Compositional Inversion for Stable Diffusion Models

Xulu Zhang[1,2], **Xiao-Yong Wei**[3,1*], **Jinlin Wu**[2,5], **Tianyi Zhang**[1],
Zhaoxiang Zhang[2,4,5], **Zhen Lei**[2,4,5], **Qing Li**[1]

[1]Department of Computing, the Hong Kong Polytechnic University, Hong Kong
[2]Center for Artificial Intelligence and Robotics, HKISI, CAS, Hong Kong
[3]College of Computer Science, Sichuan University, Chengdu, China
[4]School of Artificial Intelligence, UCAS, Beijing, China
[5]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA, Beijing, China

Abstract

Inversion methods, such as Textual Inversion, generate personalized images by incorporating concepts of interest provided by user images. However, existing methods often suffer from overfitting issues, where the dominant presence of inverted concepts leads to the absence of other desired concepts. It stems from the fact that during inversion, the irrelevant semantics in the user images are also encoded, forcing the inverted concepts to occupy locations far from the core distribution in the embedding space. To address this issue, we propose a method that guides the inversion process towards the core distribution for compositional embeddings. Additionally, we introduce a spatial regularization approach to balance the attention on the concepts being composed. Our method is designed as a post-training approach and can be seamlessly integrated with other inversion methods. Experimental results demonstrate the effectiveness of our proposed approach in mitigating the overfitting problem and generating more diverse and balanced compositions of concepts in the synthesized images. The source code is available at https://github.com/zhangxulu1996/Compositional-Inversion.

Introduction

Recently, image synthesis has witnessed remarkable performance from text-to-image diffusion models such as DALL•E (Ramesh et al. 2021), Stable Diffusion (Rombach et al. 2022), Imagen (Saharia et al. 2022). These models typically consist of two modules: semantic embedding and diffusion. Given a simple text prompt like "*a cat chasing butterflies*", the semantic embedding module represents the semantics as embeddings, while the diffusion module transforms the embeddings into images that incorporate the desired concepts (e.g., *cat*, *butterflies*). However, these models produce concepts in a general sense, resulting in randomly assigned appearances for the *cat*. This limitation becomes apparent when users seek specific concepts, such as their own cat. It raises challenges to these models in the era of pursuing personalized customization.

Textual Inversion (TI) (Gal et al. 2022) remains a core technology to address this limitation. The underlying hypothesis is that an optimal point exists within the embedding

space that can represent the semantics of a specific concept, even if it is difficult to describe in words. TI evaluates the distance of current embedding to the optimal point through back-propagated gradients based on the reconstruction loss from a few sample images provided by the user. Instead of updating the model weights like in regular training, TI updates the values of the current embedding towards the optimal while keeping the weights fixed. The post-training feature enables personalization for a significantly wider range of users and researchers, as it demands fewer computational resources compared to the extensive requirement for pre-training or fine-tuning diffusion models. We hereafter use a star to denote an inverted specific concept (e.g., *cat**), commonly referred to as a pseudo word in literature.

Despite the presence of promising outcomes, the composition of inverted concepts with other concepts proves to be challenging. As shown in Fig. 1, the results maintain fidelity to the user samples for *cat**, but the concept *butterflies* is absent. This occurs because the method primarily emphasizes the reconstruction loss while disregarding the compositional aspect of the target concept in relation to others. Similar findings are reported in (Tewel et al. 2023) which suggests the dominance of the inverted concepts in the generation process encroaches upon the spotlight of other concepts. However, this is simply attributed to an over-fitting problem, with underlying rationale remaining unexplored.

This paper represents an initial endeavor to delve into the underlying reasons and offer straightforward solutions from an internal perspective. Specifically, we have discovered that Textual Inversion leads to the inverted concepts being out of distribution (OOD). Modern models are always trained on large-scale dataset like LAION (Schuhmann et al. 2022) containing text-image pairs on the billions scale. Most existing concepts have thus been trained to be compositional to others due to their frequent occurrence in the dataset. It forms a core distribution where the pretrained concepts are easily combinable. We have evaluated the compositionality of each concept by combining with others in prompts and testing the probability of their presence in resulting images using object detection. In Fig. 2, the core distribution becomes evident through the visualization of the probabilities based on their coordinates in the embedding space. This visual representation clearly showcases the OOD issue of the

Figure 1: Image synthesis using traditional inversion methods and the proposed compositional inversion: concepts of *butterflies*, *street*, and *spaceship* are absent when composed with concepts inverted with traditional methods.
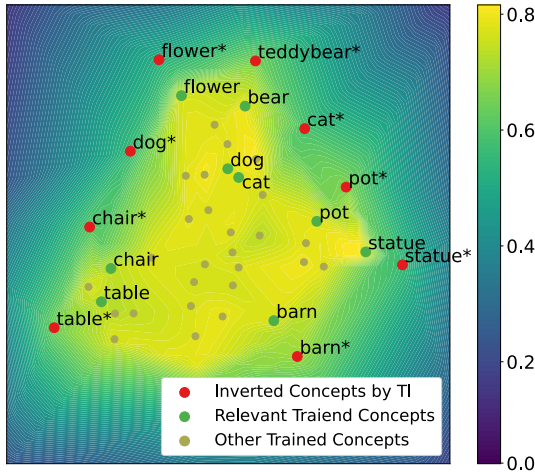

Figure 2: Visualization of compositionality in the embedding space with the evident core distribution and the OOD.

Figure 3: Development of the relative attention similarity and attention maps of various types of concepts.

inverted concepts. The OOD results from the calculation of reconstruction loss that is spanned the entire image rather than the target concept region. It makes other concepts in the background being "inverted", leading to the degradation in the purity of the semantics within the inverted concept. In Fig. 2, due to the distraction of background semantics, the inverted concept *dog** converges to an OOD area instead of the theoretically more appropriate neighborhood around the concept *dog*. This observation is further supported by the fact the average entropy of the inverted embeddings has been increased by 3% from that of the pretrained concepts.

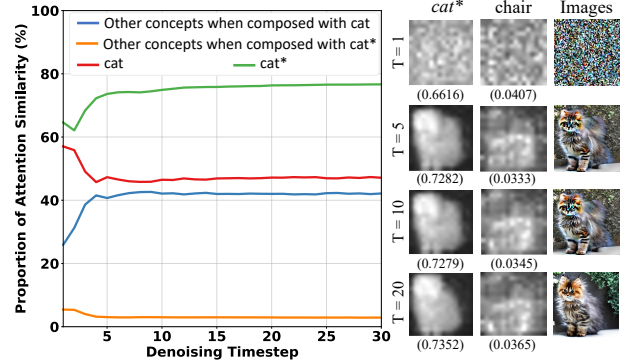The larger entropy consequently causes the dominance of the inverted concept over others in the diffusion module. The diffusion module utilizes Transformer blocks to transfer text semantics into visual content, where embeddings are employed to construct the $\mathbf{K}$, and the $\mathbf{Q}$ is typically initialized with random noise. Therefore, the presence of a concept heavily relies on the cross-attention of its embedding to the random noise. As the larger entropy of the inverted concept implies a broader span of dimensions to store semantics, it may have a higher probability of obtaining greater attention similarity compared to other concepts. In Fig. 3, we present statistics of the development of the attention similarity between $\mathbf{K}$ and $\mathbf{Q}$ over iterations. The divergence between the inverted concept and others is much more pronounced than that between a pretrained concept and others. The cross-attention mechanism iteratively integrates the attention map of the inverted concept into other concepts, ultimately re-

sulting in the absence of other concepts or their replacement with the inverted concept (e.g., Fig. 3).

Based on the aforementioned analysis, we propose a compositional inversion approach comprising two components: a **semantic inversion** component which guides the embedding search towards the core distribution, and a **spatial inversion** component which regularizes the attention maps to avoid the dominance of the inverted concepts. The framework of the proposed method is shown in Fig. 4.

Related Work

Text-to-image synthesis has earned significant attention for its potential applications in content creation, virtual reality, and computer graphics. The objective is to bridge the semantic gap and enable machines to understand and generate images that align with text prompts. For several years, generative adversarial networks (GANs) (Goodfellow et al. 2014; Karras, Laine, and Aila 2019) have been the dominant approach (Zhu et al. 2019; Tao et al. 2022). With recent improvements in DDPM (Ho, Jain, and Abbeel 2020) and DDIM (Song, Meng, and Ermon 2020), text-conditioned diffusion models have made remarkable progress. Building upon the latent images, the Latent Diffusion Model (LDM) (Rombach et al. 2022) was introduced and further extended to Stable Diffusion (Rombach et al. 2022), which is regarded as one of the most promising models for text-to-image synthesis. Another notable framework, Imagen (Saharia et al. 2022) takes a different approach by diffusing pixels directly using a pyramid structure, without relying on latent images. DALL•E2 (Ramesh et al. 2022) uses a prior network that takes text embedding as input to produce an image embedding as the input of the diffusion model.

Inversion for Customization and Personalization

As aforementioned, it is a demanding feature for the models to generate images containing specific concepts of interest (CoI) implied by user samples. This requires models' capacity to "invert" the samples into concept embeddings, which can be used in future prompts for customized generations. Textual Inversion (Gal et al. 2022) is one of the initial methods that directly searches for the optimal solution in the embedding space to address this issue. However, the remaining methods, although employing similar approaches of searching for inverted embeddings, rely on either retraining or fine-tuning for this purpose. For instance, DreamBooth (Ruiz et al. 2023) retrains the entire Imagen for constructing embeddings for CoI, while Custom Diffusion (Kumari et al. 2023), Perfusion (Tewel et al. 2023), SVDiff (Han et al. 2023), and Cones (Liu et al. 2023) only fine-tune partial parameters of the Stable Diffusion model. To mitigate language drift and overfitting problems, a large number of images from the same CoI class are typically utilized as regularization during the training/fine-tuning process.

Compositionality of Inverted Concepts

Current methods in the field of compositionality primarily focus on combining inverted concepts with each other rather than with a broader range of pretrained concepts. This approach, known as multi-concept composition, is related to but distinct from the scope of this paper. Existing methods include Custom Diffusion, SVDiff, Cones, and Perfusion. Custom Diffusion archives this by merging the outputs of multiple models that have been fine-tuned to invert various CoI. It can be considered as a model-level composition approach. SVDiff manually combines objects selected from different CoI concepts as training images, enabling the model to learn to compose them. Cones evaluates the neurons' contributions to the fidelity of inverted concepts and deactivates those with minor contributions during composition. Perfusion fuses the **V** components of inverted concepts to balance their contribution to generation. These methods all rely on training/fine-tuning, which requires effort and expertise to gather the regularization images. In contrast, the compositional inversion proposed in this paper is a post-training approach that can be applied to any trained or fine-tuned models and thus is compatible to all the aforementioned methods. Furthermore, the proposed method can be employed for the composition of both pretrained and inverted concepts, making this paper a study of compositionality in a more general sense.

Spatial Guidance in Text-to-Image Synthesis

In terms of imposing spatial constraints, there is another category of methods specifically designed for controlling the contours, shapes, or layouts of objects. ControlNet (Zhang, Rao, and Agrawala 2023) trains a new branch that incorporates spatial constraints as input and injects them into each layer of the diffusion module for customized synthesis. Prompt-to-prompt (Hertz et al. 2022) enables object-specific editing by replacing the attention map in the cross-attention module. GLIGEN (Li et al. 2023) designs a gated self-attention layer to incorporate spatial conditions, such as bounding boxes. Layout-control (Chen, Laina, and Vedaldi 2023) employs a training-free approach that ensures higher activation values of the attention maps within the bounding box regions. ReCo (Yang et al. 2023) achieves layout control by encoding regional tokens as part of the text prompt. The spatial inversion module in our proposed method draws inspiration from these methods in terms of controlling the layout. However, these methods are not developed for inversion purpose but rather assume the constrains as a prior, while our focus is on automatically discovering the underlying spatial distribution without any user specifications.

Method

Preliminaries

By taking an encoder-decoder view that is similar to that of variational autoencoders (VAEs) (Kingma and Welling 2014), it is straightforward to inspect diffusion models. The encoding is more commonly called a forward process that iteratively "diffuses" a sequence of Gaussian noises $(\epsilon_t)_{t=1}^{T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into an image $\mathbf{x}_0$ using a Markov chain of T steps, producing a sequence of noisy samples $(\mathbf{x}_t)_{t=1}^{T}$ with

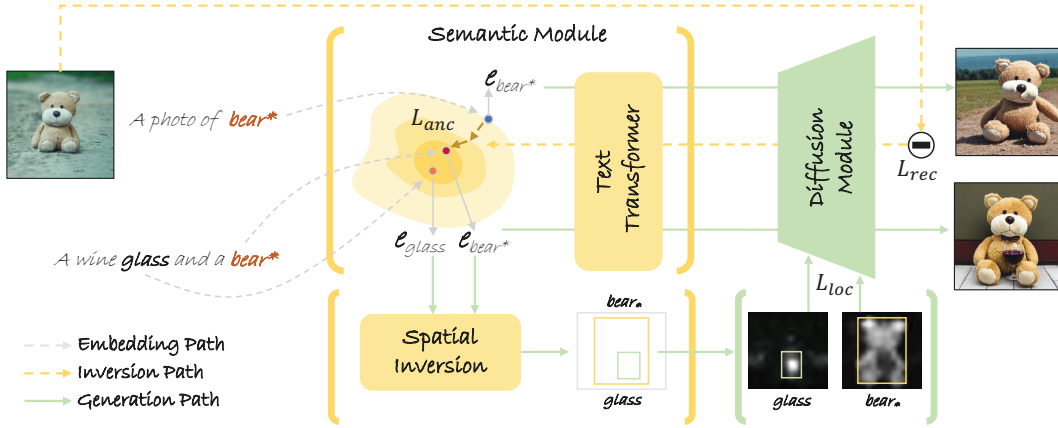$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \ 1 \le t \le T, \qquad (1)$$

Figure 4: The framework of the proposed method consisting of semantic and spatial inversion components.

where α_t controls the variance of the Gaussian noises ϵ_t. It also defines a Gaussian distribution $q(\mathbf{x}_t|\mathbf{x}_0)$ that we can use to sample latent representations for $\mathbf{x}_t$ in the generation.

The decoding is more commonly referred to as a reverse diffusion process, in which the goal is to learn another Gaussian distribution $q(\mathbf{x}_0|\mathbf{x}_t)$ so that we can reconstruct $\mathbf{x}_0$ from $\mathbf{x}_t$. Since the Markov encoding is non-reversible, the reverse diffusion is implemented by approximating $q(\mathbf{x}_0|\mathbf{x}_t)$ using a model f (e.g., a neural network) which is parameterized on θ and learns an estimated distribution p_θ. This can be formulated as a T-step "denoising" process where, at the tth step, it tries to reconstruct $\mathbf{x}_0$ by removing noise from $\mathbf{x}_t$ and results in an estimation

$$\hat{\mathbf{x}}_0 = f(\mathbf{x}_t; \theta) \sim p_\theta(\mathbf{x}_0|\mathbf{x}_t). \qquad (2)$$

The learning of the model can be done based on the loss of the estimation $\hat{\mathbf{x}}_0$ from $\mathbf{x}_0$. To implement text-to-image synthesis, text embedding $\mathbf{e}$ will also be fused with $\mathbf{x}_t$ to generate a conditioned image using Eq. (2) as $\hat{\mathbf{x}}_0 = f(\mathbf{x}_t \circ \mathbf{e}; \theta)$ where $\circ$ is a reserved fusion operator which is implemented differently in various models. The loss is then written

$$
\begin{aligned}
L_{rec} &= \mathbb{E}\left[w_t\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_2^2\right], \\
&= \mathbb{E}\left[w_t\|f(\mathbf{x}_t \circ \mathbf{e}; \theta) - \mathbf{x}_0\|_2^2\right], \\
&= \mathbb{E}\left[w_t\|f((\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_0) \circ \mathbf{e}; \theta) - \mathbf{x}_0\|_2^2\right],
\end{aligned}
\qquad (3)
$$

where w_t is a time-step dependent weight.

As the generated content is indeed controlled by the only input $\mathbf{e}$, we can generate desired content as long as we know its text embeddings. This is easy for pretrained concepts (e.g., *cat*) because the learners have seen enough samples during the training, but hard for specific concepts (e.g., my own cat). To address this issue, Textual Inversion is a method to backtrack the text embeddings of specific concepts (Gal et al. 2022). It feeds a few samples of the target concept (e.g., 3–5 images of the user's cat) and updates the pseudo-concept embedding ($\mathbf{e}_{cat*}$). It is formulated as

$$\mathbf{e}_* = \arg\min_{\mathbf{e}} L_{rec}. \qquad (4)$$

Semantic Inversion

As visualized in Fig. 2, Textual Inversion will make the new (pseudo-)embeddings OOD and incompatible to other concepts in the embedding space, because it does not have enough interactions with others during the post-training learning. Our idea is then straightforward to improve its interactions for better compositionality.

To this end, we select a set of general concepts as anchors (e.g., *dog*, *car*, *chair*, *building*) and collect their text embeddings $\{\mathbf{e}_{anc}\}$. These concepts can be found from existing benchmark dataset like COCO (Lin et al. 2014) and even be combined for a wider coverage of semantic references (Deng et al. 2009; Wei and Yang 2012, 2011). We will use the anchor concepts as attractors to guide the search of the pseudo-embedding towards the core distribution. A loss regularization is introduced as

$$L_{anc} = \mathbb{E}\left[\frac{1}{\|\{\mathbf{e}_{anc}\}\|} \sum_{\mathbf{e}_i \in \{\mathbf{e}_{anc}\}} \delta_i \|\mathbf{e} - \mathbf{e}_i\|_2^2\right], \text{ r.s.t } \|\delta\|_0 < c$$

$$(5)$$

where $\delta = \{\delta_i\}$ is a weighting vector to control the strength of the ith attractor, and the constraint $\|\delta\|_0 < c$ limits the number of active attractors to c. This is to avoid the distraction from irrelevant attractors. For example, when searching for a pseudo-embedding for a cat related concept, active attractors like *cat*, *pet* are preferred over irrelevant ones like *car*, *airplane*. We implement the weighting vector δ using sparse coding (Olshausen and Field 1996).

Once a pseudo-word $S*$ has been assigned with the embedding $\mathbf{e}_*$, it can be used in the same way as a real word in a prompt for image generation (e.g., "*a S* cat sitting next to a dog*"). Each word in a prompt is more commonly referred to as a token. The model first assigns an embedding for each token and feeds these embeddings into a text Transformer where they are refined into the actual token embeddings that will be used as conditions in the generation (or reverse) process. To simplify the description, we will still use the symbol $\mathbf{e}$'s to represent these refined token embeddings.

Methods	Comp. w/ Pretrain Concepts			Comp. w/ Inverted Concepts	
	Text-Align.	CoI Likelihood	Image-Align.	Text-Align.	Image-Align.
Textual Inversion (Gal et al. 2022)	0.603	0.032	**0.784**	0.606	0.656
+ Semantic Inversion	0.645	0.121	0.762	0.633	**0.664**
+ Spatial Inversion	0.631	0.116	0.749	0.620	0.645
+ Semantic + Spatial	**0.702**	**0.284**	0.732	**0.662**	0.658
Custom Diffusion (Kumari et al. 2023)	0.695	0.226	**0.802**	0.702	**0.700**
+ Semantic Inversion	0.701	0.352	0.760	**0.706**	0.681
+ Spatial Inversion	**0.738**	0.425	0.727	0.689	0.652
+ Semantic + Spatial	0.734	**0.459**	0.683	0.703	0.628
DreamBooth (Ruiz et al. 2023)	0.716	0.431	**0.734**	0.691	**0.695**
+ Semantic Inversion	0.720	0.436	0.718	0.704	0.683
+ Spatial Inversion	0.750	**0.534**	0.657	0.705	0.632
+ Semantic + Spatial	**0.753**	0.529	0.646	**0.710**	0.616

Table 1: Evaluation of performance by composing with pretrained and inverted concepts, with ablation of semantic and spatial inversion components. The best results are in bold font.

Spatial Inversion

To make the image generation conditioned on the token embeddings in the reverse process, a popular way is to use transformer blocks. More specifically, an attention map will be calculated for each token embedding to indicate its appearance or how it is attended in the resulting image (e.g., location, shape, details). This is implemented by the cross-attention mechanism as

$$\mathbf{A}_i = softmax\left(\frac{\phi(\mathbf{x}_t)\kappa(\mathbf{e}_i)^\top}{\sqrt{d_k}}\right), \qquad (6)$$

where ϕ and κ are the image feature extractor and text feature extractor respectively and d_k is the dimensionality of $\kappa(\mathbf{e}_i)$. Several pioneer works have found that the appearance of the token can be controlled by manipulating this attention map (Hertz et al. 2022; Parmar et al. 2023; Chen, Laina, and Vedaldi 2023). Therefore, we can regulate the attention maps to be attended on the right tokens to avoid the situation that the pseudo tokens dominate the generation process. In spatial inversion, we propose a method to recover the coherent locations of a pseudo token (e.g., S*) and concepts being combined (e.g., *dog*) in a prompt. The locations are then used to regulate the attention maps of tokens.

We implement the location recovery by training an MLP model which takes two token embeddings as the input and outputs the locations as

$$\mathbf{l}_i, \mathbf{l}_j = MLP(\mathbf{e}_i, \mathbf{e}_j), \qquad (7)$$

where $\mathbf{l}_i, \mathbf{l}_j \in \mathbb{R}^4$ are coordinates of the bounding boxes of the attended areas of the two tokens. We simplify the method by only considering noun tokens and construct a vocabulary of frequently used nouns (Lin et al. 2014). The nouns are then combined as prompts and used to generate images. The object detection (Carion et al. 2020) is conducted on the resulting images to find the bounding boxes of these nouns which serve as the ground truth for the training. For tokens in the vocabulary, we assign it to the noun that is the nearest one in the embedding space. The MLP is then able to recover the locations of any given tokens.

With the location bounding boxes, we convert each of them into an attention mask $\mathbf{M}_i$ which is with the same size as that of $\mathbf{A}_i$. We can then manipulate the attention maps by introducing a location regularization loss as

$$L_{loc} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\sum(\mathbf{M}_i \circ \mathbf{A}_i)}{\sum \mathbf{A}_i}\right). \qquad (8)$$

It encourages the tokens to be attended on the locations indicated by the masks and penalizes deviations. It is calculated at the first 10 reverse steps to update the latent variable $\mathbf{x}_t$.

Experiments

To evaluate the performance of our proposed methods, we conduct experiments by combining the inverted concepts with both pretrained and inverted concepts.

Datasets. We construct a comprehensive dataset by accumulating almost all open-sourced concepts used in previous studies (Kumari et al. 2023; Gal et al. 2022; Ruiz et al. 2023). It consists of 10 concepts of 2 animal, 2 furniture, 2 object/container, 1 house, 1 plant, 2 toy categories. To test the generalizability, we generate prompts by combining the inverted concepts with 80 categories from the COCO dataset (Lin et al. 2014) using the conjunction word "and". This results in 1600 prompts and generates 16000 images. We also combine the inverted concepts to each other, resulting in 90 prompts and 900 images generated. This is also aligned to the multi-concept composition task in previous studies.

Evaluation Metrics. We utilized three evaluation metrics: 1) *Text-alignment* which quantifies the extent to which a generated image accurately represents the semantics of the text prompt, as determined by the CLIP similarity (Radford et al. 2021). 2) *CoI Likelihood* which measures the probability that CoIs present in the results using an object detector (DETR (Carion et al. 2020) based on ResNet101 (He et al. 2016) and pretrained on the COCO dataset). 3) *Image-alignment* which evaluates the extent to which the generated images are visually similar to the user samples, as determined by the cosine similarity of their CLIP image features.

Figure 5: Examples of composing inverted concepts *cat** and *dog** with pretrained concepts *backpack* and *book*.

Baselines. We employ 3 poplar state-of-the-art (SOTA) methods as the baselines including 1) *Textual Inversion (TI)* (Gal et al. 2022) which focuses on fine-tuning the text embedding exclusively. We employ the Stable Diffusion version, using the parameters reported by the authors in their paper. 2) *DreamBooth* (Ruiz et al. 2023) which fine-tunes all parameters of the U-Net architecture. As DreamBooth does not fine-tune the text embedding, we integrate TI into DreamBooth to apply the proposed method in this paper. 3) *Custom Diffusion* (Kumari et al. 2023) which aims to fine-tune partial parameters in the cross-attention modules. For the composition of inverted concepts, we adopt the joint training strategy, as it has been highlighted in the paper as the best-performing approach. The third-party implementations from HuggingFace are used for all the aforementioned methods. In the fine-tuning and inference stages, we follow the usual practice to use the *S** and superclass token to represent the inverted concept (e.g., "*cat* cat*").

Performance

The results are presented in Table 1. The proposed method exhibits improvements over SOTA methods in terms of 16.4% (787.5%), 5.6% (103.1%), 5.2% (22.7%) on Text-Align (CoI Likelihood) compared to TI, Custom Diffusion, and DreamBooth, respectively. There is only a slight trade-off of 6.6%, 14.8%, and 12.0% on Image-Align when compared to the three methods. The performance gain on CoI Likelihood reaches 52.9% when composing with pretrained

concepts, indicating a significant improvement. Another observation is that the augmented TI achieves a comparable performance to the original Custom Diffusion and Dream-Booth. This is surprising because SOTA performance is achieved without any fine-tuning of network parameters.

Fig. 5 shows two examples of composing inverted concepts with pretrained concepts. The proposed method clearly improves the performance in the presence of the pretrained concepts. Note that the semantic inversion module primarily emphasizes semantic completeness, occasionally resulting in the generation of low-probability scenes (such as half a cat in a backpack or a dog reading a book). On the other hand, the spatial inversion module tends to generate scenes that align with more common statistical occurrences.

Fig. 6 presents two examples of composing the inverted concepts to each other. The presence of the CoIs is also significantly increased. A noticeable difference compared to the results in Fig. 5 is the larger variation in the appearance of the concepts of interest. Specifically, the generated cats, dogs, and barns exhibit a wider range of viewpoints.

User Study

To assess the computational efficiency and quality of the synthesis, we conducted a user study. We randomly selected 1,600 images generated by the proposed method and enlisted the participation of two users to rate the synthesis quality. The ratings were divided into three categories: *Excellent* represents the successful generation of two CoIs without any
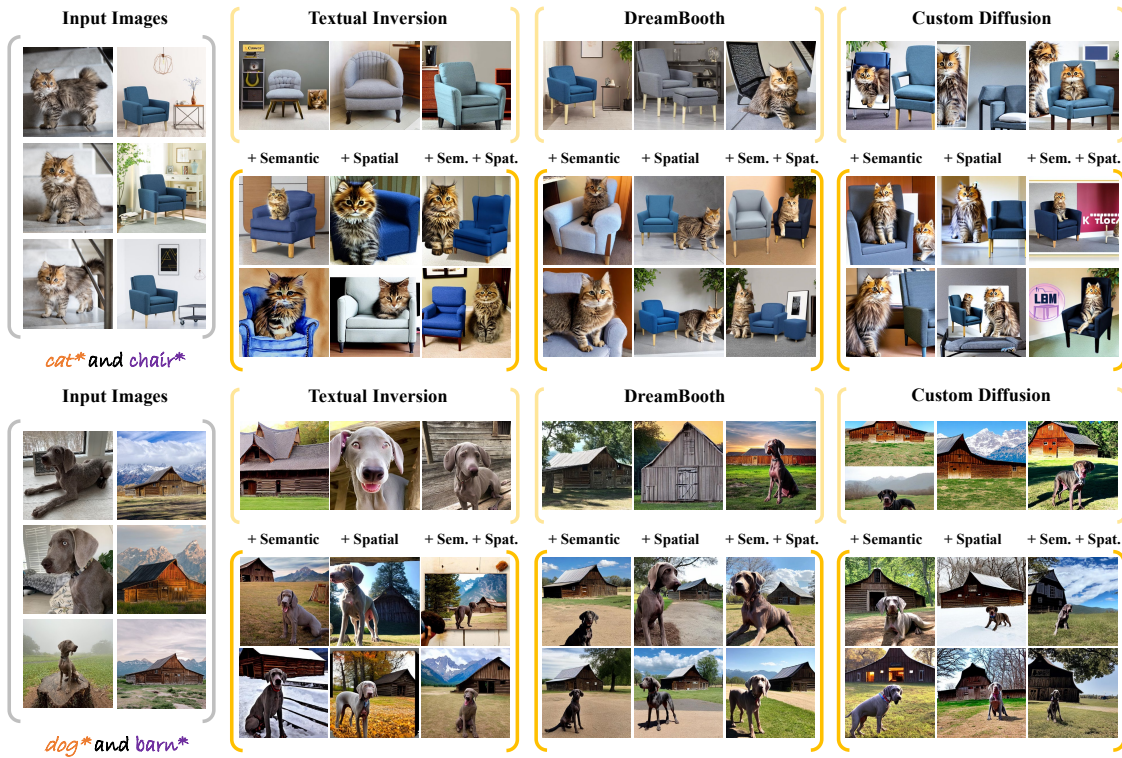
Figure 6: Examples of composing inverted concepts of *cat**, *chair**, *dog**, and *barn** to each other.
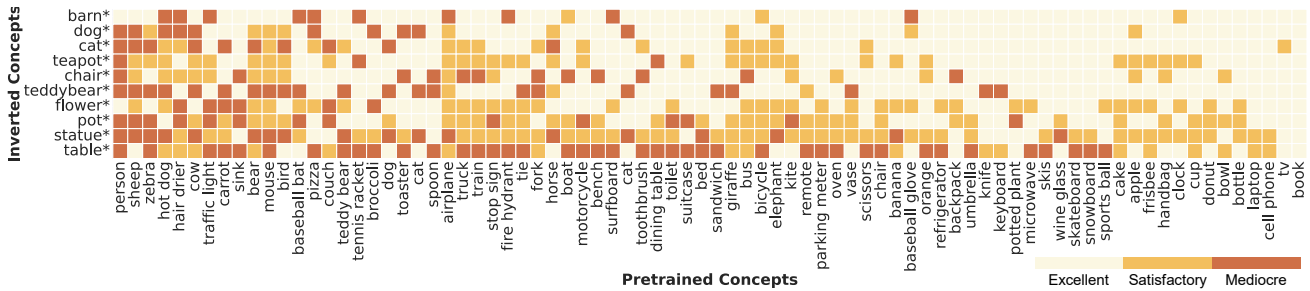

Figure 7: Assessment of compositional synthesis quality through user evaluations.

unnatural details; *Satisfactory* indicates that the CoIs were generated in an acceptable manner, though some minor flaws may be present; and *Mediocre* signifies the presence of obvious unreasonable details or missing CoIs. The results of the user study are presented in Fig. 7. The high capability of the proposed method in generating quality images is clearly evident, as indicated by a probability of 81.9% for receiving ratings above the *Satisfactory*.

Additionally, we assume that the quality rating serves as an indicator of compositionality. When the probability of generating high-quality images through the composition of two concepts is higher, it suggests that those concepts are easier to compose. In Fig. 7, it becomes apparent that rigid objects (e.g. $book$, tv) are more straightforward to compose. This observation is supported by the fact that 9 out of the 10 rightmost concepts in the figure are rigid objects. This finding aligns with our understanding as rigid objects pos-

sess more consistent appearances and visual characteristics. In contrast, non-rigid objects like animals (e.g., cow, $sheep$) are challenging to compose, as indicated by the fact that 7 out of the 10 leftmost concepts are non-rigid objects.

Conclusion

We have identified the mechanism that causes the overfitting and dominance of the inverted concepts in generation. To address the issue, we propose a compositional inversion method which consists of two modules of semantic and spatial inversions. The semantic inversion guides the inversion towards the core distribution to ensure better coherence with other concepts, while the spatial inversion discovers the underlying layout distribution for CoIs and uses it to regularize the attention maps. The experimental results have validated the effectiveness of the method.

Acknowledgments

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Chen, M.; Laina, I.; and Vedaldi, A. 2023. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2022. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

Liu, Z.; Feng, R.; Zhu, K.; Zhang, Y.; Zheng, K.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*.

Olshausen, B. A.; and Field, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583): 607–609.

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

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

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 16515–16525.

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

Wei, X.-Y.; and Yang, Z.-Q. 2011. Coached active learning for interactive video search. In *Proceedings of the 19th ACM international conference on Multimedia*, 443–452.

Wei, X.-Y.; and Yang, Z.-Q. 2012. Coaching the exploration and exploitation in active learning for interactive video retrieval. *IEEE Transactions on Image Processing*, 22(3): 955–968.

Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14246–14255.

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

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5810.