DisEnvisioner: Disentangled and Enriched Visual Prompt for Customized Image Generation

Under-review

https://disenvisioner.github.io

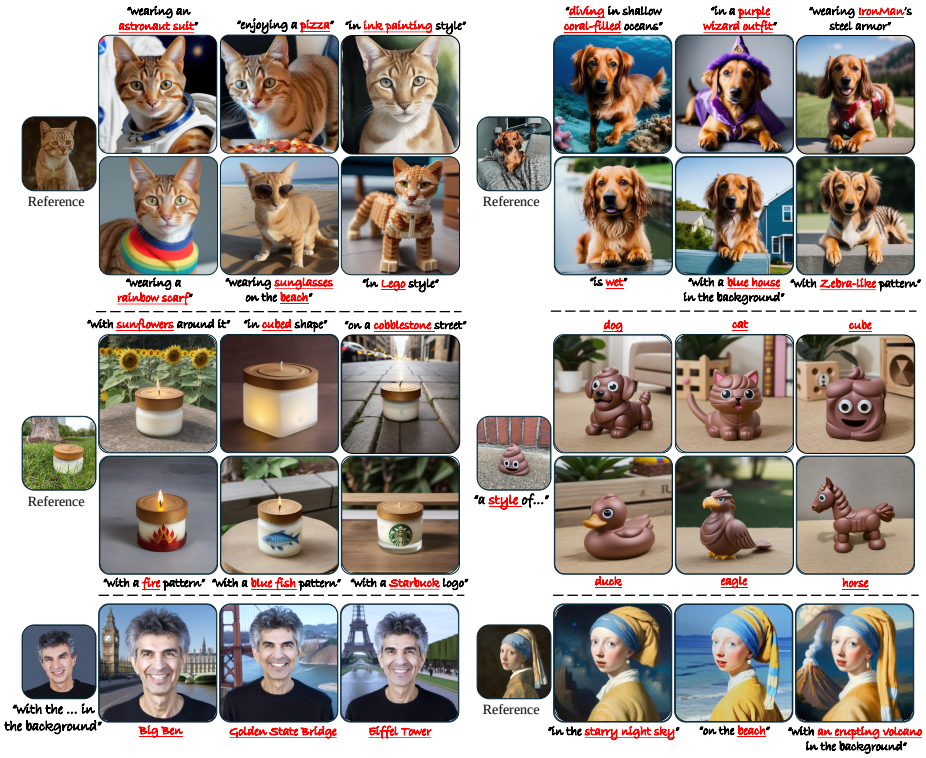


Fig. 1: Customization examples of DisEnvisioner. Without cumbersome tuning or relying on multiple reference images, our DisEnvisioner is capable of generating a variety of exceptional customized images. Characterized by its emphasis on the interpretation of subject-essential attributes, DisEnvisioner effectively discerns and enhances the subject-essential feature while filtering out irrelevant attributes, achieving superior personalizing quality with reduced inference time.

Abstract. In the realm of image generation, creating imagery of customized subject from visual prompt with additional textual instruction emerges as a promising endeavor, attracting the forefront of innovation and interest. However, existing methods, both tuning-based and tuning-free, struggle with interpreting the subject-essential attributes from the visual prompt. This leads to subject-irrelevant attributes infiltrating the generation process, ultimately compromising the personalization quality

in both editability and ID preservation. In this paper, we present **DisEnvisioner**, a novel approach for extracting subject-essential features, enabling exceptional customization performance, in a **tuning-free** manner and using only **a single image**. Specifically, the feature of the subject and other irrelevant components are effectively separated via aggregating image features into distinctive visual tokens, enabling a much more precise interpretation. Aiming to advanced customization performance, we further enrich the disentangled elements, which are sculpted into a much more granular representation. Experiments demonstrate the superiority of our approach over existing methods not only in instruction response but also the ID consistency. These highlight the effectiveness and efficiency of DisEnvisioner, paving the way for practical image customization applications.

Keywords: Visual Disentanglement and Enrichment · Zero-shot Customization · Text-to-Image Generation

1 Introduction

By training with billions of image-text pairs, state-of-the-art text-to-image generation models, *e.g.*, DALL·E [29], Imagen [33], UnCLIP [28], Stable Diffusion (SD) [30], and PixArt-α [4], have demonstrated remarkable proficiency in generating contextually aligned images from textual descriptions. Despite their unprecedentedly creative capabilities, **customized image generation** poses a much more intricate challenge. This task aims to synthesize life-like imagery that accurately response natural language instructions (referred to as **editability**) while preserving the subject identity in reference images (referred to as **ID consistency**), has garnered great attention from both academia and industry [1, 3, 5, 7, 8, 10, 14, 18, 20, 23, 25, 32, 34, 36–38, 41].

To enhance the quality of personalization, accurate interpretation of the given image is crucial. This involves effectively extracting **subject-essential attributes** from the reference image while minimizing the influence of **subject-irrelevant attributes**. Failure to do so may result in 1) overemphasis on irrelevant details: generated images may prioritize irrelevant information, sidelining the textual instructions and compromising the overall editability of the output; 2) diminished subject identity: key details essential to the object pertinent to the subject may be inadequately preserved, leading to a degradation of the subject's identity consistency.

Existing methods, both tuning-based [3, 7, 8, 18, 32] and tuning-free [20, 23, 25, 38, 41], struggle with interpreting the subject-essential attributes, particularly in scenarios involving a single image. Specifically, prevailing tuning-based methods [3, 8, 18, 32], like DreamBooth [32] and DisenBooth [3], rely heavily on multiple reference images to bind the subject concept into the model. Despite their impressive results, these methods are not only hindered by the time-consuming tuning process, but also the lack of subject attributes extraction in single-image scenarios that leads to compromised customization quality, as illustrated in Fig. 2. Recent tuning-free methods [5, 20, 23, 25, 34, 38, 41], offering

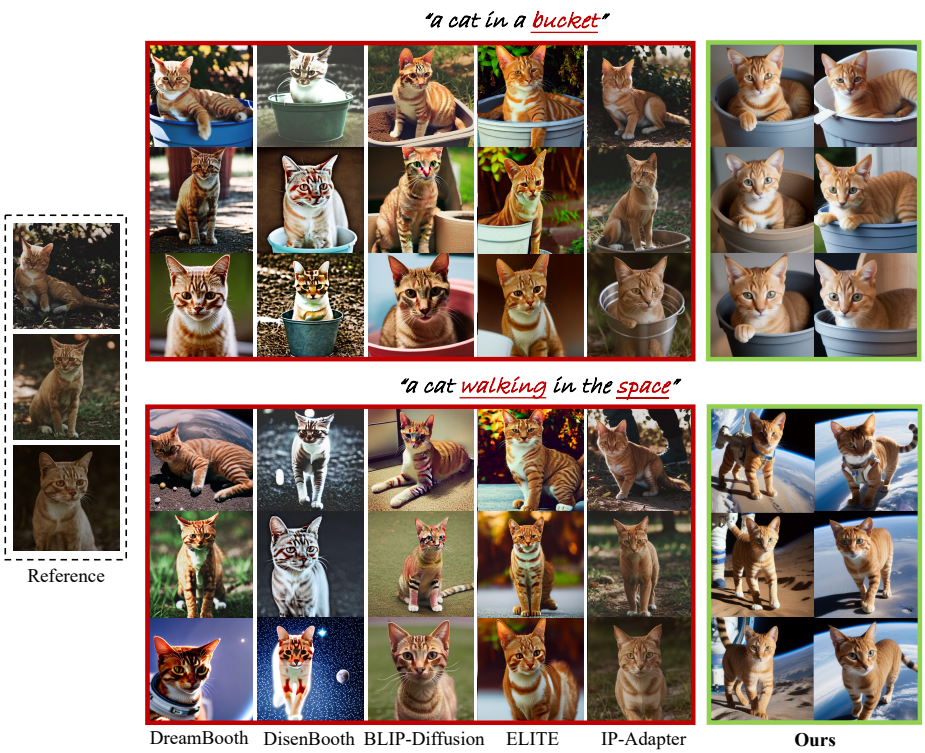


Fig. 2: Comparisons between our method and other existing methods [3, 20, 32, 38, 41] under single-image setting. By employing multiple reference images of the same subject captured under varied conditions, we observe that extraneous factors affect customization quality in both editability and identity consistency. For instance, exact posture replication leads to scene distortions, and background impedes the transition to novel "space" scene.

a notable boost in the inference speed, try to explore the customized image generation given only single image. Nonetheless, methods like ELITE [38] and Subject-Diffusion [25], heavily rely on additional segmentation masks, which is still inadequate in subject feature extraction due to a lack of semantic interpretation of the visual prompt. In addition, IP-Adapter [41], BLIP-Diffusion [20], and PhotoVerse [5] consider the global feature of the given image as subject-essential, inevitably introducing the subject-irrelevant information that diminishes the personalizing quality. As specified in Fig. 2, the customization quality of these methods are notably compromised by irrelevant factors (*e.g.*, the bushes in the background), especially in the *"walking in the space"* scenario.

Motivated by the above analysis, we propose **DisEnvisioner**, a novel framework meticulously designed to addresses the core issues via attribute disentanglement and enrichment, leading to unprecedented control and quality in image personalization. As illustrated in Fig. 3, the image is tokenized into compact disentangled features, *i.e.*, subject-essential and subject-irrelevant tokens, through

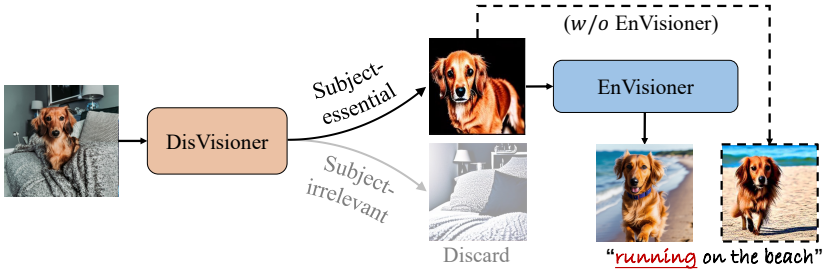


Fig. 3: Overview of our proposed DisEnvisioner, which consists of two key components: *i.e.*, DisVisioner and EnVisioner. In the inference phase, subject-irrelevant features are discarded to minimize harmful disturbances.

the image tokenizer [39] in DisVisioner. Then, the subject-irrelevant feature is ignored, facilitating the model to focus on precise subject editing according to textual instructions, eliminating the interference by irrelevant factors. Based on the disentangled features, the EnVisioner employs a specialized projector for enriching the subject feature into more granular representations. The enriched features are then injected into the frozen pre-trained SD model [30] for customized generation, significantly boosting the generation fidelity and the consistency of subject identity.

In summary, our key contributions are as follows.

- We emphasize the critical role of subject-essential attribute in customized image generation, which is the foundation of faithful subject concept reconstruction and reliable editability, thereby ensuring high-quality customization across diverse scenarios.
- We present DisEnvisioner, a straightforward yet effective framework designed for image customization. Utilizing visual disentanglement and enrichment, DisEnvisioner excels in identifying and enhancing subject-essential attributes derived from a single reference image, achieving superior image customization performance without extensive tuning.
- Comprehensive experiments validate DisEnvisioner's excellence in adhering to instructions and maintaining identity, which demonstrate its superior personalization capabilities and efficiency compared to previous approaches.

2 Related Works

2.1 Text-to-Image Generation

In the field of text-to-image generation, the evolution of methodologies has transitioned from Generative Adversarial Networks (GANs) [9, 11, 15–17, 42–44] to advanced Diffusion Models [4, 12, 26, 28–30, 33]. Early GAN-based models like StackGAN [43, 44], AttnGAN [40], and XMC-GAN [42] introduced multi-stage generation and attention mechanisms to improve image fidelity and alignment with textual descriptions. A significant breakthrough, DALL·E [29], utilizes a

transformer and auto-regressive model to merge text and image data, trained on 250 million pairs for high-quality, intuitive image synthesis. Following this, a series of diffusion-based methods such as GLIDE [26], DALL·E2 [28], and Imagen [33] have been introduced, offering enhanced image quality and textual coherence. The Latent Diffusion Model (LDM) [30], trained on 5 billion pairs, further enhanced training efficiency without compromising performance, becoming a community standard. Despite remarkable strides in generating images from textual descriptions, current methodologies fall short in rendering customized visual concepts from reference images. In our paper, we dive into customized image generation, extending the capabilities of existing text-to-image techniques to craft personalized visual concepts.

2.2 Customized Image Generation

Existing methods in the field of customized image generation primarily fall into two categories: tuning-based and tuning-free. Tuning-based methods [3, 7, 8, 10, 13, 18, 32] involve fine-tuning a pre-trained generative model with several reference images of a specific concept during test-time. Among these, DreamBooth [32] fine-tunes the entire diffusion model to accurately align the target concept with a unique identifier. Custom Diffusion [18] balances the fidelity and computational cost by selectively fine-tuning the key, value mapping parameters in cross-attention layers. DisenBooth [3] leverages a shared textual embedding for subject identity and separate visual embeddings for surroundings, still heavily depends on multiple reference images to ward off the effects of irrelevant variables. DreamTuner [13] balances the generative efficiency and capability by incorporating a general subject-encoder, enabling customization with a single reference image. Despite the effectiveness of tuning-based techniques, their high computational demand and scalability challenges limit the practicality. In contrast, tuning-free methods [5, 14, 20, 23, 25, 34, 37, 38, 41], leveraging off-the-shelf pre-trained generative models, employ advanced encoders to introduce customized visual representations into image generation. InstantBooth [34] and PhotoMaker [23] derives visual representations from images encoders based on multiple reference images. BLIP-Diffusion [20] leverages the capabilities of BLIP-2 [21] to acquire text-aligned image representations for precise personalization. ELITE [38] introduces a learning-based encoder with global and local mapping networks designed for integrating the feature of visual prompt into the customization process. IP-Adapter [41] introduces a decoupled cross-attention mechanism, treating textual and visual prompts separately for enhanced generation qualify and flexibility. While tuning-free methods improve inference speed, they struggle with editability and identity consistency due to entangled elements and coarse visuals. Consequently, this paper proposes a novel tuning-free approach using visual disentanglement and enrichment for improved personalization performance in single-image scenarios.

3 Method

The objective of customized image generation is to synthesize lifelike images that adhere to the instructions while preserving the subject's identity. To tackle the pivotal challenge of reducing the impact of attributes irrelevant to the subject, we introduce DisEnvisioner—a novel approach that emphasizes disentanglement and enrichment. As depicted in Fig. 3, DisEnvisioner initiates the process by disentangling the image features into the essential and the irrelevant by our DisVisioner module. Subsequently, to bolster the consistency of the subject's identity, the EnVisioner module is employed to refine the essential features into a more detailed representation. Our discussion commences with an overview of the foundational concepts (Section 3.1), followed by an in-depth exploration of the DisVisioner (Section 3.2) and EnVisioner (Section 3.3).

3.1 Preliminaries

Diffusion models [4, 12, 26, 28, 29, 33] represent the current state-of-the-art in generative modeling for high-fidelity image generation. These models employ a dual-phase mechanism involving forward diffusion and reverse denoising processes. The forward process incrementally introduces noise to the data, transforming it into a Gaussian random noise by a Markov chain over a fixed number of steps T. In the reverse phase, a learned neural network is utilized to predict and subtract the added noise at each step, thereby recovering the original data from the noise. This dual-phase mechanism enables diffusion models to achieve impressive results in generating high-quality images with fine details and realism.

In this study, we adopt Stable Diffusion (SD) [30], a latent diffusion model built upon UNet [31], as our foundational generative model. Firstly, an auto-encoder $\{\mathscr{E}(\cdot), \mathscr{D}(\cdot)\}$ is trained to map between RGB space and the latent space, *i.e.*, $\mathscr{E}(\mathbf{x}) = \mathbf{z}$, $\mathscr{D}(\mathscr{E}(\mathbf{x})) \approx \mathbf{x}$. And the textual conditions are obtained using a pre-trained CLIP text encoder $\boldsymbol{c} = \psi_\theta^{\mathrm{T}}(y)$, where y is the given prompt. The training objective is:

$$L_{\mathrm{LDM}}(\theta) = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{c}, \boldsymbol{\epsilon}_t \sim \mathcal{N}(0,1), t \in [1,T]} \Big[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \boldsymbol{c}, t)\|_2^2 \Big], \tag{1}$$

where $\boldsymbol{\epsilon}_t$ represents the noise introduced during the forward process, and $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \boldsymbol{c}, t)$ is the predicted noise. Cross-attention is adopted to introduce textual condition into SD. Latent image feature $\boldsymbol{f} \in \mathbb{R}^{(\mathrm{H} \times \mathrm{W}) \times d_{\mathrm{q}}}$ and text condition $\boldsymbol{c} \in \mathbb{R}^{n_y \times d_{\mathrm{k}}}$ are initially projected to obtain query $\boldsymbol{Q} = \boldsymbol{w}_{\texttt{to\_q}} \circ \boldsymbol{f}$, key $\boldsymbol{K} = \boldsymbol{w}_{\texttt{to\_k}} \circ \boldsymbol{c}$, and value $\boldsymbol{V} = \boldsymbol{w}_{\texttt{to\_v}} \circ \boldsymbol{c}$, where $\boldsymbol{w}$ are weights of the corresponding mapping layers. Then, the cross-attention is calculated as:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}. \tag{2}$$

During inference, the model iteratively constructs images $\mathbf{x}_0$ from random noises $\mathbf{z}_T$ through reverse denoising: $\hat{\mathbf{x}}_0 = \mathscr{D}(\hat{\mathbf{z}}_0)$.

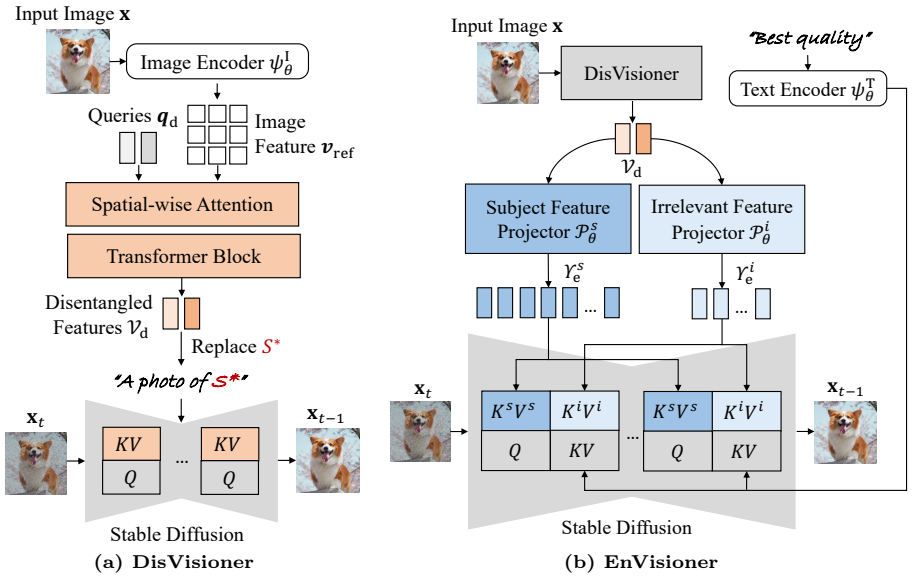


Fig. 4: Training pipeline of our proposed DisEnvisioner. Our approach is structured into two stages: (a) DisVisioner firstly decomposes the feature of the subject and other irrelevant components by aggregating the image feature $\boldsymbol{v}_{\mathrm{ref}}$ into two distinct and orthogonal tokens $\boldsymbol{v}_{\mathrm{d}}$. (b) EnVisioner subsequently refines and sculpts the disentangled features $\mathcal{V}_{\mathrm{d}}$ into more granular representations to produce high ID-consistency images with the input image $\mathbf{x}_{\mathrm{ref}}$. Only colored modules (orange and blue) are trainable.

3.2 DisVisioner

As revealed in [20, 38, 41], the customized image can be encoded as a sequence of tokens via inversion. In our work, we propose to isolate the subject-essential token from irrelevant components using an image tokenizer [39]. Image tokenizer is a method to aggregate the image features into compact visual tokens, with each token corresponds to a distinct visual component. As illustrated in Fig. 4a, given a reference image $\mathbf{x}_{\mathrm{ref}}$, we employ the CLIP image encoder to extract image features $\boldsymbol{v}_{\mathrm{ref}} = \psi_\theta^{\mathrm{I}}(\mathbf{x}_{\mathrm{ref}}) \in \mathbb{R}^{(\mathrm{H}\times\mathrm{W})\times d_{\mathrm{k}}}$. Subsequently, the image tokenizer network $M(\cdot)$ is utilized to extract disentangled visual tokens:

$$\mathcal{V}_{\mathrm{d}} = M(\boldsymbol{q}_{\mathrm{d}}, \psi_\theta^{\mathrm{I}}(\mathbf{x}_{\mathrm{ref}})) \tag{3}$$

where $\boldsymbol{q}_{\mathrm{d}} \in \mathbb{R}^{(n_s+n_i)\times d_{\mathrm{q}}}$ is the queries for feature aggregation, n_s for the subject feature while n_i for the irrelevant. In image tokenizer $M(\cdot)$, a spatial-wise cross-attention mechanism is adopted to aggregate the image features, where $\boldsymbol{q}_{\mathrm{d}}$ is the query and $\psi_\theta^{\mathrm{I}}(\mathbf{x}_{\mathrm{ref}})$ serves as the key and value. Consequently, the image features are separated into two aspects of features according to $\boldsymbol{q}_{\mathrm{d}}$. In order to determine the sequence of disentangled features, we initialize $\boldsymbol{q}_{\mathrm{d}}$ with the CLIP prior as revealed in [22]. The n_s queries for subject feature are initialized by random and

other n_i queries with the tokenized class name. Transformer blocks are followed to refine the disentangled features to obtain $\mathcal{V}_\mathrm{d}$ as the output of our DisVisioner.

To train the image tokenizer $M(\cdot)$, we insert the disentangled features $\mathcal{V}_\mathrm{d}$ into textual feature space of the prompt by replacing the textual embedding of placeholder $S*$, and adopt the Eq. 1 as the training objective for the target of reconstruction. Following [18, 38], the entire SD model is frozen except the `to_k` and `to_v` layers of the cross-attention module for correct interpretation of the new disentangled tokens. In our experiment, we set $n_s = 1$ and $n_i = 1$ for subject-essential and irrelevant features, respectively. Excessive tokens will lead to inaccurate disentanglement, thus impairing the response to the textual instruction and ID consistency (please refer to Sec. 4.4 for more details).

Benefiting from image tokenizer, the subject-essential features are accurately compressed into the tokens while well separated from irrelevant factors via the spatial-wise attention space. When customized image generation, the subject-irrelevant token will be discarded for exclude the unwanted disturbance, facilitating the instruction editing and ID fidelity.

3.3 EnVisioner

While DisVisioner effectively extracts the subject-essential feature into a single token, it may be inadequate for capturing the detailed nuances of the customized subject. In EnVisioner, we introduce a novel approach wherein we map the disentangled features into a sequence of granular tokens using new projectors $P^s(\cdot)$ and $P^i(\cdot)$. The utilization of separate projectors ensures the disentanglement between the subject-essential and irrelevant tokens.

As illustrated in Fig. 4b, the disentangled tokens $\mathcal{V}_\mathrm{d}$ from DisVisioner is enhanced into multiple tokens:

$$\Upsilon_\mathrm{e}^s = P^s \circ \boldsymbol{\tau}_\mathrm{d}^s,\ \Upsilon_\mathrm{e}^i = P^i \circ \boldsymbol{\tau}_\mathrm{d}^i, \tag{4}$$

where $\tau_\mathrm{d}^s, \tau_\mathrm{d}^i \in \mathbb{R}^{1\times d}$ denote the disentangled subject and the irrelevant feature through DisVisioner. $\Upsilon_\mathrm{e}^s \in \mathbb{R}^{n_s'\times \mathrm{d}}$ is the enriched subject tokens from τ_d^s, and $\Upsilon_\mathrm{e}^i \in \mathbb{R}^{n_i'\times \mathrm{d}}$ is the subject-irrelevant tokens projected from τ_d^i. By enriching the subject token into multiple ones, the abstract disentangled concept is further enriched and enhanced into a more granular representation, improving especially the ID consistency between the synthesized image and the reference image.

Similar to DisVisioner, we also employ Eq. 1 as the training objective to train the projectors in EnVisioner, while keeping the entire SD model and DisVisioner frozen. Additionally, we introduce separate additional cross-attention layers for subject-essential tokens and irrelevant ones. This separate injection strategy further ensures that the subject feature will not be interfered with by other factors. Specifically, given the latent diffusion feature $\boldsymbol{f} \in \mathbb{R}^{(\mathrm{H}\times\mathrm{W})\times d_\mathrm{q}}$ and text instruction $\boldsymbol{c} \in \mathbb{R}^{n_y\times d_\mathrm{k}}$, the cross-attention output $\boldsymbol{f'}$ is derived through three decoupled cross-attention layers, which can be described via the following

Equation:

$$\begin{aligned} \boldsymbol{f}' = \text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \lambda_s \,\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}^s, \boldsymbol{V}^s) + \lambda_i \,\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}^i, \boldsymbol{V}^i), \\ \text{where } K^s = \boldsymbol{w}^s_{\texttt{to\_k}} \circ \Upsilon^s_{\text{e}}, V^s = \boldsymbol{w}^s_{\texttt{to\_v}} \circ \Upsilon^s_{\text{e}}, K^i = \boldsymbol{w}^i_{\texttt{to\_k}} \circ \Upsilon^i_{\text{e}}, V^i = \boldsymbol{w}^i_{\texttt{to\_v}} \circ \Upsilon^i_{\text{e}}, \end{aligned} \tag{5}$$

the Attention$(\cdot)$ is defined in Eq. 2, $\boldsymbol{w}^s_{\texttt{to\_k}}$, $\boldsymbol{w}^s_{\texttt{to\_v}}$, $\boldsymbol{w}^i_{\texttt{to\_k}}$ and $\boldsymbol{w}^i_{\texttt{to\_v}}$ are trainable matrices. During training, weights λ_s and λ_i are fixed at 1.0. During inference, for the purpose of effectively ignoring irrelevant feature Υ^i_{e} , we set $\lambda_i = 0$.

4 Experiments

4.1 Experimental Setup

Training Dataset. The *training set* of OpenImages V6 [19] is employed for training the DisEnvisioner. It contains about 14.61M annotated boxes across 1.74M images. Based on this dataset, we construct 6.82M {prompt, image} pairs for training. The images are cropped and resized (256×256 for DisVisioner and 512×512 for EnVisioner) according to the bounding box annotations. The text prompts are obtained by randomly selecting a CLIP ImageNet template [27] and integrating the annotated class names into it.

Evaluation Dataset and Metrics. The evaluation is carried out on the DreamBooth [32] dataset, which comprises 30 subjects and 158 images in total (4∼6 images per subject). For quantitative evaluation, 25 editing prompts [32] are used for each image. We inference 40 times for each {prompt, image} pair, generating 158,000 customized images for evaluation across 6 metrics.

In alignment with previous methodologies [18,23,32,38], we utilize: 1) CLIP Text-alignment (**C-T**) to measure the instruction response fidelity; 2) CLIP Image-alignment (**C-I**) and 3) DINO Image-alignment (**D-I**) [2] to evaluate the ID consistency with the reference image. Additionally, we introduced two novel metrics aimed at quantifying the effectiveness of minimizing the impact of irrelevant image components. These are: 4) Internal Variance of CLIP Image-alignment (**C-IV**) and 5) Internal Variance of LPIPS (**L-IV**), which measure the variance of the customization results given images containing the same subject under different conditions. Lower values of CLIP-IV and LPIPS-IV indicate a reduced influence of irrelevant subject attributes. In terms of efficiency, we record the 6) Inference-time (**T**) on single NVIDIA A800 GPU for evaluation.

Implementation Details. Our DisEnvisioner is built upon Stable Diffusion v1.5 , employing OpenCLIP ViT-H/14 model as the image/text encoder. During training, DisVisioner is configured with a batch size of 160, a learning rate of 5e-7 at the resolution of 256. We set the token number $n_s = 1$ and $n_i = 1$, for subject-essential feature and subject-irrelevant respectively. The EnVisioner employs a batch size of 40, a learning rate of 1e-4 at the resolution of 512. The enriched token number is $n'_s = 16$ and $n'_i = 4$, with attention scale $\lambda_s = 1.0$ and $\lambda_i = 1.0$. All experiments are conducted on 8 NVIDIA A800 GPUs using

Table 1: Quantitative comparisons with existing methods. The evaluation metrics include text alignment (C-T), identity preservation (C-I, D-I), resistance to irrelevant (C-IV, L-IV), and efficiency (T). Our DisEnvisioner demonstrates better comprehensive performance than other methods. Top results are in bold; second-best are underlined.* For equity, we assess only tuning and inference times, omitting I/O operations. ‡ Setting `ip_scale=0.4` (default `0.6`) for balanced editability and ID consistency.

Method	C-T↑	C-I↑	D-I↑	C-IV↓	L-IV↓	T↓ (s)*
DreamBooth [32]	0.286	0.842	0.849	0.039	0.614	1.12e+3
DisenBooth [3]	0.303	0.760	0.781	0.041	0.586	2.42e+3
ELITE [38]	0.287	0.792	0.770	0.036	0.553	4.12
IP-Adapter, s=0.6‡ [41]	0.275	**0.883**	**0.912**	0.033	0.572	1.98
IP-Adapter, s=0.4‡ [41]	0.309	0.809	0.825	0.029	0.573	1.98
BLIP-Diffusion [20]	0.295	0.785	0.765	0.029	**0.502**	**1.10**
Ours	**0.315**	0.828	0.802	**0.026**	0.522	1.96

Table 2: User study. Participants rank the methods based on four criteria using a round-robin format, with the final scores are normalized before being recorded.

Method	TA↑	IA↑	RE↑	IQ ↑
DreamBooth [32]	0.159	0.152	0.159	0.158
DisenBooth [3]	0.157	0.143	0.156	0.140
ELITE [38]	0.161	0.147	0.154	0.147
IP-Adapter, s=0.6 [41]	0.153	0.177	0.152	0.181
BLIP-Diffusion [20]	0.159	0.162	0.163	0.165
Ours	**0.211**	**0.219**	**0.216**	**0.209**

the AdamW optimizer [24] with a weight decay of 0.01. To enable classifier-free guidance [6], we use a probability of 0.05 to drop the condition, both textual and visual. During inference, λ_i is set to 0 to eliminate irrelevant feature. In addition, we use 50 steps of DDIM sampler [35] and the scale of classifier-free guidance is set to 5.0.

4.2 Quantitative Results

We compare our DisEnvisioner against five leading methods according to metrics specified in Sec. 4.1. For the tuning-based methods, we chose DreamBooth [32] and DisenBooth [3], and implement them within the single-image setting. Our tuning-free comparison covers all available open-source methods, they are ELITE [38], BLIP-Diffusion [20], and IP-Adapter [41]. As demonstrated in Tab. 1, our approach successfully eliminates subject-irrelevant information, enhancing editability significantly. Moreover, since we focus solely on subject-essential features, ignoring the surrounding irrelevant components, the DisEnvisioner tends to generate surrounding environment that aligns more closely with the textual instructions rather than the visual prompt, like the *"in the bucket"* case in Fig. 2. This characteristic keeps our image alignment scores (C-I and D-I) moderate,

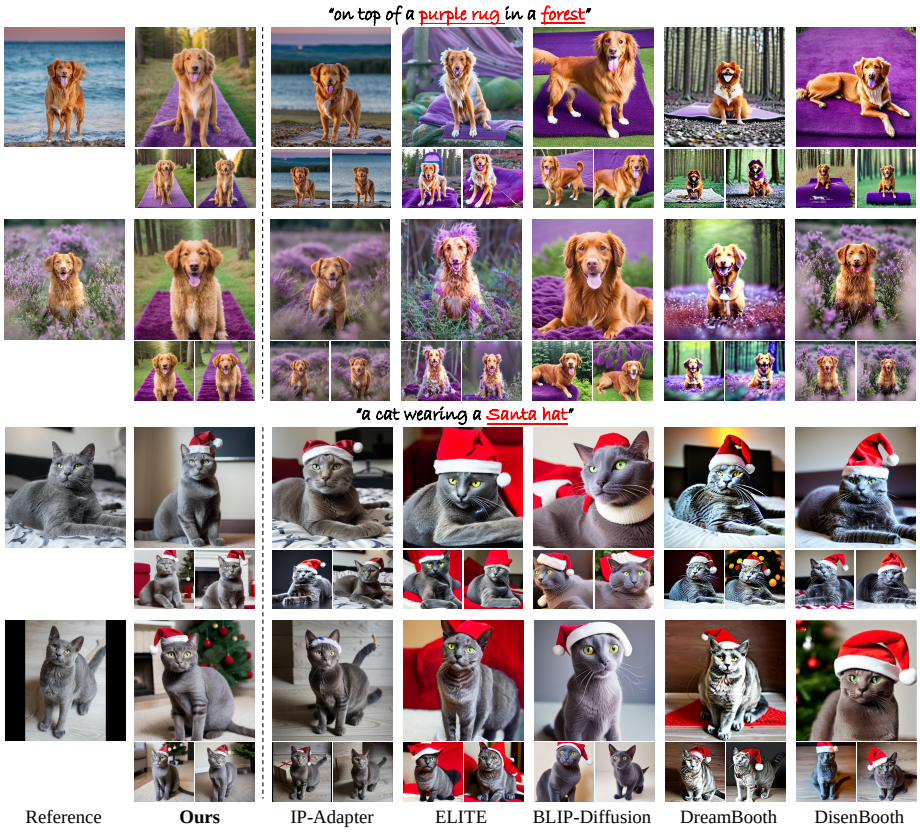


Fig. 5: Qualitative comparison on live subjects. Comparing two live subjects across various scenarios, our method excels in quality, editability, and identity fidelity. Notably, it preserves the subject's posture unaffected by reference images, showcasing our strength in capturing essential features.

unlike IP-Adapter [41] and DreamBooth [32], which exhibit higher scores due to the dominance of surrounding information from the visual prompt into the customization process. In terms of efficiency, thanks to our lightweight and effective design, eliminating the test-time tuning, only 1.96s is required for each customized generation.

We also conduct a user study using a round-robin format, where participants grade each method given the generation outputs across three images of the same subject, in a total of 5 rounds.The metrics are text adherence (**TA**), identity alignment (**IA**), extraneous element resistance (**RE**), and image quality (**IQ**). A total of 69 users, and 345 rounds are recorded. As detailed in Table 2, our method outperforms all baselines in all metrics.

4.3 Qualitative Results

To obtain deeper insight into the proposed DisEnvisioner, we visualize its synthesized images against selected five prevailing methods across two images per

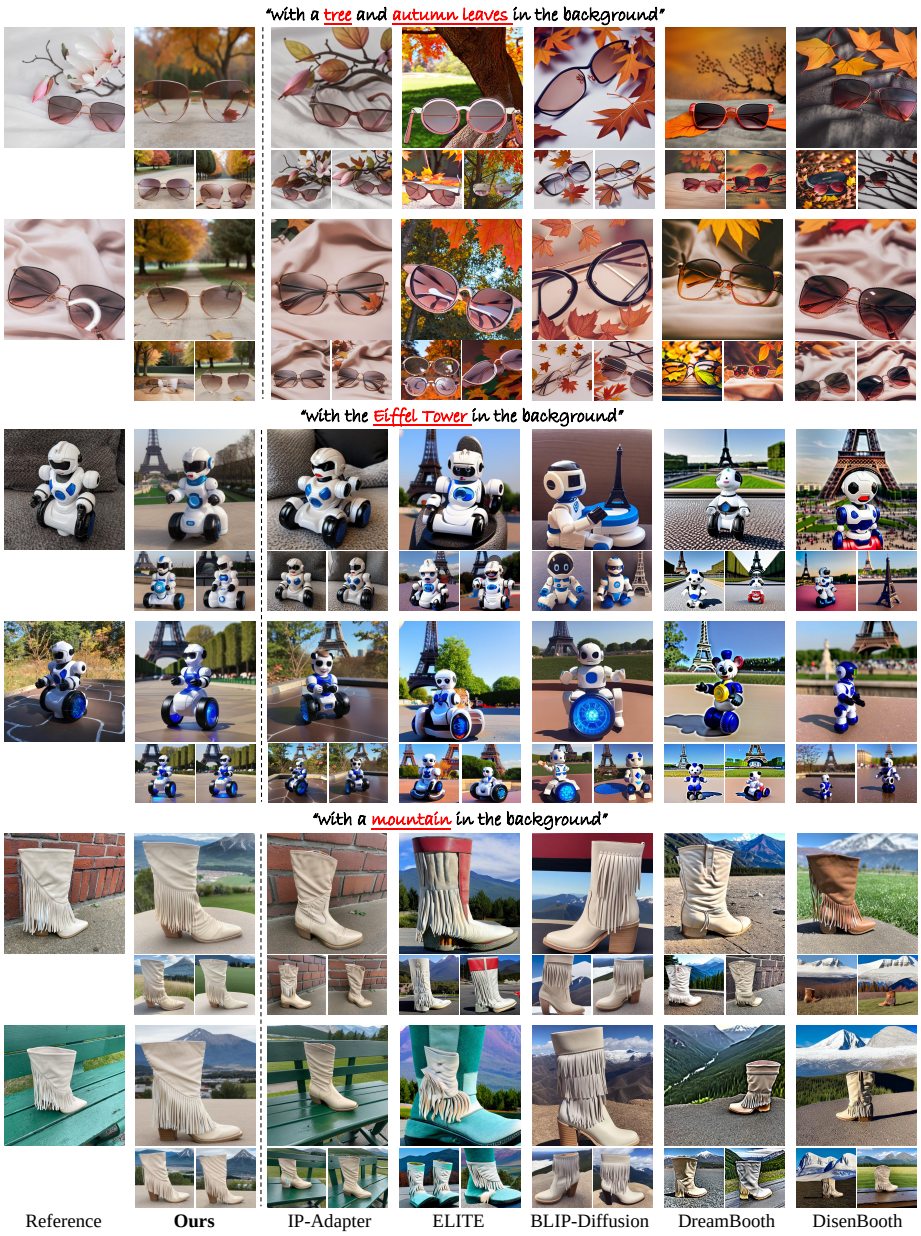
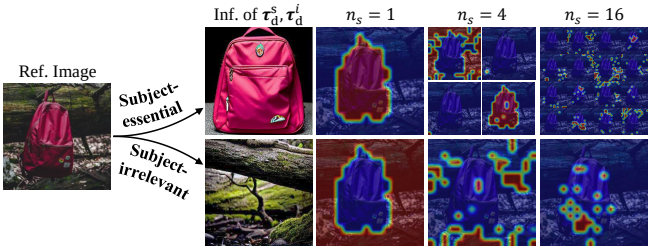


Fig. 6: Qualitative comparison on non-live objects. Besides offering editability based on textual instructions, our method excels in preserving high-level identity, notably in customized objects like robots and boots. Furthermore, images generated by our method are minimally affected by irrelevant elements.



(a) The effect across varying number of subject tokens n_s.

(b) More visualizations of feature disentanglement.

Fig. 7: Comparison between different n_s and more results. (a) Demonstrates that $n_s = 1$ outperforms other configurations with precise attention map and token inference. (b) Showcases DisVisioner's ability to accurately discern subject-essential attributes across diverse scenarios.

subject. Fig. 5 clearly demonstrates DisEnvisioner's superiority in producing high-quality, editable images with strong identity fidelity and resistance to irrelevant elements. Notably, the consistency in animal postures alongside the minimal impact of irrelevant backgrounds across references under the same condition, showcase our robustness to extraneous elements, *i.e.*, only the animal-essential are extracted and preserved. This excellence also applies to non-live objects, as seen in Fig. 6, where exceptional customizations like robots and boots remain true to their essence, free from irrelevant noise.

4.4 Ablation Study

The influence about the number of subject tokens in DisVisioner. As depicted in Fig. 7, the visualized attention maps reveal that with 16 tokens, all tokens fail to accurately identify the region of interest. Reducing the number of tokens to 4 improves focus on essential features but still includes irrelevant features, such as the upper-left token when $n_s = 4$. However, with a single token, the model precisely identifies and represents subject-essential attributes, *i.e.*, clear attention maps and token inference for both subject-essential and subject-irrelevant features. This suggests that an excess of tokens may hinder accurate feature extraction. Additionally, we present more editing results guided by textual instructions in Fig. 8a. The challenge of responding accurately to textual instructions is emerged when the extraction of subject-essential features is imprecise, *i.e.*, resulting in the predominance of subject-irrelevant features during the generation process. Our DisVisioner proposes an optimal strategy that involves encapsulating a complete subject within a single token, thereby facilitating effective extraction of subject-essential features.

(a) Editability under different n_s. **(b)** Effect of EnVisioner. **(c)** Training strategy.

Fig. 8: Ablation Study. (a) Setting $n_s = 1$ achieves precise disentanglement. (b) More granular details are captured by our EnVisioner. (c) Two-stage training can effectively disentangle and enrich the subject for customized generation. "EV" denotes the proposed EnVisioner module. "stg." stands for training stage. The prompt used in this experiment is "*a photo of S* on the beach*".

The effect of enriched subject representation. As shown in Fig. 8b, we generate images of customized subjects using the same prompt to explore the effect of enriched subject representation in our proposed EnVisioner. The images enhanced by EnVisioner exhibit finer details and higher identity consistency, with notably boosted overall image quality. In contrast, the absence of EnVisioner compromises the quality of customization significantly, highlighting its importance in providing detailed subject representations.

The comparison with single-stage and current two-stage training. For single-stage training, we combine the training processes of DisVisioner and EnVisioner, aiming for simultaneous learning the disentanglement and enrichment of subject-essential features. In Fig. 8c, the single-stage strategy fails to represent the subject and merely response to text instructions. We hypothesize that the challenge of simultaneously balancing the learning for both disentanglement and enrichment is significant. Nonetheless, the two-stage training strategy of our DisVisioner—separating disentanglement and enrichment—proves to be more effective in achieving high-quality customization.

5 Conclusion

In this paper, we propose **DisEnvisioner**, which is characterized by its emphasis on the interpretation of subject-essential attributes for high-quality image customization. DisEnvisioner effectively identifies and enhances the subject-essential feature while filtering out other irrelevant information, enabling exceptional image personalization without cumbersome tuning or relying on multiple reference images. Through both the quantitative and qualitative evaluations, alongside the user study, we demonstrate DisEnvisioner's superior performance in customization quality and efficient inference time, offering a promising solution for practical applications.

References

1. Arar, M., Voynov, A., Hertz, A., Avrahami, O., Fruchter, S., Pritch, Y., Cohen-Or, D., Shamir, A.: Palp: Prompt aligned personalization of text-to-image models. arXiv preprint arXiv:2401.06105 (2024) 2
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 9
3. Chen, H., Zhang, Y., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374 (2023) 2, 3, 5, 10
4. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 2, 4, 6
5. Chen, L., Zhao, M., Liu, Y., Ding, M., Song, Y., Wang, S., Wang, X., Yang, H., Liu, J., Du, K., et al.: Photoverse: Tuning-free image customization with text-to-image diffusion models. arXiv preprint arXiv:2309.05793 (2023) 2, 3, 5
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 10
7. Dong, Z., Wei, P., Lin, L.: Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. arXiv preprint arXiv:2211.11337 (2022) 2, 5
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 2, 5
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 4
10. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. arXiv preprint arXiv:2303.11305 (2023) 2, 5
11. He, J., Zhou, Y., Zhang, Q., Peng, J., Shen, Y., Sun, X., Chen, C., Ji, R.: Pixelfolder: An efficient progressive pixel synthesis network for image generation. arXiv preprint arXiv:2204.00833 (2022) 4
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 4, 6
13. Hua, M., Liu, J., Ding, F., Liu, W., Wu, J., He, Q.: Dreamtuner: Single image is enough for subject-driven generation. arXiv preprint arXiv:2312.13691 (2023) 5
14. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023) 2, 5
15. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021) 4
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 4
17. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 4

18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023) 2, 5, 8, 9
19. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 9
20. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36** (2024) 2, 3, 5, 7, 10
21. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 5
22. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:2304.05653 (2023) 7
23. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. arXiv preprint arXiv:2312.04461 (2023) 2, 5, 9
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
25. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410 (2023) 2, 3, 5
26. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 4, 5, 6
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 9
28. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 2, 4, 5, 6
29. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 2, 4, 6
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 4, 5, 6
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 6
32. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023) 2, 3, 5, 9, 10, 11

33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 2, 4, 5, 6
34. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023) 2, 5
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 10
36. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023) 2
37. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024) 2, 5
38. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023) 2, 3, 5, 7, 8, 9, 10
39. Wu, H., Wang, M., Zhou, W., Hu, Y., Li, H.: Learning token-based representation for image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2703–2711 (2022) 4, 7
40. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1316–1324 (2018) 4
41. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023) 2, 3, 5, 7, 10, 11
42. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 833–842 (2021) 4
43. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 5907–5915 (2017) 4
44. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1947–1962 (2018) 4

DisEnvisioner: Disentangled and Enriched Visual Prompt for Customized Image Generation Supplementary Material

Under-review

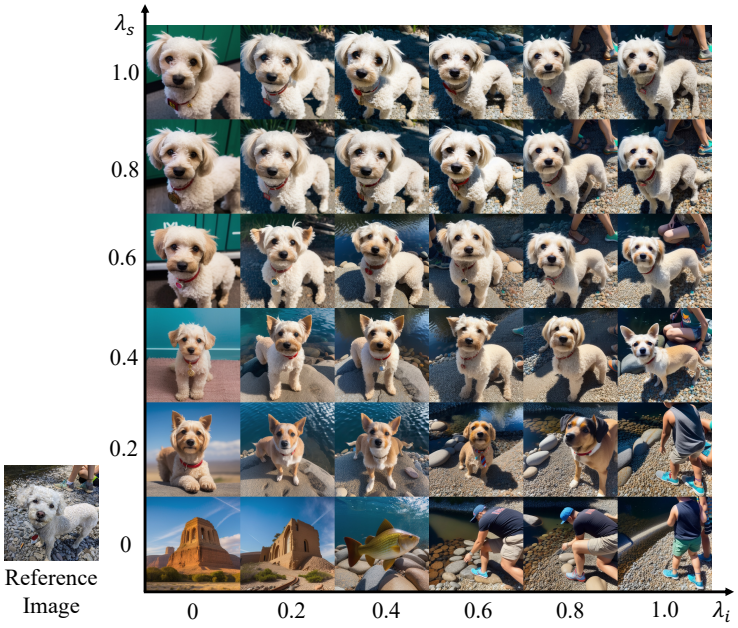
https://disenvisioner.github.io

A Effect of λ_s and λ_i

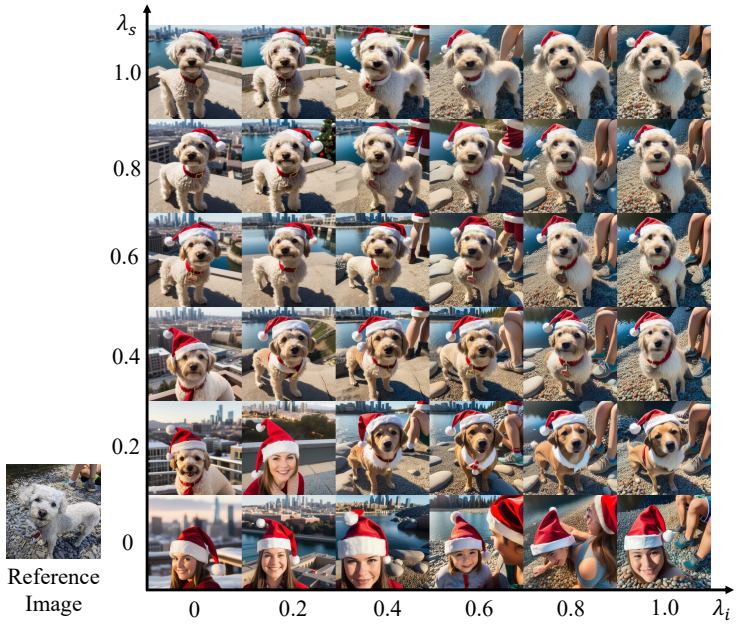
As defined in Eq. 5 of the main paper, the weights λ_s and λ_i serve to modulate the integration of information that is essential and irrelevant to the subject in the given visual prompt. To thoroughly assess their effect, we adjust their values systematically from 0 to 1.0 throughout the image customization process.

We generate images under varying settings of λ_s and λ_i employing both empty and non-empty (editing) prompts. Fig. 1 demonstrates that as λ_i decreases progressively (moving from the right to the left columns), the presence of subject-irrelevant disturbances in the images notably declines. Conversely, enhancing λ_s (moving from the bottom to the top rows) brings more pronounced consistency in subject identity between the reference and generated image. When both λ_s and λ_i are reduced to their minimum value, *i.e.*, $\lambda_s = 0$ and $\lambda_i = 0$, the generated images are solely influenced by the textual prompt, without incorporating any information from the reference image. To further explore the role of additional information in the prompt, we generate images with specific class names included in the prompts. As illustrated in Fig. 1b and Fig. 2, particularly in terms of identity consistency, no matter the category-guidance (for instance, specifying the class name "dog") is provided or not, it does not alter the customization quality. This indicates that our approach effectively deciphers and extracts subject-essential attributes from the reference image. By doing so, it renders additional semantic information redundant, which is then eliminated.

It is also evident that when image generation focuses exclusively on solely subject-essential features ($\lambda_s = 1.0$ and $\lambda_i = 0$) or solely on purely subject-irrelevant features ($\lambda_s = 0$ and $\lambda_i = 1.0$), the reproduction of the subject and the extraneous surrounding content is achieved independently, devoid of any interference from one another. This phenomenon confirms the proficiency of our DisEnvisioner in precisely segregating and enriching subject-essential and non-essential features. It highlights the DisEnvisioner's exceptional customization performance without the need for test-time tuning, and relying solely on a single reference image.



(a) With empty prompt.



(b) With non-empty prompt: "wearing a santa hat with a city in the background".

Fig. 1: Effect of varying λ_s and λ_i with empty and non-empty prompts without providing class names.

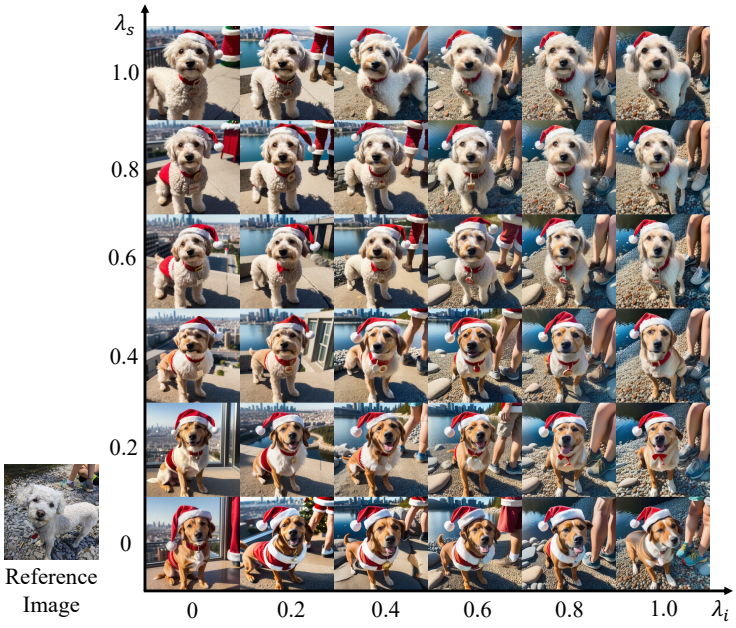


Fig. 2: Effect of varying λ_s and λ_i with providing class name in prompt. The prompt is "a **dog** is wearing a santa hat with a city in the background".

B Experimental Details

B.1 Grading of User Study

During each round of the user study, rather than **ranking** our DisEnvisioner and five other existing methods from the best to the worst, users are expected to **assign grades** from 0 to 5 to each method according to specific metrics. In practice, all participants have the complete freedom to grade any method with any score based on their personal judgment. After a total of 345 rounds of evaluation, the best-performing method often receives the highest scores, while scores for other methods are frequently identical due to similar customization quality. Additionally, it is uncommon for users to assign scores as low as 0 or 1. Although the grading differences among methods are not particularly large, DisEnvisioner consistently outperforms others competitors across all evaluation criteria.

B.2 Training Data

In our experiments, we utilize the *training set* of OpenImages V6 [1] as our training dataset. Based on this dataset, we construct {prompt, image} pairs for training. As depicted in Fig. 3, the training images are derived by cropping and resizing the raw images in accordance with the bounding box annotations. To

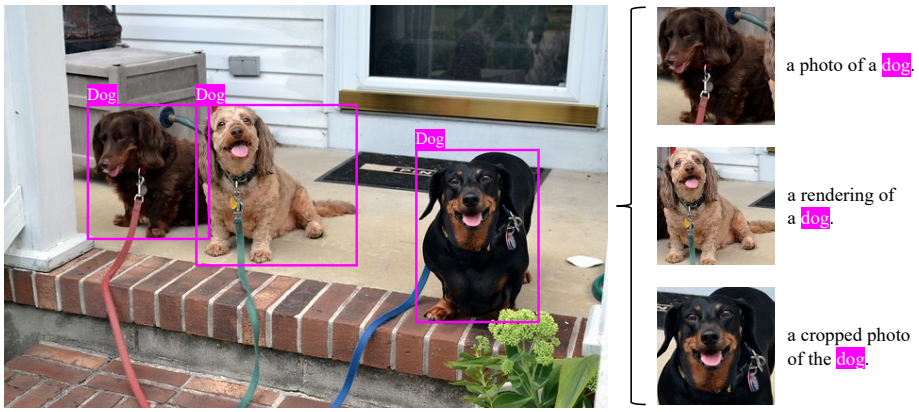


Fig. 3: Examples of training data. The training images are derived by cropping and resizing the raw images in accordance with the bounding box annotations.

ensure the quality of the training images, we further filter the cropped images: a cropped image is considered as unsatisfactory and therefore excluded if its area is greater than 80% or less than 2% of the original image's area. As a result, out of 14.61 million annotated bounding boxes, we obtain **6.82 million** {prompt, image} pairs. The text prompts are selected randomly from a CLIP ImageNet template [2] and integrated with labelled class names. The complete list of CLIP templates is provided below:

– "a photo of a S^*",
– "a rendering of a S^*",
– "a cropped photo of the S^*",
– "the photo of a S^*",
– "a photo of a clean S^*",
– "a photo of a dirty S^*",
– "a dark photo of the S^*",
– "a photo of my S^*",
– "a photo of the cool S^*",
– "a close-up photo of a S^*",
– "a bright photo of the S^*",
– "a cropped photo of a S^*",
– "a photo of the S^*",
– "a good photo of the S^*",
– "a photo of one S^*",
– "a close-up photo of the S^*",
– "a rendition of the S^*",
– "a photo of the clean S^*",
– "a rendition of a S^*",
– "a photo of a nice S^*",
– "a good photo of a S^*",

Fig. 4: Examples of testing data. We randomly selected 8 subjects of the 30, and three images are shown for each subject.

– "a photo of the nice S^*",
– "a photo of the small S^*",
– "a photo of the weird S^*",
– "a photo of the large S^*",
– "a photo of a cool S^*",
– "a photo of a small S^*"

B.3 Testing Data

For evaluation, we adopt images and editing prompts from DreamBooth [3]. It contains a total of 158 images spanning 30 diverse categories, including dog, cat, robot, boot, etc. Fig. 4 showcases a selection of these image samples. The the complete set of editing prompts for live subjects is detailed below:

– "a S^* in the jungle"
– "a S^* in the snow"
– "a S^* on the beach"
– "a S^* on a cobblestone street"
– "a S^* on top of pink fabric"
– "a S^* on top of a wooden floor"
– "a S^* with a city in the background"
– "a S^* with a mountain in the background"
– "a S^* with a blue house in the background"
– "a S^* on top of a purple rug in a forest"
– "a S^* with a wheat field in the background"
– "a S^* with a tree and autumn leaves in the background"
– "a S^* with the Eiffel Tower in the background"
– "a S^* floating on top of water"
– "a S^* floating in an ocean of milk"
– "a S^* on top of green grass with sunflowers around it"
– "a S^* on top of a mirror"

– "a S^* on top of the sidewalk in a crowded street"
– "a S^* on top of a dirt road"
– "a S^* on top of a white rug"
– "a red S^*"
– "a purple S^*"
– "a shiny S^*"
– "a wet S^*"
– "a cube shaped S^*"

Additionally, we also enumerate the full set of editing prompts for non-live subjects:

– "a S^* in the jungle"
– "a S^* in the snow"
– "a S^* on the beach"
– "a S^* on a cobblestone street"
– "a S^* on top of pink fabric"
– "a S^* on top of a wooden floor"
– "a S^* with a city in the background"
– "a S^* with a mountain in the background"
– "a S^* with a blue house in the background"
– "a S^* on top of a purple rug in a forest"
– "a S^* wearing a red hat"
– "a S^* wearing a santa hat"
– "a S^* wearing a rainbow scarf"
– "a S^* wearing a black top hat and a monocle"
– "a S^* in a chef outfit"
– "a S^* in a firefighter outfit"
– "a S^* in a police outfit"
– "a S^* wearing pink glasses"
– "a S^* wearing a yellow shirt"
– "a S^* in a purple wizard outfit"
– "a red S^*"
– "a purple S^*"
– "a shiny S^*"
– "a wet S^*"
– "a cube shaped $S*$"

References

1. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 3
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4

3. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023) 5