Non-confusing Generation of Customized Concepts in Diffusion Models

Wang Lin [1 *] **Jingyuan Chen** [1 2] **Jiaxin Shi** [3 *] **Yichen Zhu** [1] **Chen Liang** [4] **Junzhong Miao** [5] **Tao Jin** [1]
Zhou Zhao [1] **Fei Wu** [1] **Shuicheng Yan** [6] **Hanwang Zhang** [6 7]

Abstract

We tackle the common challenge of inter-concept visual confusion in compositional concept generation using text-guided diffusion models (TGDMs). It becomes even more pronounced in the generation of customized concepts, due to the scarcity of user-provided concept visual examples. By revisiting the two major stages leading to the success of TGDMs—1) contrastive image-language pre-training (CLIP) for text encoder that encodes visual semantics, and 2) training TGDM that decodes the textual embeddings into pixels—we point that existing customized generation methods only focus on fine-tuning the second stage while overlooking the first one. To this end, we propose a simple yet effective solution called CLIF: contrastive image-language fine-tuning. Specifically, given a few samples of customized concepts, we obtain non-confusing textual embeddings of a concept by fine-tuning CLIP via contrasting a concept and the over-segmented visual regions of other concepts. Experimental results demonstrate the effectiveness of CLIF in preventing the confusion of multi-customized concept generation. Project page: https://clif-official.github.io/clif.

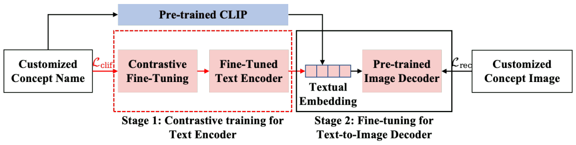

Pre-trained CLIP
Customized Concept Name
$\mathcal{L}_{\text{clif}}$
Contrastive Fine-Tuning
Fine-Tuned Text Encoder
Textual Embedding
Pre-trained Image Decoder
$\mathcal{L}_{\text{rec}}$
Customized Concept Image
Stage 1: Contrastive training for Text Encoder
Stage 2: Fine-tuning for Text-to-Image Decoder

Figure 1. The black line and box denote the prevailing pipeline of customized generation methods. Our contribution is to contrast the textual embeddings of customized concepts in the Text Encoder stage, which is shown in the red line and dashed box.

1. Introduction

We are interested in customizing a text-guided diffusion model (TGDM), *e.g.*, Stable Diffusion (Rombach et al., 2022), to generate compositions of user-provided concepts. For example, as shown in Figure 3, given a few images of *Hector Rivera* and *Tang Seng*, we can generate imaginary compositions by the prompt "*Hector Rivera* snuggled up in *Tang Seng*". Existing customized generation methods are based on fine-tuning a pre-trained TGDM, where the tunable parameters include the textual embeddings of the new concept names (Gal et al., 2022; Voynov et al., 2023) and/or LoRAs (Hu et al., 2021) on the generation backbone (Ruiz et al., 2023; Kumari et al., 2023). In this way, the fine-tuned TGDM is expected to memorize the visual concepts and generalize them to unseen compositions. However, an ever-elusive challenge of the generalization is the inter-concept confusion shown in Figure 3. This visual defect is even more severe when the interaction is spatially cluttered, such as "snuggling" and "riding motorcycle".

Recent findings of visualizing the role of the image-text cross-attention in pre-trained TGDM (Tewel et al., 2023; Patashnik et al., 2023) show that the textual embeddings (V-values) control "what to draw" and the cross-attention map (Q-K softmax) tells "where to draw". Inspired by this, we believe that the cause of the confusion is mainly due to the confusing textual embeddings of concepts. To see this, we revisit the two stages in training TGDM (Figure 1 and Section 3):

- **Text Encoder**: we obtain the text encoder from contrastive image-language pre-training (CLIP) on large-scale image-text pairs (Radford et al., 2021; Schuhmann et al., 2021). In this way, the token embedding of a concept token carries its visual features.
- **Text-to-Image Decoder**: the textual embedding is decoded into pixels by the cross-attention between textual and visual embeddings in the U-net decoder. In Section 2, we discuss that almost all the customized generation methods focus on this stage.

We can show the confusion of common concepts in the existing vocabulary of Stable Diffusion by measuring the confusion degree of each concept. We use a sentence of two concepts as the prompt for generation (*e.g.*, "a cat and a

[*] Equal contribution [1] Zhejiang University [2] Corresponding Author [3] Huawei Cloud Computing [4] Tsinghua University [5] Harbin Institute of Technology [6] Skywork AISingapore [7] Nanyang Technological University. Correspondence to: Jingyuan Chen <jingyuanchen@zju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

dog”), and calculate the probability of their presence in the generated image using an object detector. As illustrated in Figure 2, when the two concept token embeddings are far away (*e.g.*, “octopus” and “cat”), the composition is rarely confused; when they are close, confusion is common.

Can we use existing methods (Chefer et al., 2023; Li et al., 2023a; Huang et al., 2023; Zhang et al., 2023; Mou et al., 2023) of de-confusing the above TGDM-known concepts to mitigate the confusion of TGDM-unknown, customized concepts? Unfortunately, the answer is no. This is because those methods rely on the assumption that the visual features and textual embeddings are well-aligned—but it does not hold for few-shot customized concepts. Then, what about the de-confusing methods especially designed for fine-tuning customized concepts (Gu et al., 2023; Po et al., 2023)? Still no, because fine-tuning the second stage should not contrast the textual embeddings of different concepts too sharply to prevent overfitting, *e.g.*, making the pre-trained TGDM lose the original ability of text control.

We propose a simple yet effective method called **CLIF**: Contrastive Language-Image *Fine-tuning*, to tackle the confusion challenge directly in the first stage by contrasting the textual embeddings of customized concepts (Figure 1). We first present an over-segmented concept dataset that augments the visual examples of customized concepts into a large number of language-image contrastive data (Section 4.1). Thus, by applying contrast fine-tuning of the text encoder on the augmented data, we can fundamentally eliminate the confusion in the concept token embeddings (Section 4.2). Then, in the second stage, we reconstruct the images of concepts to fine-tune both the text embeddings with the text encoder frozen and the Unet of TGDM (Section 4.3).

In Section 5, to demonstrate the non-confusing effectiveness of CLIF, we jointly customize 18 user-provided characters and compare CLIF with prior SOTA methods as shown in Figure 3. Different from prior methods (Gu et al., 2023), these 18 characters are more fair and proper for evaluation, because they are from less popular movies, which is rare for a pre-trained TGDM. We conduct extensive ablations to analyze how each CLIF’s breakdowns mitigate confusion in multi-concept generation.

2. Related Work

Concept Customization. The goal of customized generation is to implant the user-provided visual examples of concepts into a pre-trained TGDM to generate various renditions of the concepts vividly guided by text prompts. Existing works can be categorized into three types based on the fine-tuned modules of the text-guided diffusion model (image decoder): 1) Text embedding (Gal et al., 2022; Voynov et al., 2023; Yuan et al., 2023; Alaluf et al., 2023): this type fine-tunes the text embeddings of customized concepts to align with input images while freezing the diffusion model. 2) Decoder (*e.g.*, U-net) (Ruiz et al., 2023): this type explicitly binds the concept with rare words by fine-tuning the entire diffusion model. Additionally, (Ryu, 2023) adopts a low-rank adapter (LoRA) (Hu et al., 2021) for concept tuning, which is lightweight and can achieve comparable fidelity to full-weight tuning. 3) Joint methods (Kumari et al., 2023) that fine-tune the above two. However, these three modules are only in the image decoder stage of TGDM. In contrast, our CLIF recalls the neglected first stage and fine-tunes both of the stages, thereby improving the ability to customize and compose more diverse concepts.

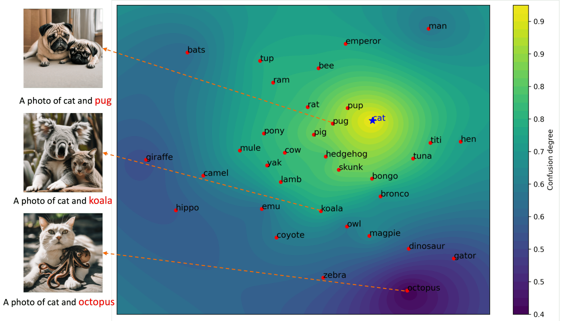


Figure 2. Visualization of confusion in embedding space with “cat” as an anchor point, see Appendix for details. It shows an evident correlation between confusion and embedding distance.

Confusion in Multi-Concept Customization. Some of the above works (Kumari et al., 2023; Gu et al., 2023; Liu et al., 2023) focus on injecting multiple concepts into TGDM, facing the challenge of inter-concept confusion. To tackle the challenge, current approaches can be categorized into the following 3 types: 1) Generative semantic nursing (Chefer et al., 2023; Li et al., 2023a; Hertz et al., 2022): this type optimize or edit the cross-attention maps in the generative process during inference time. Although for common concepts these methods can correct minor errors in attention maps and enable more accurate multi-concept synthesis, they are not salvageable for the confused attention maps generated by TGDM-unknown customized concepts. 2) Spatial control (Huang et al., 2023; Zhang et al., 2023; Mou et al., 2023): this type directly integrates spatial layout as a pixel-level specification. However, obtaining additional spatial conditions is costly and proves ineffective in complex interactions involving large overlapping areas. 3) Token embedding (Balaji et al., 2022; Liu et al., 2022; Feng et al., 2022): this type aims to improve the prompt alignment on the text side by combining T5 and CLIP text encoders or utilizes language parsers to associate attributes solely with the corresponding concepts. In contrast, our approach does not require additional components, and we directly fine-tune the concept token embeddings to alleviate the confusion and achieve better prompt alignment.

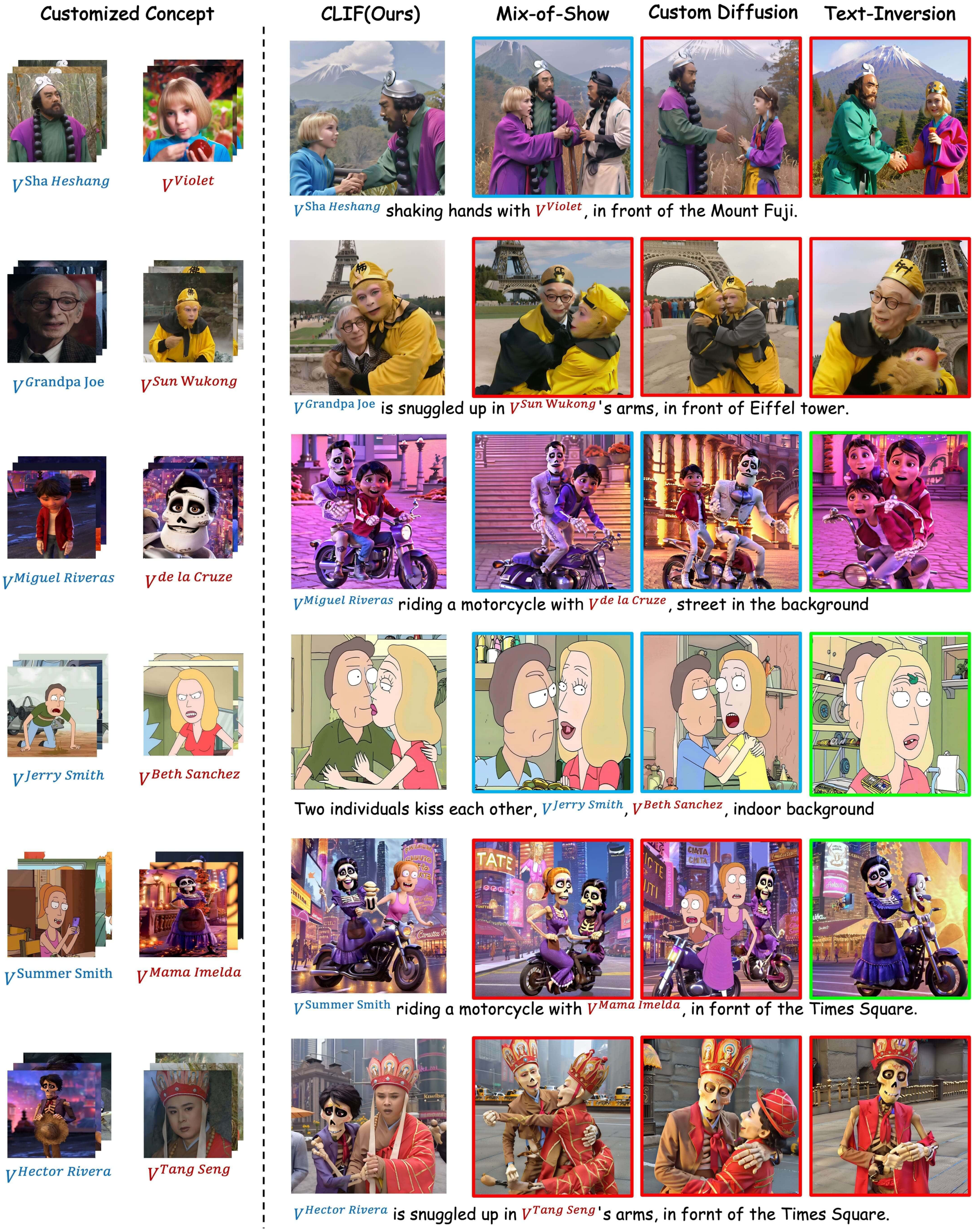

Customized Concept
CLIF(Ours)
Mix-of-Show
Custom Diffusion
Text-Inversion
$V^{Sha\ Heshang}$
V^{Violet}
$V^{Sha\ Heshang}$ shaking hands with V^{Violet}, in front of the Mount Fuji.
$V^{Grandpa\ Joe}$
$V^{Sun\ Wukong}$
$V^{Grandpa\ Joe}$ is snuggled up in $V^{Sun\ Wukong}$'s arms, in front of Eiffel tower.
$V^{Miguel\ Riveras}$
$V^{de\ la\ Cruze}$
$V^{Miguel\ Riveras}$ riding a motorcycle with $V^{de\ la\ Cruze}$, street in the background
$V^{Jerry\ Smith}$
$V^{Beth\ Sanchez}$
Two individuals kiss each other, $V^{Jerry\ Smith}$, $V^{Beth\ Sanchez}$, indoor background
$V^{Summer\ Smith}$
$V^{Mama\ Imelda}$
$V^{Summer\ Smith}$ riding a motorcycle with $V^{Mama\ Imelda}$, in fornt of the Times Square.
$V^{Hector\ Rivera}$
$V^{Tang\ Seng}$
$V^{Hector\ Rivera}$ is snuggled up in $V^{Tang\ Seng}$'s arms, in fornt of the Times Square.

Figure 3. **Visualization of multi-concept customization for challenging cases.** When the concepts to be customized belong to categories with high semantic similarity (all belonging to "humans" superclass), or when there is large regional overlap (*e.g.*, the second and third rows) or combinations across styles (*e.g.*, 2D combined with 3D characters in the fifth and sixth rows), the baseline methods suffer from, identity loss (red border), attribute leaking (blue border), or concept missing (green border), which are effectively circumvented by CLIF.

3. Preliminary

CLIP as Text Encoder. For a concept with name c (might be associated with some text prompts p such as "a photo of c"), the Text Encoder transforms it into a textual embedding V^* that carries the concept's visual features as:

$$\mathrm{V}^* := \texttt{Text-Encoder}(p), \tag{1}$$

where $:=$ denotes that V^* is the token embedding selected from the position of c in p. Existing methods employ CLIP to train Text Encoder by contrastive pre-training on a large-scale dataset of text-image pairs. Specifically, for each text-image pair (p, q) in a batch, the goal of CLIP is to maximize the similarity between the text and the corresponding image while minimizing the similarity with another non-matching image. The training loss can be formulated as:

$$\mathcal{L}_{clip} = -\log \frac{\exp(\mathrm{s}(\texttt{Text-Encoder}(p), f(q)))}{\sum_{n^-} \exp(\mathrm{s}(\texttt{Text-Encoder}(p), f(q^-)))}, \tag{2}$$

where $f(\cdot)$ is image encoder and $\mathrm{s}(\cdot,\cdot)$ represents the similarity in the feature space, usually using cosine similarity.

By minimizing this loss on large-scale text-image pairs, the generated token embeddings will be well-aligned with the corresponding visual features.

Stable Diffusion as Text-to-Image Decoder. After obtaining the concept token embedding V^*, Text-to-Image Decoder generates an image x_{gen} with an initial noise map $\varepsilon \sim \mathcal{N}(0, 1)$ as:

$$\mathrm{x}_{gen} = \texttt{Image-Decoder}_{x_\theta}(\varepsilon, \mathrm{V}^*). \tag{3}$$

Specifically, we use Stable Diffusion (Rombach et al., 2022) as the image decoder x_θ. The concept token embedding V^* is decoded into images by the cross-attention between textual and visual embeddings in the U-net decoder. The cross-attention layers project V^* into keys $\mathbf{K}$ and values $\mathbf{V}$, while the queries $\mathbf{Q}$ are derived from the intermediate features of U-net. The attention maps are then calculated by $\mathbf{A} = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d}})$, where d denotes the hidden state dimension. Finally, pixel features are comprised of the values $\mathbf{V}$, weighted by the attention maps $\mathbf{A}$ as: $\mathbf{A} \cdot \mathbf{V}$. (Tewel et al., 2023) find that the keys $\mathbf{K}$ control the compositional structure of the generated image, and the values $\mathbf{V}$ control the appearance of image components.

Stable Diffusion is trained using a squared error loss to denoise a variably-noised image or latent code z_t as:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathrm{x},\mathrm{V}^*,\varepsilon,t}\left[\|x_\theta(z_t, \mathrm{V}^*) - \mathrm{x}\|_2^2\right], \tag{4}$$

where x is the ground-truth image, and $z_t = \sqrt{\alpha_t}\mathrm{x} + \sqrt{1-\alpha_t}\varepsilon$ is the noisy input at time-step t where α_t is related to a fixed variance schedule. Such training objective can be simplified as a reconstruction loss.

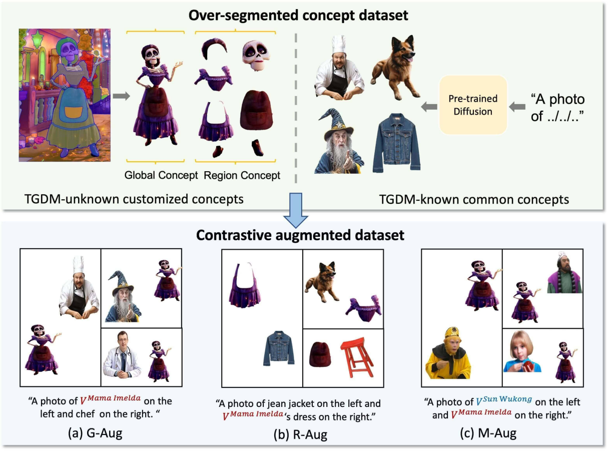

Over-segmented concept dataset
Global Concept
Region Concept
TGDM-unknown customized concepts
Pre-trained Diffusion
"A photo of ../../.."
TGDM-known common concepts
Contrastive augmented dataset
"A photo of $V^{Mama\ Imelda}$ on the left and chef on the right. "
(a) G-Aug
"A photo of jean jacket on the left and $V^{Mama\ Imelda}$'s dress on the right."
(b) R-Aug
"A photo of $V^{Sun\ Wukong}$ on the left and $V^{Mama\ Imelda}$ on the right."
(c) M-Aug

Figure 4. Pipeline of training data curation. We mix the customized concepts and common concepts at instance-level and segmentation-level, to help decouple multi-concept token embeddings which can eliminate the confusion issues.

4. Method: CLIF

We aim to compose multiple customized concepts in one image with complex interaction. To eliminate the confusion issues of baselines (Gu et al., 2023; Kumari et al., 2023) in Figure 3, we re-visit the two stages of TGDM and then propose a two stage fine-tuning method as shown in Figure 5, to make the embeddings of different concepts more contrastive. To support an effective contrastive fine-tuning with very limited user-provided concept images, we design an over-segmented method with multi-granularity for training data curation to ensure the learned concept embeddings are separated globally and locally.

4.1. Training Data Curation

Upon revisiting the two stages of TGDM, it becomes apparent that the textual embeddings of newly added customized concepts (*i.e.*, TGDM-unknown concepts) in existing customized generation methods are solely trained in the Text-To-Image Decoder stage and not in the Text Encoder stage. They overlook that during image decoder training, the reconstruction loss $\mathcal{L}_{rec}$ aims to reconstruct visual features at the pixel level and should not be mistaken for the contrastive learning loss $\mathcal{L}_{clip}$, which decouples relationships between concepts. This results in the under-trained concept embedding and the projected V being confusing, ultimately leading to confusion in generated multi-concept images.

Our idea is to fine-tune the customized concept embedding by contrastive learning similar to CLIP's to reduce its confusion. To this end, we propose a simple technique to construct a large number of image-text pairs for fine-tuning the text encoder and image decoder because the customized concepts are derived from the user-provided limited image data. Specifically, we decompose the confusion into three issues: 1) **Identity Preservation**, 2) **Attribute Binding**, and 3) **Concept Attendance**. To this end, as shown in Figure 4, we construct three augmentation data: 1) Global Augmentation (**G-Aug**), 2) Region Augmentation (**R-Aug**), and 3) Mix Augmentation (**M-Aug**), to address the above issues respectively. Below, we describe in detail the motivation and process for constructing each type of data.

G-Aug for Identity Preservation We draw the following two observations regarding existing approaches: **1)** In Figure 2, we find that the degree of confusion and the Euclidean distance between embeddings show a correlation, which suggests that it is necessary to pull embeddings that use the same initialization (*e.g.*, man) farther apart in the embedding space. **2)** In Figure 3, we observe that confusing concept embeddings will fail to preserve identity information, such as the erroneous fusion of *TangSeng* and *Hector Rivera*'s visual appearance.

Based on the aforementioned observations, we attribute the identity loss to inter-concept confusion. To reduce it, we propose global augmentation. Specifically, 1) we first segment the concept from the original images using SAM (Kirillov et al., 2023), to filter irrelevant contexts like the background; 2) then we utilize a pre-trained diffusion model to generate some general concepts such as policeman, dog, denim jacket, etc., as new contexts; 3) finally, we combine these segmented concepts with the general concepts to generate a large number of text-image pairs with different contexts.

The global augmentation is designed to fine-tune the text encoder. With the supervision contained in text paired with images, the concept token carries its visual features, excluding the visual features of other concepts, and inter-concept confusion is mitigated.

R-Aug for Attribute Binding. In Figure 3, we observe that the semantic components of a concept can also be confused, leading to attribute leakage, *e.g. Jerry Smith*'s top incorrectly uses the blue color of his pants.

Based on this observation, we attribute the attribute leakage to intra-concept confusion. To reduce it, we propose regional augmentation. Specifically, 1) we first use GPT-4 to caption as much as possible the characterization in the concept such as hair, necklaces, hats, and so on; 2) then, we further segment the global concepts to get the regional concept, and label it with the results from GPT-4; 3) finally, we follow the global concepts' process to generate a large number of region-based text-image pairs.

The regional augmentation is designed to fine-tune text encoder and is complementary to global augmentation, working together to achieve a non-confusing text embedding.

M-Aug for Concept Attendance. We make the following observations about missing concepts: **1)** In Figure 3, we observe that there are often dominant concepts, while other non-dominant concepts *e.g.*, *Miguel Ricveras* and *Summer Smith*, are not always successfully generated in images, sometimes missing entirely. **2)** We find the text-to-image decoder has defined the dominance in one of the concept embedding vectors beforehand, which we call dominant bias. For example, when given the prompt "a photo of a cat and a pug", stable diffusion tends to generate two pugs due to the bias in the pre-train data. The presence of dominant bias results in missing concepts in multiple concept generation where non-dominant concepts are often lost or produce redundant dominant concepts.

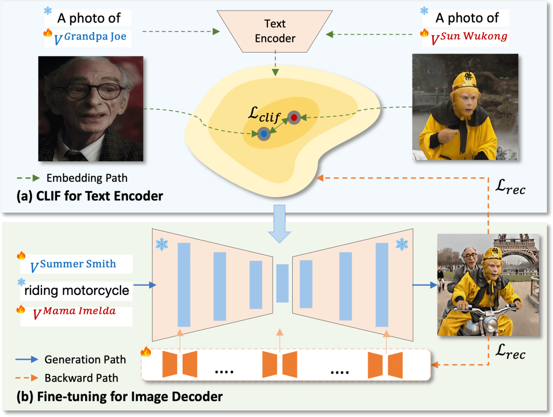


Figure 5. Our two stage framework for multi-concept learning. We first fine-tune the text encoder to get contrastive concept embeddings, and then fine-tune the text-to-image decoder to synthesizing non-confusing images.

Based on the aforementioned observations, we attribute concept missing to the dominant bias. To reduce it, we propose mixed augmentation. Specifically, 1) we first segment the concept from the original images similar to global augmentation; 2) then, we randomly scale and place the segmented concept with another one on either left or right side of the image and generate corresponding text prompts.

The mixed augmentation is designed to fine-tune the text-to-image decoder. To present the model with correctly mixed image samples, the text-to-image decoder is enforced to synthesize multi-concepts equally.

4.2. CLIF for Text Encoder

We investigate embedding tuning (Gal et al., 2022; Voynov et al., 2023) in concept customization. Given a text prompt containing the customized concept *Jerry Smith* or *Willy Wonka*, the supercategory embedding, *e.g.*, "man", is used to initialize both concepts.

Our goal is to eliminate confusion in text embeddings by fine-tuning the customized concepts contrastively. The fine-tuning approach is similar to CLIP's training, optimizing a symmetric cross-entropy loss over these image-text similar-

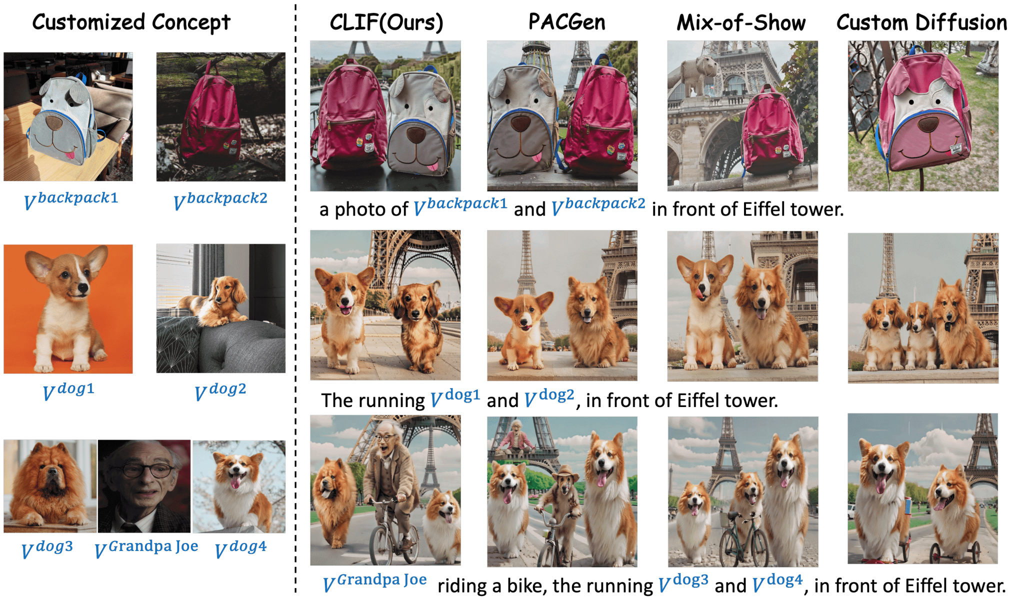

Customized Concept
CLIF(Ours)
PACGen
Mix-of-Show
Custom Diffusion
V^backpack1
V^backpack2
a photo of V^backpack1 and V^backpack2 in front of Eiffel tower.
V^dog1
V^dog2
The running V^dog1 and V^dog2, in front of Eiffel tower.
V^dog3
V^Grandpa Joe
V^dog4
V^Grandpa Joe riding a bike, the running V^dog3 and V^dog4, in front of Eiffel tower.

Figure 6. Comparison of different methods on DreamBench.

ity scores as follows:

$$\mathcal{L}_{clif} = -\log \frac{\exp(\mathrm{s}(\texttt{Text-Encoder}(p_a), f(q_a)))}{\sum_{q_a^-} \exp(\mathrm{s}(\texttt{Text-Encoder}(p_a), f(q_a^-))))}, \tag{5}$$

where $\mathrm{s}(p_a, q_a)$ is the similarity score between the augmented image q_a and corresponding prompt p_a, while $\mathrm{s}(p_a, q_a^-)$ is the similarity score for negative pair in the batch.

4.3. Fine-tuning for Text-to-Image Decoder

After applying contrastive fine-tuning to the customized concepts in the text encoder, we have obtained decoupled concept embeddings c and can generate customized images with well-maintained identity information.

Our goal is to generate images that contain multi-customized concepts. However, existing weight fine-tuning techniques for the diffusion model are insufficient to achieve this goal by dominant bias. With the mixed concepts augmentation, we freeze the text encoder and jointly train multiple concepts with a shared LoRA W_θ in the U-net θ. Since embeddings are already separated in the text encoder, during joint training the model can avoid concept conflict (Gu et al., 2023).

5. Experimental Results

5.1. Experimental Setup

Task and Dataset. We aim to address the challenge of preventing the confusion of multiple customized concepts. To comprehensively verify the effectiveness of CLIF, we consider concepts as *characters* comprising a range of visual elements (*e.g.*, "face", "hat", and "clothes") that need to be preserved. We curate a dataset consisting of 18 representative characters, including 9 real-world, 4 3D-animated, and 5 2D-animated. Each of them possesses unique visual appearances that must be preserved in the customized generation. In our experiment, we will demonstrate the ability of CLIF to generate imaginative compositions of these characters with complex interactions involving spatial clutter (*e.g.*, "snuggling" and "riding motorcycle").

Baselines. We compare CLIF against state-of-the-art baselines: Text-Inversion (Gal et al., 2022), Custom Diffusion (Kumari et al., 2023), Dreambooth (Ruiz et al., 2023), and Mix-of-Show (Gu et al., 2023). Moreover, to demonstrate the generalizability of the proposed strategy in CLIF, we integrate it with Text-Inversion and Custom Diffusion. Note that for fair comparison, all methods do not incorporate additional spatial constraints as in (Zhang et al., 2023).

5.2. Qualitative Comparison

We compare CLIF with Mix-of-Show, Custom Diffusion (Custom for short), and Text-Inversion (TI for short) for multi-concept customized generation in Figure 3. The baseline methods suffer from, identity loss (highlighted in red box), attribute leaking (highlighted in blue box), or concept missing (highlighted in green box). TI and Custom implicitly delegate the task of disentangling multi-concept token embeddings to the Text-to-Image decoder. However, this paradigm is limited to the reconstruction loss which only encodes the concept's visual features into the embeddings without contrasting it with other token embeddings. On the other hand, Mix-of-Show uses gradient fusion to merge multiple separately fine-tuned concepts, which aims to preserve the single concept identity in the fused model rather than decoupling each other.

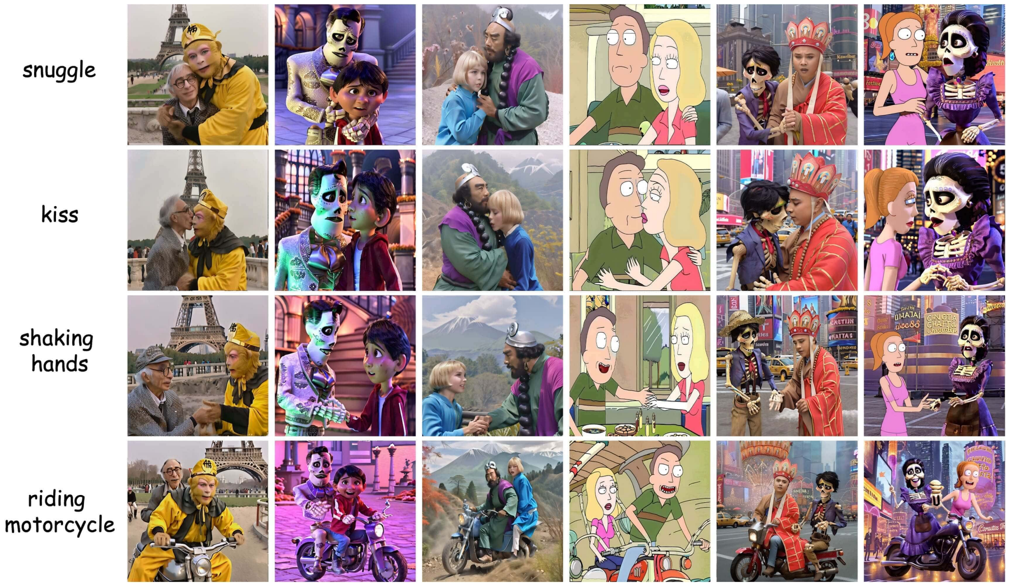


Figure 7. Results for multi-concept customized generation using CLIF. Our approach is able to generate non-confusion images containing multiple characters with complex interactions, without the need for additional spatial constraints (*e.g.*, layout, mask, sketch).

To demonstrate the generalization of our method, we conduct experiments on Dreambench. We incorporate an additional baseline PACGen(Li et al., 2023b) which is a multi-subject driven generation method in Dreambench. As shown in Figure 6, the experimental results demonstrate that our approach generalizes well to Dreambench.

Based on the results of CLIF for multi-concept customization in Figure 7, we can find that: 1) Benefiting from contrastive fine-tuning, CLIF can accurately generate each character even when all the customized concepts belong to the same superclass (*i.e.*, "human"). This indicates CLIF's ability to differentiate between similar concepts within the same category; and 2) In the case of multi-concept generation with complex interactions (such as "snuggling" and "kissing"), previous approaches exhibit more serious confusion problems. Due to the spatial clutter between concepts, these approaches tend to draw visual features of multiple concepts in duplicate areas, which results in the wrong combination or loss of concepts. CLIF can handle complex interactions natively through contrastive fine-tuned text embedding, without relying on additional spatial constraints, making it a more cost-effective solution. 3) These results demonstrate the effectiveness of directly contrasting the textual embeddings of customized concepts in the first stage, which reduces the confusion significantly. Generating more than 2 objects can be achieved by simply extending our dataset. However, it has some limitations which are discussed in Appendix A.6.

5.3. Quantitative Comparison

Following Custom Diffusion (Kumari et al., 2023), we utilize the text/image encoder of CLIP to assess text alignment and image alignment. A detailed evaluation setting is provided in Appendix A.5.1.

Table 1. Text-alignment and image-alignment vary between single-concept and multi-concept generation scenarios.

Methods	Text Alignment		Image Alignment	
	Single	Multi	Single	Multi
TI	0.604(-2.6%)	0.507(-9.2%)	0.726(-6.1%)	0.708(-5.3%)
DreamBooth	0.617(-1.3%)	0.523(-7.6%)	0.754(-3.3%)	0.711(-5.0%)
Custom	0.622(-0.8%)	0.511(-8.8%)	0.749(-3.8%)	0.715(-4.6%)
Mix-of-Show	0.629(-0.1%)	0.526(-7.3%)	0.757(-3.0%)	0.713(-4.8%)
TI+CLIF	0.631(+0.1%)	0.528(-7.1%)	0.751(-3.6%)	0.726(-3.5%)
Custom+CLIF	0.657(+2.7%)	0.535(-6.4%)	0.774(-1.3%)	0.730(-3.1%)
CLIF (ours)	0.630	0.599	0.787	0.761

Table 2. Quantitative ablation study in single-concept and multi-concept generation scenarios.

Methods	Text Alignment		Image Alignment	
	Single	Multi	Single	Multi
CLIF	0.630	0.599	0.787	0.761
w/o G-Aug	0.604(-2.6%)	0.566(-3.3%)	0.751(-3.6%)	0.729(-3.2%)
w/o R-Aug	0.627(-0.3%)	0.591(-0.8%)	0.774(-1.3%)	0.753(-0.8%)
w/o M-Aug	0.612(-1.8%)	0.579(-2.0%)	0.768(-1.9%)	0.740(-2.1%)

Based on the results presented in Table 1, we can find that: 1) For single-concept, compared with TI which encodes all concept details within the text embedding, CLIF and other baseline methods benefit from tuning the diffusion weight and exhibit superior image alignment; 2) For single-concept, CLIF exhibits superior image alignment compared to baseline methods. This is attributed to the contrastive fine-tuning of textual embeddings in CLIF. Contrastive fine-tuning not only helps mitigate confusion but also aligns the image and the concept, enabling the token embeddings of concept names to capture more detailed visual features. Furthermore, CLIF maintains comparable text alignment, indicating that our approach can enhance high identity preservation with-

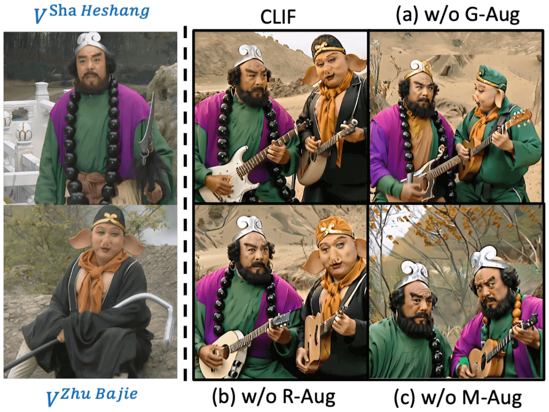


Figure 8. Visualization of ablation results for augmented data type.

out compromising composability; 3) For multi-concept, the superior performance of CLIF across all metrics highlights its ability to effectively capture concept characteristics and preserve concept distinct identity in multi-concept compositions; and 4) The integration of CLIF with TI and Custom shows a remarkable improvement, indicating that contrastive fine-tuning of the text embedding is indeed effective and generalizable.

5.4. Ablation Study

As mentioned in Section 4.1, three critical capabilities must be addressed to mitigate the confusion problem in customized generation, namely, *identity preservation*, *attribute binding*, and *concept attendance*. The following ablation experiments will demonstrate how CLIF is specifically designed to enhance these capabilities.

Effectiveness of Global Augmentation (G-Aug). In Figure 8, it can be observed that without global augmentation, the generation results suffer from identity departure (*e.g.*, *Zhu Bajie* incorrectly uses *Sha Heshang*'s coat color). In contrast, our CLIF successfully mitigates this issue by pushing the textual embeddings of the two concepts farther apart through contrastive supervision, which prevents the Text-to-Image Decoder from confusing the characters' visual appearances during generation. This interpretation is also supported by the quantitative results in Table 2, which indicate that the image alignment in multi-concept scenario decreases from 0.761 to 0.729 (-3.2%).

Effectiveness of Region Augmentation (R-Aug). Even when multiple concepts are decoupled from each other, the generated concepts may still struggle with attribute leakage (*e.g.*, *Zhu Bajie*'s hat incorrectly using the color of his tie). Therefore, we use region augmentation to fine-tune each component of the embedding by binding each sub-region of the concept to the token embedding, which not only decouples the confusion of similar components between concepts but also decouples the confusion within concepts.

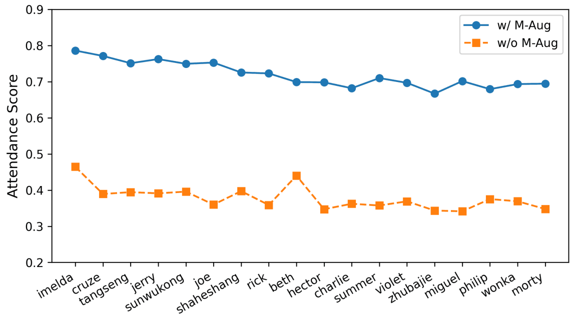


Figure 9. Effects of using M-Aug for concept missing.

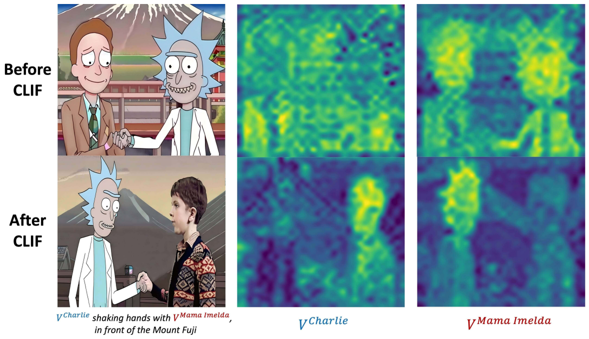

Before CLIF
After CLIF
$V^{Charlie}$ shaking hands with $V^{Mama\ Imelda}$, in front of the Mount Fuji
$V^{Charlie}$
$V^{Mama\ Imelda}$

Figure 10. Visualization of attention map for concept embeddings.

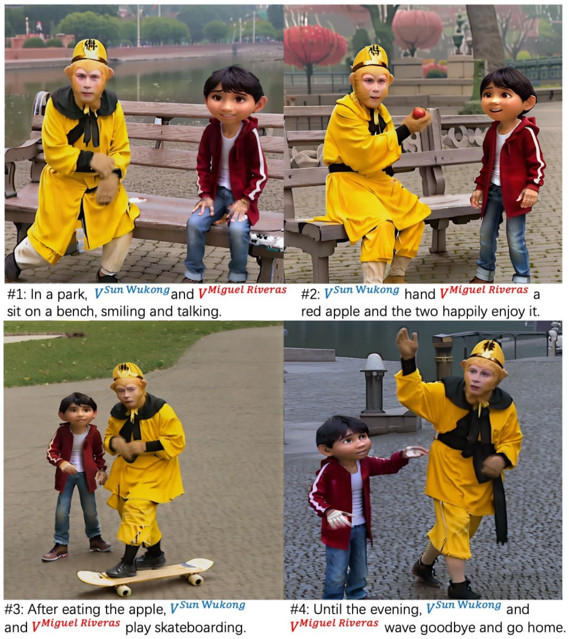
According to the results in Table 2, region augmentation improves attribute binding during multi-concept generation, boosting both single and multi-concept image alignment.

Effectiveness of Mix Augmentation (M-Aug). A common problem in multi-concept generation is concept missing. As shown in Figure 8, two instances of *Sha Heshang* were repeatedly generated while *Zhu Bajie* was missing. We attribute this issue to dominant bias and address it through mix augmentation. To highlight the impact of mix augmentation on addressing dominant bias, we propose an intuitive metric *attendance*, as described in Appendix A.3. The results are shown in Figure 9, where the attendance score for all 18 customized concepts is significantly improved, demonstrating the effectiveness of mix augmentation.

Cross-Attention in Diffusion. We visualize concept tokens' cross-attention maps before and after CLIF in Figure 10. The results indicate that contrastive fine-tuning effectively decouples the token embeddings of multiple customized concepts, eliminates confusion in cross-attention maps, and generates high-quality images.

5.5. Application to storytelling

In addition, after obtaining the non-confusion customized token embedding by contrastive fine-tuning, the user can further apply it to the tasks such as story generation, video generation, and so on. As shown in Figure 11, users can combine different customized characters to generate imaginative creations.



#1: In a park, $V^{Sun\ Wukong}$ and $V^{Miguel\ Riveras}$ sit on a bench, smiling and talking.

#2: $V^{Sun\ Wukong}$ hand $V^{Miguel\ Riveras}$ a red apple and the two happily enjoy it.

#3: After eating the apple, $V^{Sun\ Wukong}$ and $V^{Miguel\ Riveras}$ play skateboarding.

#4: Until the evening, $V^{Sun\ Wukong}$ and $V^{Miguel\ Riveras}$ wave goodbye and go home.

Figure 11. A comic story generated by CLIF.

6. Conclusions

Our CLIF approach marks a significant advancement in the customized generative models. By fine-tuning both the text encoder and text-to-image decoder stages, CLIF successfully addresses the persistent challenge of concept confusion, particularly in multi-concept generation scenarios. This technique preserves the integrity of each concept, ensuring that each retains its unique identity even amidst complex and cluttered interactions. Our extensive experiments and ablation studies underscore the efficacy of CLIF, establishing it as a powerful and versatile tool for customized concept generation. The improvements observed in both single and multi-concept customizations indicate the broad applicability and potential of our method in various creative and practical applications.

Impact Statement

Ethical Impacts This study does not raise any ethical concerns. The research does not involve subjective assessments or the use of private data. Only publicly available datasets are utilized for experimentation.

Expected Societal Implications We aim to address the confusion in customized concepts and facilitate the generation of higher-quality customized multi-concept images. A primary ethical concern is the potential misuse of this technology, notably in creating deepfakes, which can result in misinformation, privacy violations, and other harmful outcomes. To mitigate these risks, the establishment of robust ethical guidelines and continuous monitoring is essential. The concern raised here is a common one, not just for our method but across various multi-concept customization techniques. A viable strategy to lessen these risks might be the implementation of tactics akin to those used in anti-dreambooth(Van Le et al., 2023). This approach involves adding minor noise disturbances to the shared images, thereby hindering the customization process. Furthermore, embedding invisible watermarks in the generated images can serve as a deterrent against misuse and ensure that they are not used without due acknowledgment.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62037001, No.62307032) and the Shanghai Rising-Star Program (23QA1409000).

References

Alaluf, Y., Richardson, E., Metzer, G., and Cohen-Or, D. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42 (6):1–10, 2023.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Gu, Y., Wang, X., Wu, J. Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., and Zhou, J. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and 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.

Li, Y., Keuper, M., Zhang, D., and Khoreva, A. Divide & bind your attention for improved generative semantic nursing. *arXiv preprint arXiv:2307.10864*, 2023a.

Li, Y., Liu, H., Wen, Y., and Lee, Y. J. Generate anything anywhere in any scene. *arXiv preprint arXiv:2306.17154*, 2023b.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022.

Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., and Cao, Y. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., and Cohen-Or, D. Localizing object-level shape variations with text-to-image diffusion models. *arXiv preprint arXiv:2303.11306*, 2023.

Po, R., Yang, G., Aberman, K., and Wetzstein, G. Orthogonal adaptation for modular customization of diffusion models. *arXiv preprint arXiv:2312.02432*, 2023.

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.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and 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.

Ryu, S. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Tewel, Y., Gal, R., Chechik, G., and Atzmon, Y. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

Van Le, T., Phung, H., Nguyen, T. H., Dao, Q., Tran, N. N., and Tran, A. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.

Voynov, A., Chu, Q., Cohen-Or, D., and Aberman, K. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

Yuan, G., Cun, X., Zhang, Y., Li, M., Qi, C., Wang, X., Shan, Y., and Zheng, H. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

A. Appendix

A.1. Dataset and Implementation Details

A.1.1. Dataset Details

Global Augmentation Data. For each customized concept we collect 10-20 images. We apply global data augmentation through a three-step process.

- **step-1.** We first segment the original images using the SAM model. Considering that each image is character-oriented, we filter the segmented images with an area less than 10%. For the remaining segmented images, we simply use the CLIP model to classify whether it is a person or not.
- **step-2.** Then we generate 100 common concepts including characters, animals, objects, *etc.* by GPT-4. In order to prevent overfitting, we generat 20 images for each concept and require a solid color background in the generated text prompts to facilitate the subsequent segmentation process.
- **step-3.** Finally, we randomly combine our customized concepts with each generated image.

In practice, we construct about 40,000 image-text pairs for contrastive fine-tuning.

Region Augmentation Data. For region-augmented data, considering the fact that there are numerous regions that can be segmented from each image, but the majority of these regions contain noise, we opt to use an intact concept image rather than the original image in order to extract all potential regions. This approach ensures that the segmented region images are all integral components of the concept. Specifically, we extract 3-10 sub-regions for each concept. To augment the data for these sub-regions, we have divided the process into two steps.

- **step-1.** First, we use GPT-4 to characterize as many objects as possible present in the segmented intact concept image, such as hair, necklaces, hats, and so on. This provides us a list of objects on which we label each segmented region image using CLIP.
- **step-2.** Then we use the pre-trained diffusion model to generate images of the objects in the list and the segmented obtained regions are combined to obtain the enhanced image data and text prompts.

In practice, we construct about 10,000 region image-text pairs.

Mix Augmentation Data. Based on global augmented data, we randomly combine them with each other. To optimize training time efficiency, we limit the combination to two images per concept with other concepts. This resulted in a total of 612 mix augmented data points being used in our study.

A.1.2. Implementation Details

Implementation Details. The implementation process for the text encoder involves fine-tuning it on augmented data, following a similar approach as CLIP, with a learning rate of 1e-4. Once this is completed, the text encoder is then frozen, and the fine-tuning process continues for the text embedding along with the LoRA layer. As part of the LoRA tuning, we integrate the LoRA layer into the linear layer within all attention modules of the U-net, utilizing a rank of $r = 8$. The Adam optimizer is utilized for both text embeddings and diffusion model parameters, with a learning rate of 2e-4.

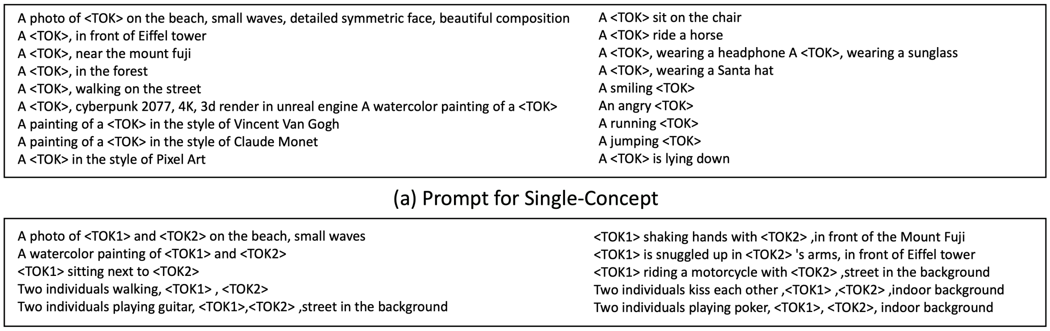
Sample Details. All experiments and evaluations make use of the DDPM with 50 sampling steps. To ensure consistency and filter out undesired variations in diffusion models, we follow the approach outlined in (Gu et al., 2023) by employing the same negative prompt for both our method and the comparison methods during sampling. The negative prompt used is "long body, low-res, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality."

Running Times. The process of tuning concept embeddings in text encoder typically requires approximately 4-5 hours using four Nvidia-A100 GPUs, accounting for variations in data volume. In the case of shared LoRA weight tuning, it takes 10 hours on four Nvidia-A100 GPUs to tune 18 concepts within the pre-trained model.

A.2. Measurement of Confusion Degree

To highlight why we need to fine-tune the embeddings of customized tokens, we visualize the confusion between token embeddings in the hidden space, using "cat" as an anchor point.

First, we extract the features of each token and perform dimensionality reduction by t-SNE to approximate their relationship



A photo of <TOK> on the beach, small waves, detailed symmetric face, beautiful composition
A <TOK>, in front of Eiffel tower
A <TOK>, near the mount fuji
A <TOK>, in the forest
A <TOK>, walking on the street
A <TOK>, cyberpunk 2077, 4K, 3d render in unreal engine A watercolor painting of a <TOK>
A painting of a <TOK> in the style of Vincent Van Gogh
A painting of a <TOK> in the style of Claude Monet
A <TOK> in the style of Pixel Art

A <TOK> sit on the chair
A <TOK> ride a horse
A <TOK>, wearing a headphone A <TOK>, wearing a sunglass
A <TOK>, wearing a Santa hat
A smiling <TOK>
An angry <TOK>
A running <TOK>
A jumping <TOK>
A <TOK> is lying down

(a) Prompt for Single-Concept

A photo of <TOK1> and <TOK2> on the beach, small waves
A watercolor painting of <TOK1> and <TOK2>
<TOK1> sitting next to <TOK2>
Two individuals walking, <TOK1> , <TOK2>
Two individuals playing guitar, <TOK1>,<TOK2> ,street in the background

<TOK1> shaking hands with <TOK2> ,in front of the Mount Fuji
<TOK1> is snuggled up in <TOK2> 's arms, in front of Eiffel tower
<TOK1> riding a motorcycle with <TOK2> ,street in the background
Two individuals kiss each other ,<TOK1> ,<TOK2> ,indoor background
Two individuals playing poker, <TOK1>, <TOK2>, indoor background

(b) Prompt for Multi-Concept

Figure 12. Summarization of our evaluation prompts for each concept.

in the high-dimensional hidden space. Then we create the text prompt "a photo of a cat and a Tok" and use it to generate images as conditions, generating 4 images for each prompt. We then utilize a pre-trained object detector[1] to identify the two concepts, defining the confusion score as:

$$\varphi = 1 - [(|Box^{Cat}_{Cat} - Box^{Tok}_{Cat}| + |Box^{Cat}_{Tok} - Box^{Tok}_{Tok}|)/2], \tag{6}$$

where Box^{Cat}_{Cat} denotes the confidence score of the cat in the bounding box that is detected to be a cat and Box^{Tok}_{Cat} denotes the confidence score of the Tok in the bounding box that is detected to be a cat. If the two concepts are completely decoupled, φ = 0; when the two concepts are completely confused such as cats and cats, φ = 1.

A.3. Measurement of Attendance

To evaluate the improvement of mix augmentation data for missing concepts, we design *attendance* based on image alignment. Specifically, for a concept, we measure the similarity between the concept's reference image and the image generated from the text prompt containing that concept and others. In practice, for each concept, we measure its attendance in the images generated separately from the other 17 concepts and take the average as the final attendance score.

A.4. Scalability Compared to Decentralized Learning-based Approaches

There are two ways to perform multi-concept customized generation, joint training and decentralized learning. Custom Diffusion suggests that co-training multiple concepts yields better results. However, the disadvantage is that the co-training approach lacks scalability, and each newly added concept needs to be re-trained with the already fine-tuned concepts.

Another approach is decentralized learning, such as Mix-of-Show, where each concept is learned independently, and then multiple concepts are fused to obtain a new generative model. This approach has the advantage of avoiding repetitive training of concepts but also brings additional training overhead for fusion. For instance, in Mix-of-Show, 10-15 minutes of training is required to fuse 3 concepts. Furthermore, there is an extra storage overhead, as a new generative model needs to be fused and stored for each combination of concepts. For example, in this paper, we have 18 customized concepts, resulting in C^{18}_{2} = 153 merged models for all two-by-two combinations.

In summary, both co-training and decentralized learning have their advantages and disadvantages in terms of efficiency and performance. However, our research primarily focuses on addressing the issue of confusion in handling multiple concepts.

A.5. Quantitative and Qualitative Evaluation

A.5.1. Evaluation Setting

Our evaluation focuses on investigating each concept in the single-concept generation and the multi-concept generation. To assess the performance, we employ the evaluation metric, which includes image-alignment and text-alignment, as outlined in Custom Diffusion(Kumari et al., 2023). Specifically, for text-alignment, we evaluate the text-image similarity of the sampled image with the corresponding sample prompt in the CLIP feature space(Radford et al., 2021) by the CLIP-Score toolkit[2]. For image-alignment, we evaluate the pairwise image similarity between the sampled image and the target concept

[1] https://portal.vision.cognitive.azure.com/demo/generic-object-detection

[2] https://github.com/jmhessel/clipscore

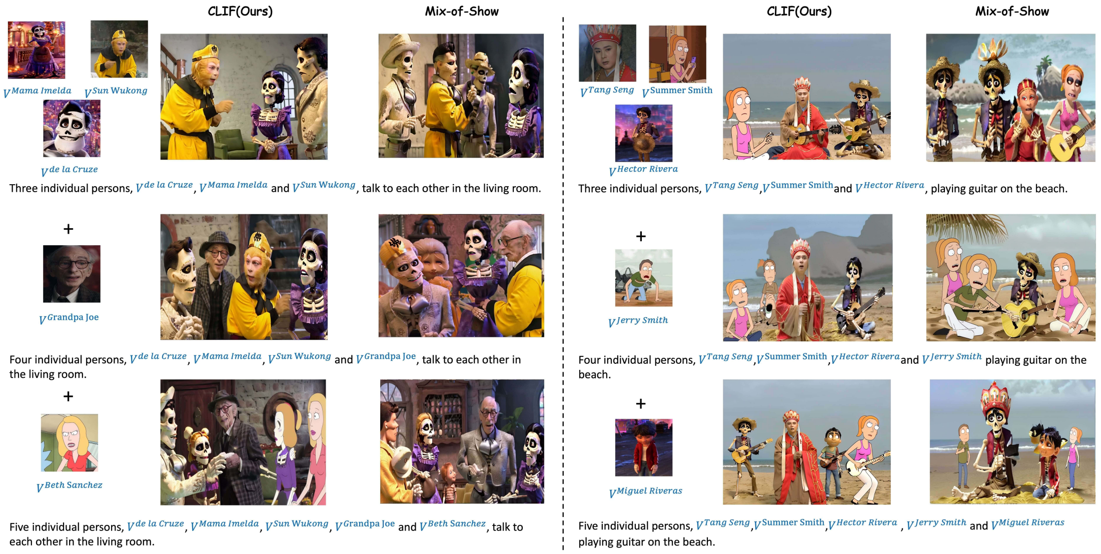

CLIF(Ours)
Mix-of-Show
$V^{\text{Mama Imelda}}$
$V^{\text{Sun Wukong}}$
$V^{\text{de la Cruze}}$
Three individual persons, $V^{\text{de la Cruze}}$, $V^{\text{Mama Imelda}}$ and $V^{\text{Sun Wukong}}$, talk to each other in the living room.
+
$V^{\text{Grandpa Joe}}$
Four individual persons, $V^{\text{de la Cruze}}$, $V^{\text{Mama Imelda}}$, $V^{\text{Sun Wukong}}$ and $V^{\text{Grandpa Joe}}$, talk to each other in the living room.
+
$V^{\text{Beth Sanchez}}$
Five individual persons, $V^{\text{de la Cruze}}$, $V^{\text{Mama Imelda}}$, $V^{\text{Sun Wukong}}$, $V^{\text{Grandpa Joe}}$ and $V^{\text{Beth Sanchez}}$, talk to each other in the living room.
CLIF(Ours)
Mix-of-Show
$V^{\text{Tang Seng}}$
$V^{\text{Summer Smith}}$
$V^{\text{Hector Rivera}}$
Three individual persons, $V^{\text{Tang Seng}}$,$V^{\text{Summer Smith}}$and $V^{\text{Hector Rivera}}$, playing guitar on the beach.
+
$V^{\text{Jerry Smith}}$
Four individual persons, $V^{\text{Tang Seng}}$,$V^{\text{Summer Smith}}$,$V^{\text{Hector Rivera}}$and $V^{\text{Jerry Smith}}$ playing guitar on the beach.
+
$V^{\text{Miguel Riveras}}$
Five individual persons, $V^{\text{Tang Seng}}$,$V^{\text{Summer Smith}}$,$V^{\text{Hector Rivera}}$, $V^{\text{Jerry Smith}}$ and $V^{\text{Miguel Riveras}}$ playing guitar on the beach.

Figure 13. More concepts customized results.

data in the CLIP Image feature space.

For single-concept generation, we utilize 20 evaluation prompts, some of which are borrowed from previous work (Gu et al., 2023). We sample 50 images for each prompt, ensuring reproducibility by fixing the seed within the range of [1, 50]. This yields a total of 1,000 images for single-concept. The evaluation prompts for each concept are presented in Figure 12 (a).

For multi-concept generation, we utilize 10 evaluation prompts, which contain some complex interactions to better evaluate the model's capabilities. We train 18 concepts together, pairing them together to generate $C_2^{18} = 153$ combinations. The evaluation prompts for each concept are presented in Figure 12 (b). We sample 10 images for each prompt and each combination. This yields a total of 15,300 images for multi-concept generation.

A.5.2. More Concepts Results

We supplement results for the customized generation of 3-5 concepts, revealing a gradual increase in confusion as more concepts are added. However, our method still remains clearly superior to other methods, as shown in Figure 13. While CLIF alleviates the problem of confusion between concepts, we are constrained by the ability of CLIP's text encoder to comprehend long texts and the ability of the Diffusion model to follow text prompts, which leads to the inability of customized generation with more concepts.

We present additional multi-concept generation results of CLIF in Figure 14. CLIF demonstrates a superior ability to preserve concept identity and offers a wider range of customized concepts.

A.6. Limitation and Future Work

The limitation involves the generation of more concepts. *e.g.*, 4, 5 or more. While CLIF alleviates the problem of confusion between concepts, we are constrained by the ability of CLIP's text encoder to comprehend long texts and the ability of Stable Diffusion to follow text prompts, leading to the inability of customized generation with more concepts.

CLIF empowers diffusion models to generate customized concepts without confusion, an advancement that can enhance story generation by maintaining character identity information, which is essential for enabling character interaction. Moreover, CLIF is not restricted to image generation models; it can also mitigate confusion in the generation of personalized videos or 3D assets. This technology could lead to more creative and efficient production processes for the media industry advertising and marketing.

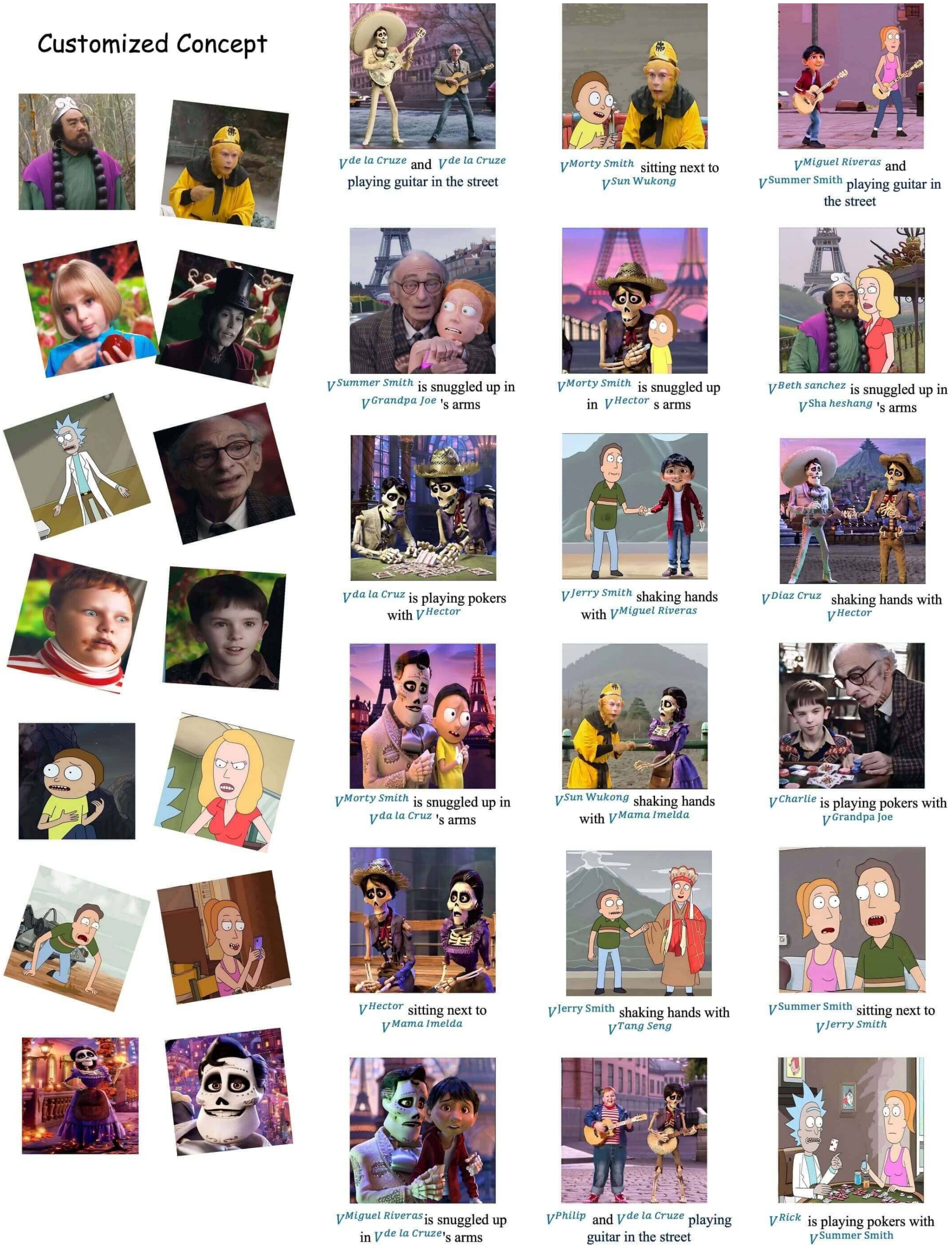

Customized Concept
$V^{\text{de la Cruze}}$ and $V^{\text{de la Cruze}}$ playing guitar in the street
$V^{\text{Morty Smith}}$ sitting next to $V^{\text{Sun Wukong}}$
$V^{\text{Miguel Riveras}}$ and $V^{\text{Summer Smith}}$ playing guitar in the street
$V^{\text{Summer Smith}}$ is snuggled up in $V^{\text{Grandpa Joe}}$'s arms
$V^{\text{Morty Smith}}$ is snuggled up in V^{Hector} s arms
$V^{\text{Beth sanchez}}$ is snuggled up in $V^{\text{Sha heshang}}$'s arms
$V^{\text{da la Cruz}}$ is playing pokers with V^{Hector}
$V^{\text{Jerry Smith}}$ shaking hands with $V^{\text{Miguel Riveras}}$
$V^{\text{Diaz Cruz}}$ shaking hands with V^{Hector}
$V^{\text{Morty Smith}}$ is snuggled up in $V^{\text{da la Cruz}}$'s arms
$V^{\text{Sun Wukong}}$ shaking hands with $V^{\text{Mama Imelda}}$
V^{Charlie} is playing pokers with $V^{\text{Grandpa Joe}}$
V^{Hector} sitting next to $V^{\text{Mama Imelda}}$
$V^{\text{Jerry Smith}}$ shaking hands with $V^{\text{Tang Seng}}$
$V^{\text{Summer Smith}}$ sitting next to $V^{\text{Jerry Smith}}$
$V^{\text{Miguel Riveras}}$is snuggled up in $V^{\text{de la Cruze}}$'s arms
V^{Philip} and $V^{\text{de la Cruze}}$ playing guitar in the street
V^{Rick} is playing pokers with $V^{\text{Summer Smith}}$

Figure 14. More results of multi-subject generation.