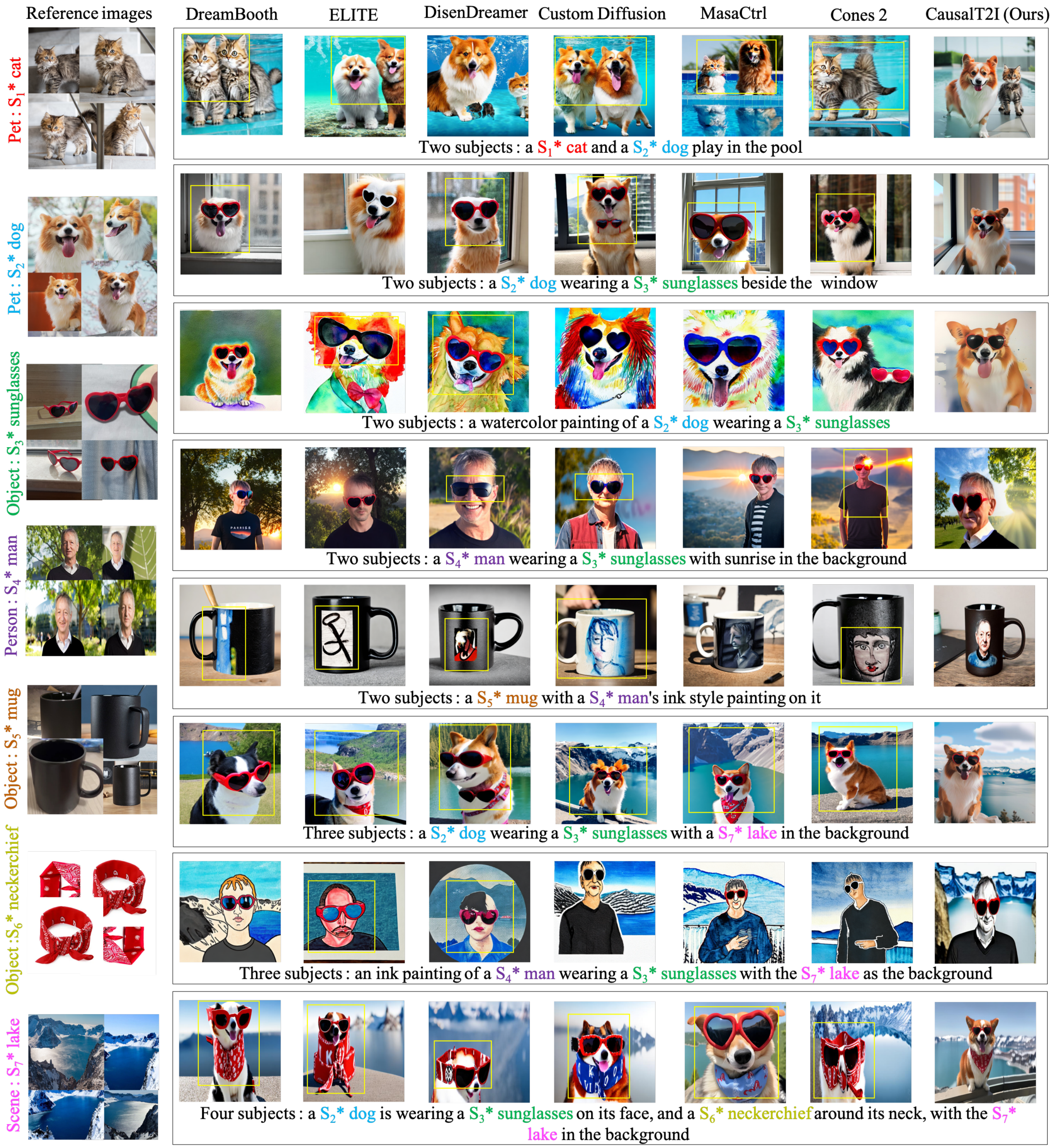

Reference images
DreamBooth
ELITE
DisenDreamer
Custom Diffusion
MasaCtrl
Cones 2
CausalT2I (Ours)
Pet : S_1* cat
Pet : S_2* dog
Object : S_3* sunglasses
Person : S_4* man
Object : S_5* mug
Object :S_6* neckerchief
Scene : S_7* lake
Two subjects : a S_1* cat and a S_2* dog play in the pool
Two subjects : a S_2* dog wearing a S_3* sunglasses beside the window
Two subjects : a watercolor painting of a S_2* dog wearing a S_3* sunglasses
Two subjects : a S_4* man wearing a S_3* sunglasses with sunrise in the background
Two subjects : a S_5* mug with a S_4* man's ink style painting on it
Three subjects : a S_2* dog wearing a S_3* sunglasses with a S_7* lake in the background
Three subjects : an ink painting of a S_4* man wearing a S_3* sunglasses with the S_7* lake as the background
Four subjects : a S_2* dog is wearing a S_3* sunglasses on its face, and a S_6* neckerchief around its neck, with the S_7* lake in the background

Fig. 4. Generated images of multiple subjects based on different text prompts and compared with various baseline methods.

(SaCD), *soft cross-attention guidance* (SCaG), and *causal denoise objective* (CdO)) of CausalT2I to show their contribution. There are two points to note: (1) *w/o SaCD*. Since the CdO and the SCaG depend on the SaCD, the others will also be ablated when the SaCD is ablated. To address this, we design an alternative ablation scheme. Specifically, we first use the SAM (Segment Anything Model) [80] to obtain pixel-level masks that indicate identity-relevant and irrelevant regions. We then use representations of these pixel-level masks to replace the feature-level embeddings produced by SaCD, allowing SCaG and CdO to function as usual. (2) *w/o CdO*. We ablate the causal denoise module, i.e., training the traditional denoise objective (no identity-irrelevant information of subjects).

We present the ablation study on the two-subject generation, as shown in Table VI. The ablation results indicate that all three components contribute to improving the performance of multi-subject generation, with SaCD playing a more crucial role in particular. We attribute this to the following two main reasons: (1) SaCD learns feature-level masks at the semantic level, enabling the model to flexibly capture both subject-

TABLE VI
ABLATION STUDY ON TWO-SUBJECT GENERATION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓
w/o SaCD	0.769	0.371	0.704	0.199
w/o SCaG	0.756	0.377	0.716	0.206
w/o CdO	0.753	0.379	0.71	0.203
CausalT2I	**0.775**	**0.387**	**0.729**	**0.184**

relevant and irrelevant information, thus enhancing image-text consistency for the subject. In contrast, pixel-level masks can only distinguish foreground from background, failing to effectively model the semantic information, such as pose. (2) The masks in SaCD are learnable and can undergo joint optimization during fine-tuning, whereas pixel-level masking follows a two-stage pipeline, which may potentially result in suboptimal performance.

The visualization ablation results on three groups of two subjects (i.e., "the cat and the dog", "the sunglasses and the dog", "the tortoise plushy and the teddy bear") customized generation are shown in Fig. 5. The results in Fig. 5 indicate that all three components can enhance the performance of multi-subject generation. To further illustrate the effectiveness of subject-aware causal disentanglement (SaCD) and soft cross-attention guidance (SCaG), the visualization of SaCD and SCaG is presented in Fig. 6. Fig. 6 indicates that our method can capture the identity-relevant embedding and generate the cross-attention maps effectively.

2) Ablation Results of Cross-attention Block: To further investigate which cross-attention block performs best, we conduct experiments by fine-tuning different cross-attention blocks on two-subject and three-subject tasks. As shown in Table VII, the up-block outperforms both the down- and mid-blocks. Training all cross-attention blocks together yields the best results. This is likely because the U-Net first extracts low-level features through the down-block, then refines these representations in the mid-block, and finally passes them to the up-block. As a result, the parameters of the up-block play a more crucial role in learning the latent space representations compared to the down- and mid-blocks.

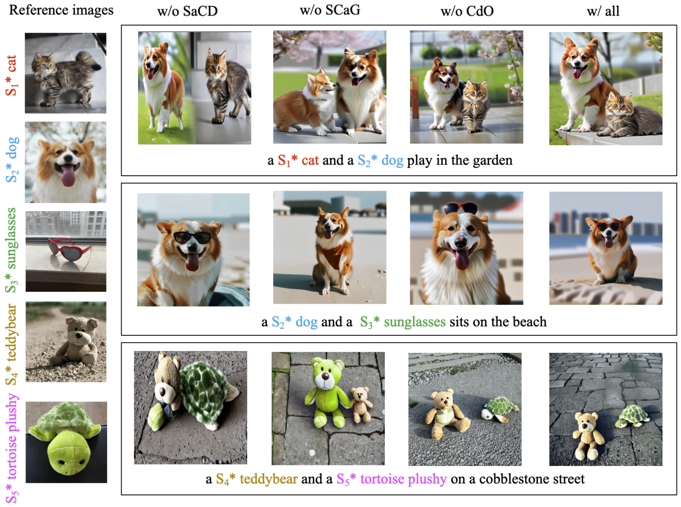

Reference images
w/o SaCD
w/o SCaG
w/o CdO
w/ all
S₁* cat
S₂* dog
S₃* sunglasses
S₄* teddybear
S₅* tortoise plushy
a S₁* cat and a S₂* dog play in the garden
a S₂* dog and a S₃* sunglasses sits on the beach
a S₄* teddybear and a S₅* tortoise plushy on a cobblestone street

Fig. 5. The ablation visualization of two-subject generation.

F. Fine-tuning Time Cost (RQ5)

To evaluate fine-tuning time costs, Table VIII presents the training time and inference time (for generating a single image) of fine-tuning-based methods for two-subject images on a single H800 GPU. As shown in Table VIII, our method incurs no significant increase in fine-tuning costs compared to DreamBooth, DisenDreamer, and Custom Diffusion. Moreover, our framework enables parameters trained on multiple subjects to be reused for generating images of individual subjects or fewer subjects without requiring retraining. For instance, parameters trained on both "cat" and "dog" can be used to generate customized images of either subject. Thus, our method enhances customized multi-subject T2I generation without significantly increasing fine-tuning time costs.

G. Hyper-parameter

Our CausalT2I framework involves one hyper-parameter, i.e., the weight of the causal disentangled objective, denoted as α. Using two challenging subject groups as examples ("the cat and the dog", "the teddy bear and the tortoise plushy"), Table IX presents the quantitative results of our hyper-parameter analysis. To mitigate any negative influence that the causal disentangled objective might have on the causal denoising objective, we initially computed the ratio of the causal disentangled objective to the causal denoising objective over the first 300 rounds. This allowed us to estimate the approximate ranges for α (roughly 0.0006-0.013), ensuring that the causal disentangled and causal denoising objectives maintain similar magnitudes. Subsequently, we fine-tuned the model using various values of α within the established ranges. As evident from Table IX, increasing α enhances the disentanglement of visual embeddings and improves image similarity, reflected in higher CLIP-I scores and LPIPS values. However, a huge value

TABLE VII
QUANTITATIVE RESULTS FOR TWO-SUBJECT AND THREE-SUBJECT TASKS WITH FINE-TUNING OF DIFFERENT CROSS-ATTENTION BLOCKS. THE SUBOPTIMAL RESULTS ARE ANNOTATED WITH UNDERLINES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Cross-Attention Block	Two-subject				Three-subject			
	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓
Up-Block	0.763	0.368	0.716	0.196	0.721	0.336	0.653	0.207
Mid-Block	0.755	0.365	0.712	0.197	0.711	0.334	0.642	0.213
Down-Block	0.752	0.358	0.706	0.204	0.707	0.321	0.62	0.219
All-Blocks	**0.775**	**0.387**	**0.729**	**0.184**	**0.74**	**0.361**	**0.687**	**0.192**

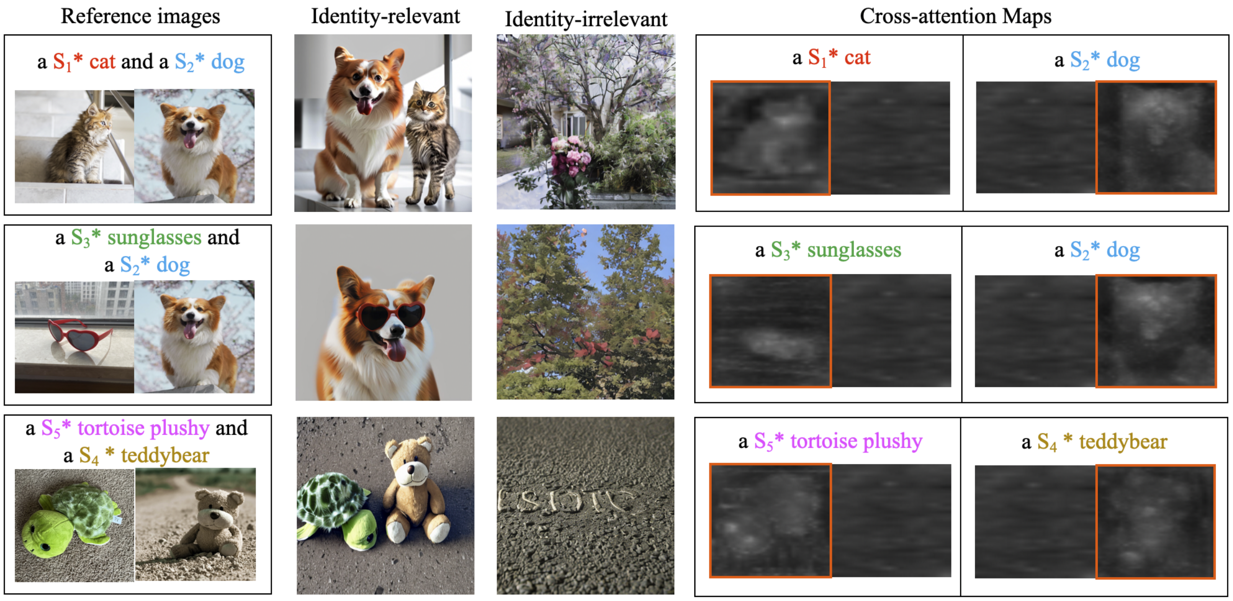


Fig. 6. The visualizations of subject-aware causal disentanglement and cross-attention map for two-subject generation.

TABLE VIII
THE TRAINING TIME AND INFERENCE TIME FOR GENERATING A SINGLE IMAGE WITH TWO SUBJECTS ON AN H800 GPU.

Methods	Training Time	Inference Time
DreamBooth	15 minutes	9 seconds
DisenDreamer	28 minutes	17 seconds
Custom Diffusion	9 minutes	8 seconds
Cones 2	65 minutes	69 seconds
CausalT2I (ours)	8 minutes	6 seconds

TABLE IX
HYPER-PARAMETER EXPERIMENTS ON α. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

α	CLIP-I ↑	CLIP-T ↑	LPIPS ↑	KID ↓
0.0	0.752	0.368	0.704	0.202
0.001	0.768	**0.387**	0.726	**0.184**
0.01	**0.772**	0.379	**0.728**	0.189
0.1	0.759	0.372	0.723	0.196

like 0.1 can cause this objective to dominate the optimization process, compromising the denoising process and resulting in lower CLIP-T scores and higher KID scores. Empirically, we found that setting α to 0.001 yielded optimal results.

V. Limitation and Discussion

Extensive experiments demonstrate the effectiveness of our CausalT2I framework in customized multi-subject T2I generation. However, some limitations remain that require further improvement and discussion.

(1) Merging Images. To help the model better learn the spatial relationships between multiple subjects in a single image, we adopt a simple image merging method. While this does not introduce significant additional training dependencies compared to approaches requiring specific layouts, it adds some complexity. In future work, we plan to explore using large multi-modal models to learn spatial layouts, which removes the need for manual composition and enhances flexibility.

(2) Subject-specific Fine-Tuning. CausalT2I requires fine-tuning on reference images, averaging around 500 iterations per subject. While CausalT2I improves multi-subject generation performance without significant overhead compared to contrastive fine-tuning baseline methods, there is room for optimization. We aim to refine the process in future iterations, focusing on generating higher-quality images with lower computational costs.

VI. Conclusion

In this paper, we propose CausalT2I, an innovative causal tuning framework for customized multi-subject text-to-image generation. Unlike existing methods, CausalT2I effectively distinguishes between causally relevant and irrelevant information associated with subjects through *subject-aware causal disentanglement*. It also employs *soft cross-attention guidance* to mitigate inter-subject interference. Additionally, the framework designs a *causal denoising objective* that optimizes the denoising process by integrating identity-preserved textual information with identity-irrelevant visual details, thereby enhancing the model's understanding of global visual contexts. Extensive experiments demonstrate that CausalT2I generates high-quality multi-subject images that accurately preserve subject identities and adhere to complex textual descriptions.

References

[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023, pp. 22 500–22 510.

[2] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2023, pp. 15 897–15 907.

[3] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023, pp. 1931–1941.

[4] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2023, pp. 22 503–22 513.

[5] Z. Liu, Y. Zhang, Y. Shen, K. Zheng, K. Zhu, R. Feng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Customizable image synthesis with multiple subjects," in *Proceedings of the Advances in Neural Information Processing Systems(NeurIPS)*, 2023, pp. 57 500–57 519.

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proceedings of the Advances in Neural Information Processing Systems(NeurIPS)*, 2022, pp. 36 479–36 494.

[7] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," *Transactions on Machine Learning Research*, 2022.

[8] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the International Conference on Machine Learning(ICML)*, 2022, pp. 16 784–16 804.

[9] Y. Ma, H. Yang, W. Wang, J. Fu, and J. Liu, "Unified multi-modal latent diffusion for joint subject and text conditional image generation," *arXiv preprint arXiv:2303.09319*, 2023.

[10] J. Chen, Y. Pan, T. Yao, and T. Mei, "Controlstyle: Text-driven stylized image generation using diffusion priors," in *Proceedings of the ACM International Conference on Multimedia(MM)*, 2023, pp. 7540–7548.

[11] Y. Lu, C. Du, Q. Zhou, D. Wang, and H. He, "Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion," in *Proceedings of the ACM International Conference on Multimedia(MM)*, 2023, pp. 5899–5908.

[12] S. Zhong, Z. Huang, W. Wen, J. Qin, and L. Lin, "Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models," in *Proceedings of the ACM International Conference on Multimedia(MM)*, 2023, pp. 567–578.

[13] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2023, pp. 7289–7300.

[14] Z. Wang, O. Li, T. Wang, L. Wei, Y. Hao, X. Wang, and Q. Tian, "Prior preserved text-to-image personalization without image regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 2, pp. 1318–1330, 2025.

[15] J. Shentu, M. Watson, and N. A. Moubayed, "Textual localization: Decomposing multi-concept images for subject-driven text-to-image generation," *arXiv preprint arXiv:2402.09966*, 2024.

[16] H. Chen, Y. Zhang, S. Wu, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation," in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2024.

[17] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones: Concept neurons in diffusion models for customized generation," in *Proceedings of the International Conference on Machine Learning(ICML)*, 2023, pp. 21 548–21 566.

[18] R. Po, G. Yang, K. Aberman, and G. Wetzstein, "Orthogonal adaptation for modular customization of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 7964–7973.

[19] J. Liang, W. Pei, and F. Lu, "Layout-bridging text-to-image synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7438–7451, 2023.

[20] X. Zhang, K. Liu, X. Wang, Z. Zhou, and H. Chen, "Rmgnet: The progressive relationship-mining graph neural network for text-to-image person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 5749–5761, 2025.

[21] H. Zhang, T. Xu, and H. Li, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2017, pp. 5908–5916.

[22] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis," in *Proceedings of the International Conference on Machine Learning(ICML)*, 2023, pp. 30 105–30 118.

[23] K. Cui, Y. Yu, F. Zhan, S. Liao, S. Lu, and E. P. Xing, "KD-DLGAN: data limited image generation via knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023, pp. 3872–3882.

[24] M. Tao, B. Bao, H. Tang, and C. Xu, "GALIP: generative adversarial clips for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023, pp. 14 214–14 223.

[25] B. Yang, X. Xiang, W. Kong, J. Zhang, and Y. Peng, "DMF-GAN: deep multimodal fusion generative adversarial networks for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 26, pp. 6956–6967, 2024.

[26] Q. Cheng, K. Wen, and X. Gu, "Vision-language matching for text-to-image synthesis via generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 25, pp. 7062–7075, 2023.

[27] B. Yuan, Y. Sheng, B. Bao, Y. P. Chen, and C. Xu, "Semantic distance adversarial learning for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 26, pp. 1255–1266, 2024.

[28] H. Tan, B. Yin, K. Xu, H. Wang, X. Liu, and X. Li, "Attention-bridged modal interaction for text-to-image generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5400–5413, 2024.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning(ICML)*, 2021, pp. 8748–8763.

[30] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2021.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proceedings of the Advances in neural information processing systems(NeurIPS)*, pp. 6840–6851, 2020.

[32] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," in *Proceedings of the Annual Conference on Neural Information Processing Systems(NeurIPS)*, 2021, pp. 8780–8794.

[33] A. Das, S. Fotiadis, A. Batra, F. Nabiei, F. Liao, S. Vakili, D. Shiu, and A. Bernacchia, "Image generation with shortest path diffusion," in *Proceedings of the International Conference on Machine Learning(ICML)*, 2023, pp. 7009–7024.

[34] Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, "Shifted diffusion for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023, pp. 10 157–10 166.

[35] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[36] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Proceedings of the Advances in neural information processing systems(NeurIPS)*, pp. 36 479–36 494, 2022.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2022, pp. 10 674–10 685.

[38] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2023.

[39] S. Zhao, D. Chen, Y. Chen, J. Bao, S. Hao, L. Yuan, and K. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," in *Proceedings of the Advances in Neural Information Processing Systems(NeurIPS)*, 2023.

[40] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "Multidiffusion: Fusing diffusion paths for controlled image generation," in *Proceedings of the International Conference on Machine Learning(ICML)*, 2023, pp. 1737–1752.

[41] J. Zhu, H. Ma, J. Chen, and J. Yuan, "Isolated diffusion: Optimizing multi-concept text-to-image generation training-freely with isolated diffusion guidance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 9, pp. 6280–6292, 2025.

[42] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 148:1–148:10, 2023.

[43] J. Ma, J. Liang, C. Chen, and H. Lu, "Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning," in *Proceedings of the ACM SIGGRAPH Conference(SIGGRAPH)*, 2024.

[44] V. Sarukkai, L. Li, A. Ma, C. Ré, and K. Fatahalian, “Collage diffusion,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(WACV)*, 2024, pp. 4196–4205.

[45] Y. Alaluf, E. Richardson, G. Metzer, and D. Cohen-Or, “A neural space-time representation for text-to-image personalization,” *ACM Transactions on Graphics*, vol. 42, no. 6, pp. 243:1–243:10, 2023.

[46] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An image is worth one word: Personalizing text-to-image generation using textual inversion,” in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2023.

[47] Z. Sun, Y. Zhou, H. He, and P. Y. Mok, “Sgdiff: A style guided diffusion model for fashion synthesis,” in *Proceedings of the ACM International Conference on Multimedia(MM)*, 2023, pp. 8433–8442.

[48] J. S. Smith, Y. Hsu, L. Zhang, T. Hua, Z. Kira, Y. Shen, and H. Jin, “Continual diffusion: Continual customization of text-to-image diffusion with c-lora,” *Transactions on Machine Learning Research*, 2024.

[49] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, “Instantbooth: Personalized text-to-image generation without test-time finetuning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 8543–8552.

[50] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, and W. Zhu, “Disendreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6860–6873, 2024.

[51] R. Zhao, M. Zhu, S. Dong, N. Wang, and X. Gao, “Catversion: Concatenating embeddings for diffusion-based text-to-image personalization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 6047–6058, 2025.

[52] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, and Y. Atzmon, “Training-free consistent text-to-image generation,” *ACM Transactions on Graphics*, vol. 43, no. 4, pp. 52:1–52:18, 2024.

[53] Y. Zhou, R. Zhang, J. Gu, and T. Sun, “Customization assistant for text-to-image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 9182–9191.

[54] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, “Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 6527–6536.

[55] N. Chen, M. Huang, Z. Chen, Y. Zheng, L. Zhang, and Z. Mao, “Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization,” in *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, 2025, pp. 2123–2131.

[56] Y. Zeng, V. M. Patel, H. Wang, X. Huang, T. Wang, M. Liu, and Y. Balaji, “Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 6786–6795.

[57] Y. Liu, C. Fan, Y. Dai, X. Chen, P. Zhou, and L. Sun, “Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 24 219–24 228.

[58] G. Ding, C. Zhao, W. Wang, Z. Yang, Z. Liu, H. Chen, and C. Shen, “Freecustom: Tuning-free customized image generation for multi-concept composition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 9089–9098.

[59] G. Kwon, S. Jenni, D. Li, J. Lee, J. C. Ye, and F. C. Heilbron, “Concept weaver: Enabling multi-concept fusion in text-to-image models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 8880–8889.

[60] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han, “Fastcomposer: Tuning-free multi-subject image generation with localized attention,” *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1175–1194, 2025.

[61] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, “Spatext: Spatio-textual representation for controllable image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2023, pp. 18 370–18 380.

[62] S. Jang, J. Jo, K. Lee, and S. J. Hwang, “Identity decoupling for multi-subject personalization of text-to-image models,” in *Proceedings of the Advances in Neural Information Processing Systems(NeurIPS)*, 2024.

[63] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu, Y. Ge, Y. Shan, and M. Z. Shou, “Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models,” in *Proceedings of the Advances in Neural Information Processing Systems(NeurIPS)*, 2023.

[64] Z. Kong, Y. Zhang, T. Yang, T. Wang, K. Zhang, B. Wu, G. Chen, W. Liu, and W. Luo, “OMG: occlusion-friendly personalized multi-concept generation in diffusion models,” in *Proceedings of the European Conference on Computer Vision(ECCV)*, 2024, pp. 253–270.

[65] H. He, H. Yang, Z. Tuo, Y. Zhou, Q. Wang, Y. Zhang, Z. Liu, W. Huang, H. Chao, and J. Yin, “Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion,” *CoRR*, vol. abs/2407.12899, 2024.

[66] Z. Guo and T. Jin, “Conceptguard: Continual personalized text-to-image generation with forgetting and confusion mitigation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2025, pp. 2945–2954.

[67] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2014.

[68] A. Karimi, K. Muandet, S. Kornblith, B. Schölkopf, and B. Kim, “On the relationship between explanation and prediction: A causal view,” in *Proceedings of the International Conference on Machine Learning(ICML)*, 2023, pp. 15 861–15 883.

[69] U. Shalit, F. D. Johansson, and D. Sontag, “Estimating individual treatment effect: Generalization bounds and algorithms,” in *Proceedings of the International Conference on Machine Learning(ICML)*, 2017, pp. 3076–3085.

[70] F. D. Johansson, U. Shalit, and D. Sontag, “Learning representations for counterfactual inference,” in *Proceedings of the International Conference on Machine Learning(ICML)*, 2016, pp. 3020–3029.

[71] P. Schwab, D. Miladinovic, and W. Karlen, “Granger-causal attentive mixtures of experts: Learning important features with neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, 2019, pp. 4846–4853.

[72] P. Schwab and W. Karlen, “Cxplain: Causal explanations for model interpretation under uncertainty,” in *Proceedings of the Advances in Neural Information Processing Systems(NeurIPS)*, 2019, pp. 10 220–10 230.

[73] M. Chen, I. Laina, and A. Vedaldi, “Training-free layout control with cross-attention guidance,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(WACV)*, 2024, pp. 5343–5353.

[74] P. Von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and T. Wolf, “Diffusers: State-of-the-art diffusion models,” 2022.

[75] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, “T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models,” in *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, 2024, pp. 4296–4304.

[76] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD gans,” in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2018.

[77] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 586–595.

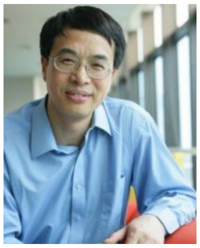
[78] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “SDXL: improving latent diffusion models for high-resolution image synthesis,” in *Proceedings of the International Conference on Learning Representations(ICLR)*, 2024.

[79] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, “Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps,” in *Proceedings of the Annual Conference on Neural Information Processing Systems(NeurIPS)*, 2022.

[80] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick, “Segment anything,” *CoRR*, vol. abs/2304.02643, 2023.

Chaoyang Li received the M.S. degree in cyberspace security from the Department of Cyber Security, University of Chinese Academy of Sciences, Beijing, China, in 2022. He is currently pursuing a Ph.D. degree in electronic information from the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His current research interests include causal learning, multi-task learning, and multi-modal large models.

Xin Wang (Memeber, IEEE) received the B.E. degree in computer science and technology from Zhejiang University, China, and the Ph.D. degree in computing science from Simon Fraser University, Canada. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. He has published over 150 high-quality research papers in top journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, ACM TOIS, ICML, NeurIPS, ACM KDD, ACM Web Conference, ACM SIGIR, and ACM Multimedia, winning three best paper awards. His research interests include multimedia intelligence and machine learning, and their applications in multimedia big data analysis. He was a recipient of the 2020 ACM China Rising Star Award, 2022 IEEE TCMC Rising Star Award, and 2023 DAMO Academy Young Fellow.



Wenwu Zhu (Fellow, IEEE) received the Ph.D. degree from New York University in 1996.

He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, the Vice Dean of the National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. Prior to his current post, he was a Senior Researcher and the Research Manager of Microsoft Research Asia. He was a Chief Scientist and the Director of Intel Research China from 2004 to 2008. He has published over 400 refereed articles and is the inventor of over 80 patents. His research interests include graph machine learning, curriculum learning, data-driven multimedia, and big data.

Dr. Zhu was with Bell Laboratories, NJ, USA, as a member of Technical Staff from 1996 to 1999. He is a fellow of AAAS, ACM, and SPIE, and a member of Academia Europaea. He received ten Best Paper Awards, including ACM Multimedia 2012 and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2001 and 2019. He serves as the Editor-in-Chief for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and IEEE TRANSACTIONS ON MULTIMEDIA (2017–2019) and the Chair of the Steering Committee for IEEE TRANSACTIONS ON MULTIMEDIA (2020–2022). He serves as the General Co-Chair for ACM Multimedia 2018 and ACM CIKM 2019.

Lingzhi Wang received the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, China, in 2023. She is currently an Associate Professor with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. Her research interests include natural language processing and the safety of large language models. She serves as an Area Chair and Senior Program Committee member for several leading conferences, including ACL, EMNLP, and IJCAI.

Ning Hu received the B.S., M.S., and Ph.D. degrees in computer science from the College of National University of Defense Technology, Changsha, China, in 1994, 1997, and 2010, respectively. He is currently a Professor with the Department of New Networks, Peng Cheng Laboratory, Shenzhen, China. His current research interests include software-defined networks, Industrial IoT, and network security. His research has been supported by the National Natural Science Foundation of China.

Qing Liao (Member, IEEE) received the Ph.D. degree from The Hong Kong University of Science and Technology, Hong Kong, China, in 2016. She is currently a Professor with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. Her research interests include data mining, artificial intelligence, and information security.