VISUALLY GUIDED DECODING: GRADIENT-FREE HARD PROMPT INVERSION WITH LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image generative models like DALL-E and Stable Diffusion have revolutionized visual content creation across various applications, including advertising, personalized media, and design prototyping. However, crafting effective textual prompts to guide these models remains challenging, often requiring extensive trial and error. Existing prompt inversion methods, such as soft and hard prompt techniques, suffer from issues like limited interpretability and incoherent prompt generation. To address these limitations, we introduce Visually Guided Decoding (VGD), a gradient-free approach that leverages large language models (LLMs) and CLIP-based guidance to generate coherent and semantically aligned prompts. VGD utilizes the robust text generation capabilities of LLMs to produce human-readable prompts while employing CLIP scores to ensure alignment with user-specified visual concepts. This method enhances the interpretability, generalization, and flexibility of prompt generation without the need for additional training. Our experiments demonstrate that VGD outperforms existing prompt inversion techniques in generating understandable and contextually relevant prompts, facilitating more intuitive and controllable interactions with text-to-image models.

1 INTRODUCTION

In recent years, image generative models such as DALL-E and Stable Diffusion have shown remarkable success in generating high-fidelity images (Ramesh et al., 2022; Rombach et al., 2022; Podell et al., 2024). These models have been widely used in a variety of applications, including visual content generation (*e.g.*, advertisement, movie, game), personalized content generation (*e.g.*, caricature, photo editing), and also prototyping (*e.g.*, architecture and product design). Previous studies have shown that, just as humans can draw an image of an object solely based on detailed descriptions (*e.g.*, criminal composite sketch), generative models can generate images of objects using a well-crafted prompt, even if they have not been trained on those specific objects (Gal et al., 2023; Everaert et al., 2023).

A well-known drawback of these approaches is the difficulty in finding a textual description (*a.k.a.,* prompt) that effectively guides the generation of the desired visual content (Hao et al., 2024). For example, to create a culinary masterpiece (see Figure 1), one should have good knowledge of exquisite cooking styles such as French cuisine or molecular gastronomy. Without the expertise to craft precise prompts, users have to rely on laborious trial and error to generate the desired images. Automation of the prompt generation process will reduce the time and effort required to generate the desired image.

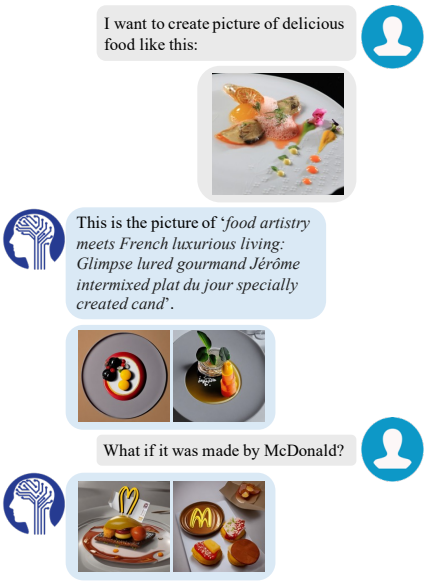


Figure 1: Visually Guided Decoding (**VGD**) works with any LLM without extra training, making it easy to integrate into a chat-based interface that offers interpretable and controllable text-to-image generation.

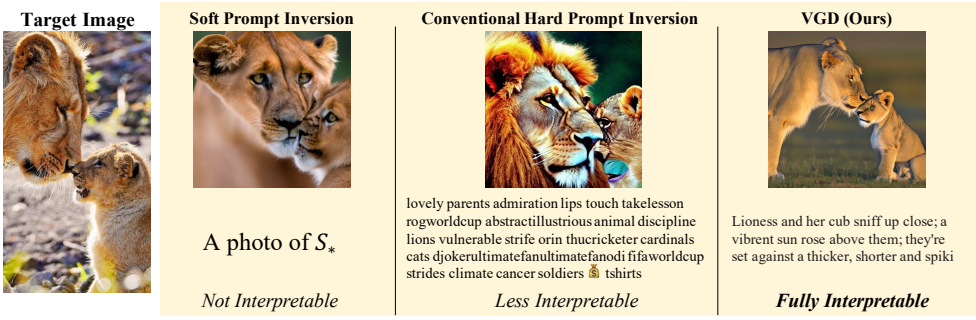
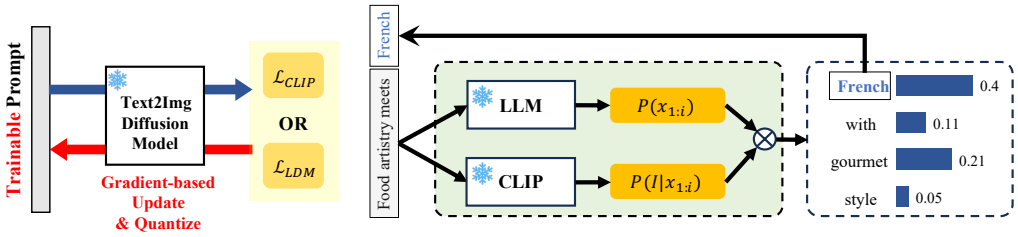


Figure 2: VGD generates fully interpretable prompts that enhance generalizability across tasks and models in text-to-image generation.



(a) Training-based Prompt Inversion (b) Gradient-free, Visually Guided Hard Prompt Inversion (**Ours**)

Figure 3: While conventional prompt inversion techniques update prompt embeddings through gradient-based optimization and quantization, VGD is a gradient-free technique that utilizes large language models and CLIP to generate relevant sentences.

To address the difficulty in finding the prompt, soft prompt inversion techniques have been proposed (Gal et al., 2023; Kumari et al., 2023; Ruiz et al., 2023; Voynov et al., 2023). In these approaches, following the prompt learning works in natural language processing (Li & Liang, 2021; Lester et al., 2021; Liu et al., 2022), an image is transformed into the embedding vector. While this so-called soft prompt contains useful information about the image, since it is a string of numbers, it is not human-readable and hence users cannot handle it easily (see Figure 2).

Recently, to mitigate the lack of interpretability and generalization capability, hard prompt inversion has been proposed (Mahajan et al., 2024; Wen et al., 2024). In this approach, soft prompts (embedding vectors) are projected to the nearest neighbors in a predefined vocabulary (see Figure 3 (a)). Specifically, the learnable input vectors for the conditioning network (Radford et al., 2021) in Stable Diffusion are trained to reconstruct user-provided images through an iterative de-noising process (*a.k.a.*, reverse diffusion process) and then mapped to the nearest word embedding in the model's vocabulary. During this training process, the long gradient path in the reverse diffusion process to the input prompt often results in a vanishing gradient problem. Further, since the hard prompt is a simple collection of words (see Figure 2), it is difficult to identify which words are essential in generating the desired images.

An aim of this paper is to propose a technique that generates a fully interpretable prompt. Essence of our technique, henceforth referred to as visually guided decoding (VGD), is to generate contextually meaningful sentence for hard prompt through the token generation process of large language models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023). To this end, during the token generation process, we employ CLIP to align the generated token with the provided images (see Figure 3 (b)). The benefit of the proposed approach is that one can fully understand the meaning of the prompt (*interpretability*) and synthesize image for various applications (*generalizability*) (*e.g.*, style transfer, image editing, and image variation) with greater control over the output image (*flexibility*). Since VGD seamlessly stitches LLMs and text-to-image models without the need for training (*i.e.*, gradient-free), VGD provides a flexible and efficient solution for chat-interfaced image generation services using LLM (*e.g.*, DALL-E extension on ChatGPT) (see Figure 1).

In our experiments, VGD qualitatively and quantitatively achieves state-of-the-art (SOTA) performance, demonstrating superior interpretability, generalizability, and flexibility in text-to-image generation compared to previous soft and hard prompt inversion methods. We also show that VGD is compatible with various LLMs including LLaMA2, LLaMA3 and Mistral.

2 BACKGROUND AND RELATED WORKS

2.1 CLIP MODEL

CLIP model aligns semantically related visual and textual content within a shared representation space (Radford et al., 2021). It comprises an image encoder $\text{CLIP}_{\text{img}}(\cdot)$ and a text encoder $\text{CLIP}_{\text{txt}}(\cdot)$. The image encoder encodes an input image I into a visual embedding $\mathbf{f}_{\text{img}} = \text{CLIP}_{\text{img}}(I)$. Similarly, The text encoder processes the input text $T = [x_1^{\text{txt}}, x_2^{\text{txt}}, ..., x_N^{\text{txt}}]$, yielding a textual embedding $\mathbf{f}_{\text{txt}} = \text{CLIP}_{\text{txt}}(T)$. CLIP is trained using a contrastive learning approach that maximizes the similarity between $\mathbf{f}_{\text{txt}}$ and $\mathbf{f}_{\text{img}}$ for matched sentence-image pair, while minimizing the similarity for unmatched pairs. The similarity is defined as:

$$\text{Sim}(\mathbf{f}_{\text{txt}}, \mathbf{f}_{\text{img}}) \coloneqq s \cdot \frac{\mathbf{f}_{\text{txt}} \cdot \mathbf{f}_{\text{img}}}{\|\mathbf{f}_{\text{txt}}\|_2 \|\mathbf{f}_{\text{img}}\|_2} \tag{1}$$

where s is a scalar scaling factor. Then, the training objective of CLIP is formulated as:

$$\underset{\text{CLIP}_{\text{img}}, \text{CLIP}_{\text{txt}}}{\textbf{Maximize}} \quad \log P\left(T|I\right) + \log P\left(I|T\right), \tag{2}$$

$$P(T|I) = \frac{\exp\left(\text{Sim}(\mathbf{f}_{\text{txt}} \cdot \mathbf{f}_{\text{img}})\right)}{\sum_{j=1}^{B} \exp\left(\text{Sim}(\mathbf{f}_{\text{txt}_j} \cdot \mathbf{f}_{\text{img}})\right)}, \quad P(I|T) = \frac{\exp\left(\text{Sim}(\mathbf{f}_{\text{txt}} \cdot \mathbf{f}_{\text{img}})\right)}{\sum_{j=1}^{B} \exp\left(\text{Sim}(\mathbf{f}_{\text{txt}} \cdot \mathbf{f}_{\text{img}_j})\right)}, \tag{3}$$

where B indicates the number of text-image pairs in the mini-batch.

2.2 TEXT-TO-IMAGE DIFFUSION MODEL

Text-to-image models generate an image I that maximizes $P(I|T)$ given textual description T. Modern text-to-image models, such as Stable Diffusion, are based on Latent Diffusion Model (LDM), which usually takes CLIP text embedding $\mathbf{f}_{\text{txt}}$ as a condition to generate an image. The output of LDM, denoted as $\epsilon_\theta(\mathbf{z}_k, \mathbf{f}_{\text{txt}}, k)$, represents the predicted denoising result after k diffusion steps, starting from a random Gaussian noise $\mathbf{z}_k$ conditioned on $\mathbf{f}_{\text{txt}}$. The training objective of LDM is as follows:

$$L_{\text{LDM}} = \mathbb{E}_{\epsilon, \mathbf{z}, \mathbf{f}_{\text{txt}}, k}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_k, \mathbf{f}_{\text{txt}}, k)\|_2^2\right], \tag{4}$$

where $\epsilon \sim N(0, \mathbf{I})$ indicates the noise used to corrupt clean latent variables. Our goal is to find the optimal text condition T that yields the image containing desired visual concepts.

2.3 PROMPT INVERSION

Soft Prompt Inversion Following prompt-tuning approach Li & Liang (2021), Soft Prompt Inversion techniques (Gal et al., 2023; Kumari et al., 2023) extend the vocabulary of the model with a new special token S_*, which embeds the objects from the provided images. During the generation process, the continuous embedding vector of S_* is prepended to the embeddings of the input text tokens (*e.g.*, "A photo of S_*", "A rendition of S_*") and then used as an input to the LDM model. While effective in generating specific visual concepts, the soft prompt is in continuous vector form which is not human-readable and difficult for users to modify to generate desired images.

Hard Prompt Inversion To mitigate the limitations of Soft Prompt Inversion, PEZ (Wen et al., 2024) and PH2P (Mahajan et al., 2024) have recently introduced hard prompt (discrete prompt) inversion approaches. On top of the soft prompt inversion methods, hard prompt inversion methods leverage projected gradient descent as a quantization technique to produce the learned soft prompt within the word embedding space. These hard prompts, once generated, often appear as a disjointed collection of words (commonly referred to as the incoherence problem), making it difficult to determine which words need to be modified to generate the desired images. Moreover, hard prompt inversion involves a complex training process due to discrete optimization and multi-stage pipelining. VGD does not need for gradient-based optimization, reducing computational requirements and streamlining the process.

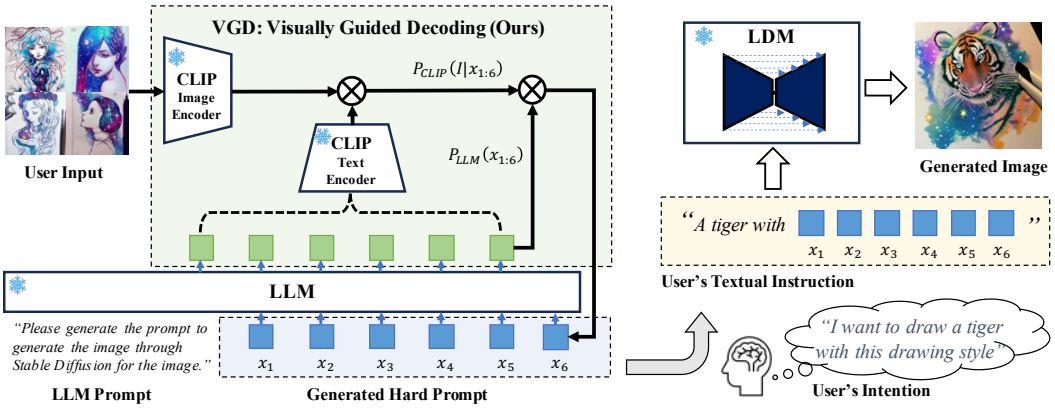


Figure 4: Overview of the proposed hard prompt inversion method, VGD, which seamlessly integrates LLM, CLIP, and LDM for user-friendly image generation process.

Image Captioning One might think that generating an image directly from captions produced by large multi-modal models (LMMs) (Li et al., 2022; 2023; Alayrac et al., 2022; Liu et al., 2024) would be an alternative option to the hard prompt inversion. However, LMM-generated captions often lack the fine details necessary for detailed image synthesis control. To complement the missing information, prompt generation services like CLIP-Interrogator [1] have been introduced to complement image captioning models such as LLaVA (see Appendix A.4 for more details).

3 Gradient Free Prompt Inversion with Language Models

3.1 Interpretability Degradation of Previous Hard Prompt Inversion

The goal of hard prompt inversion is to determine the text prompt $T = [x_1^{\text{txt}}, x_2^{\text{txt}}, ..., x_N^{\text{txt}}]$ that maximizes the LDM's probability $P(I|T)$ for a target image I (where typically $N \leqslant 77$), with each x_i^{txt} selected from the model's vocabulary. By applying Bayes' theorem, this objective can be reformulated as:

$$P(I|T) = \frac{P(I)P(T|I)}{P(T)}. \tag{5}$$

Since the prior probability of the image $P(I)$ is independent of T, finding the optimal $\hat{T}$ is equivalent to finding T that maximizes $P(T|I)/P(T)$, expressed as:

$$\hat{T} = \underset{T}{\text{argmax}} \frac{P(T|I)}{P(T)}. \tag{6}$$

As indicated in Eq. 6, the inversion process maximizes $P(T|I)$ (*a.k.a.,* image captioning objective) while simultaneously minimizing the prior probability of text, $P(T)$. Minimizing $P(T)$ can cause the hard prompt T to contain uncommon or awkward phrasing, leading to reduced interpretability. We suggest that this is the primary reason why existing hard prompt inversion techniques exhibit lower interpretability compared to our approach (see Section 4.3 for qualitative comparison).

3.2 Visually Guided Decoding

Step 1 - Problem Formulation Our goal is to find an *interpretable* text prompt without training prompt embeddings, using the *gradient-free* approach. Inspired by the noisy channel model (Jurafsky, 2000; Brown et al., 1993), widely used in machine translation and speech recognition, we model the prompt discovery process by integrating an LDM objective with a regularization term for language modeling. The objective is formulated as:

$$\hat{T} = \underset{T}{\text{argmax}}\ P(I|T)P(T)^{\alpha}, \tag{7}$$

where α is a hyperparameter that balances the influence of $P(I|T)$ and $P(T)$.

[1] https://github.com/pharmapsychotic/clip-interrogator

Step 2 - Approximation with CLIP Score Computing $P(I|T)$ by forward and reverse diffusion processes at each token generation step is computationally intractable. To alleviate this computational burden, we approximate $P(I|T)$ using CLIP, such that $P(I|T) \approx P_{\text{CLIP}}(I|T)$. We assert that this approximation is justified by the fact that Stable Diffusion utilizes the frozen CLIP text encoder. We empirically show that when a CLIP model different from the one aligned with Stable Diffusion is used, the approximation no longer holds, leading to a decline in performance (see Section 4.4 for the ablation study on CLIP image encoders).

In addition, we leverage an external language model such as LLaMA (Touvron et al., 2023) to model the $P(T) \approx P_{\text{LLM}}(T)$. This leads to our main objective:

$$\hat{T} = \underset{T}{\text{argmax}}\ P_{\text{CLIP}}(I|T) P_{\text{LLM}}(T)^{\alpha}. \tag{8}$$

Step 3 - Token-by-Token Generation We can further express Eq. 8 using a left-to-right decomposition of the text probability $P_{\text{LLM}}(T)$:

$$P_{\text{CLIP}}\left(I|x_{1:N}^{\text{txt}}\right) \prod_{i=1}^{N} P_{\text{LLM}}\left(x_i^{\text{txt}}|x_{1:i-1}^{\text{txt}}\right)^{\alpha}. \tag{9}$$

This decomposition facilitates token-by-token text generation using a beam search decoding strategy (Anderson et al., 2017; Post & Vilar, 2018; Hu et al., 2019; Holtzman et al., 2020). Specifically, we iteratively select the i-th token x_i^{txt} that maximizes the objective, Eq. 9, given previous tokens $x_{1:i-1}^{\text{txt}}$. In doing so, VGD generates prompts that are semantically aligned with the target image (*via* CLIP), linguistically coherent, and interpretable (*via* LLM).

3.3 Implementation of Visually Guided Decoding

LLM Prompting Instead of fine-tuning the LLM for each image generation task, we design specific system and user prompts tailored to different tasks. The LLM is then queried to generate tokens as if it were creating text prompts for text-to-image models (see Appendix A.1 for details).

Beam Initialization Providing the LLM with image-related words at the start of the generation process increases the probability of generating tokens that align with the image in subsequent steps. To achieve this, we first compare all text embeddings $\mathbf{f}_{\text{txt}}$ in the CLIP vocabulary with the target image embedding $\mathbf{f}_{\text{img}}$, selecting the top-M tokens $x_{1:M}^{\text{txt}}$ as the initial input prefix for the LLM. Empirically, we find that even $M = 1$ is sufficient for VGD. This process is computationally efficient since all text embeddings can be precomputed.

Beam Expansion and Pruning Since CLIP does not provide next-token probabilities, we select beam search candidates using the LLM's next-token prediction results. Specifically, we expand each beam by appending K next-token candidates with the highest probabilities $P_{\text{LLM}}(x_i^{\text{txt}}|x_{1:i-1}^{\text{txt}})$. We then prune the expanded beam candidates (K^2 candidates from K beams) to retain only K beams based on the combined score. Unless otherwise specified, we set $K = 10$ for the experiments.

Beam Search Termination The beam search terminates when either 1) beam expansion fails to improve the score (Eq. 9), or 2) a pre-determined maximum prompt length is reached. The final result $\hat{T}$ is selected among the candidates based on the score.

4 Experiments

4.1 Setup

Datasets We conduct experiments on four datasets with diverse distributions: LAION-400M (Schuhmann et al., 2021; 2022), MS COCO (Lin et al., 2014), Celeb-A (Liu et al., 2015), and Lexica.art [2]. Following PEZ (Wen et al., 2024), we randomly sample 100 images from each dataset and evaluate prompt inversion methods across 5 runs using different random seeds. See Appendix A.3 for more details.

[2] https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts

Table 1: Image quality (CLIP-I score) and prompt quality (BERTScore) comparison.

Method	#Tokens	LAION	MS COCO	Celeb-A	Lexica.art	MS COCO (BERTScore)			Lexica.art (BERTScore)		
		CLIP-I Score				Precision	Recall	F1	Precision	Recall	F1
Textual Inversion	1	0.388	0.569	0.522	0.697	-	-	-	-	-	-
LLaVA-1.5	32	0.513	0.642	0.463	0.580	0.914	0.918	0.916	0.858	0.780	0.817
+ CLIP Interrogator	∼77	0.540	0.685	0.511	0.762	0.794	0.898	0.843	0.818	0.819	0.819
LLaVA-1.5 + **VGD**	32	0.573	0.718	0.546	0.769	0.831	0.879	0.854	0.835	0.793	0.813
	∼77	0.569	0.724	0.544	0.785	0.800	0.874	0.835	0.810	0.794	0.802
PEZ	16	0.538	0.687	0.622	0.743	0.760	0.834	0.795	0.772	0.783	0.777
	32	0.530	0.685	0.619	0.745	0.736	0.830	0.780	0.752	0.784	0.768
	64	0.507	0.670	0.584	0.728	0.715	0.825	0.766	0.734	0.783	0.758
VGD (Ours)	16	0.484	0.650	0.482	0.700	0.833	0.862	0.847	0.827	0.779	0.802
	32	0.493	0.670	0.506	0.735	0.818	0.868	0.842	0.816	0.786	0.801
	64	0.511	0.678	0.514	0.753	0.787	0.863	0.823	0.799	0.789	0.794
	∼77	0.510	0.678	0.513	0.754	0.788	0.864	0.824	0.801	0.791	0.795

Figure 5: Generated images using hard prompts produced by the proposed VGD.

Baselines We compare the proposed method with the hard prompt inversion (PEZ (Wen et al., 2024)), soft prompt inversion (Textual Inversion (Gal et al., 2023)), LLaVA-1.5 (Liu et al., 2024) generated caption, and LLaVA-1.5 combined with CLIP-Interrogator. Images are generated with the Stable Diffusion 2.1-768 model across all comparisons (Podell et al., 2024).

Evaluation Metric We evaluate the quality of the prompt via image embedding similarity between the target image and an image generated using the generated hard prompt. We call this similarity-based metric CLIP-I score. To ensure fairness, we utilize different CLIP models for generation and evaluation: CLIP-ViT-H-14 for generation and larger CLIP-ViT-G-14 for similarity evaluation. Following (Mahajan et al., 2024), we also compare the contextual similarity between the generated prompt and the ground-truth annotations (captions) of the target image, using BERTScore (Zhang et al., 2020) as metric.

4.2 Image Generation with VGD Prompts

In Figure 5, we qualitatively demonstrate that VGD generates realistic and diverse images without overfitting to the original target image.

Qualitative Evaluation As shown in Table 1, the images generated by VGD are the most similar to the original images, as measured by the CLIP-I score, even without using a gradient-based optimization process like PEZ. When using LLaVA-1.5 as the language model, VGD achieves SOTA CLIP-I scores across all four datasets, surpassing the strong baseline that exploits an external database

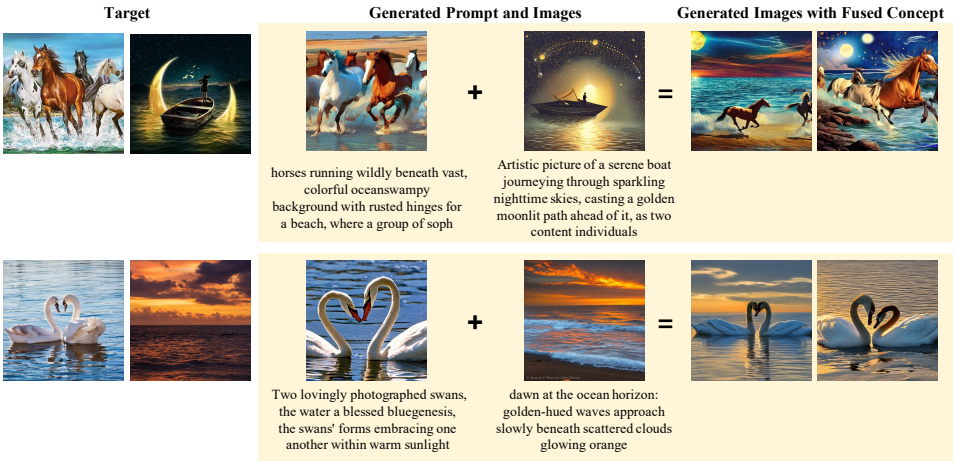


Figure 6: Multi-concept image generation with VGD. We show that the hard prompts obtained from two different images can be concatenated to fuse the semantic concepts.

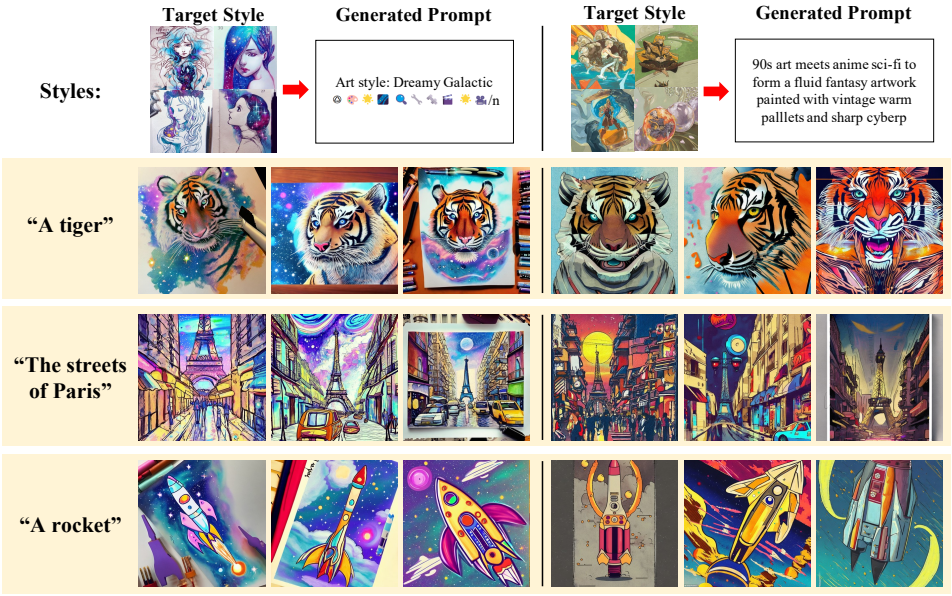


Figure 7: Decoded hard prompt for style transfer. Given several sample images with the same style, we can extract the style with a hard prompt and transfer it to other objects or scenes.

(LLaVA 1.5 + CLIP Interrogator) by a considerable margin. Furthermore, unlike PEZ, VGD exhibits a positive correlation between the CLIP-I score and the number of tokens, highlighting the effectiveness of VGD's token-by-token generation process.

Multi-concept Generation VGD is capable of fusing multiple concepts from different images without additional effort. For each image with its own distinct concept, we individually extract prompts and concatenate them to generate a single image that combines their concepts. Figure 6 presents two examples of multi-concept generation, illustrating that VGD accurately extracts the key elements from each image so that the generated image can contain multiple concepts without omitting any. We show the example of multi-concept generation of 5 images with VGD and PEZ in Figure 19 of the Appendix.

Style Transfer VGD can also be easily adapted to style transfer. Given several examples (typically 4 images) that share the same style, we extract their shared style characteristics as a single hard prompt and use this prompt to apply the style to new objects or scenes. Figure 7 shows two examples of style transfer, demonstrating that VGD correctly embeds the shared style elements in the prompt, which can be easily combined with new objects.

4.3 QUALITY OF VGD PROMPTS

Quantitative Evaluation To assess the coherence of generated prompts, we measure the BERTScore between the prompts generated by each method and the ground-truth captions. As shown in Table 1, VGD achieves a significant improvement in BERTScore compared to PEZ, especially when using more than 64 tokens. Although LLaVA-based baselines show higher BERTScore, their CLIP-I scores are worse than those of VGD. This implies that the LLaVA captions fail to describe the target image in detail, even though the text itself may be semantically similar to ground-truth captions.

Interpretability When a prompt closely aligns with the image content and is interpretable by humans, it becomes easier for users to modify the image in the desired direction by making slight edits to the prompt. To evaluate the interpretability of the generated prompts, we make minimal modifications to the prompt and observe the effects. For example, we change "winter" to "spring" and see if the generated image successfully reflects the change. As illustrated in Figure 9, we compare VGD with previous gradient-based prompt inversion techniques, namely PEZ (hard) and Textual Inversion (soft).

The results indicate that VGD exhibits superior interpretability. When we adjust the number of objects, background elements, atmosphere (seasons), and species in the prompt, VGD successfully incorporates these changes. The relevant terms are already present in the original prompt, enabling accurate image generation with the modified prompts. In contrast, the competitors struggle to produce the desired outcomes. For instance, when we try to generate four wolves, both PEZ and Textual Inversion fail to change the number of wolves included.

Prompt Distillation We can also utilize prompt inversion to reduce the length of prompts while preserving their capability, *a.k.a.* prompt distillation. Long prompts may contain redundant and unimportant information, especially when hand-crafted. This becomes problematic, particularly when there exists a limit on the number of tokens that the model can process. Using VGD, we can generate a shorter prompt that preserves the key aspects of the original prompt. As illustrated in Figure 10, the images generated with distilled prompts are very similar to the original image. In addition, images generated with distilled prompts do not show significant performance degradation on CLIP-I score compared to the original prompt, as shown in Figure 8.

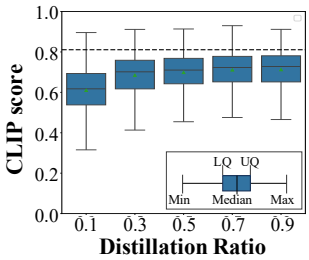


Figure 8: Prompt distillation on Lexica.art dataset.

4.4 EFFECT OF CLIP AND LANGUAGE MODEL

Table 2: Comparison of VGD variants.

Model	CLIP Score	BERTScore		
		Precision	Recall	F1
VGD	0.735	0.816	0.786	0.801
LLM only	0.487	0.811	0.773	0.791
CLIP only	0.768	0.727	0.779	0.751

Table 3: Ablation on CLIP models.

CLIP Model	#Tokens	CLIP Score
laion/ViT-H-14 (original)	32	0.735
openai/ViT-B-32	32	0.655
openai/ViT-L-14	32	0.626

Role of CLIP and LLM To investigate the effect of using both CLIP and LLM, we modify VGD in two ways and compare these variants on the Lexica.art dataset. In the 'LLM only' variant, VGD selects the next token based solely on $P_{\text{LLM}}(x_i^{\text{txt}}|x_{1:i-1}^{\text{txt}})$ without considering CLIP score, operating like typical LLM text decoding. In the 'CLIP only' variant, VGD determines the next token using only the CLIP model. Table 2 shows that the 'LLM only' variant suffers from a drastic performance degradation in the CLIP score because it does not consider the target image at all. On the other hand, the 'CLIP only' variant achieves the highest CLIP score but the lowest BERTScore, suggesting that the generated prompt is less coherent and harder to interpret. In comparison, the proposed VGD demonstrates strong and balanced performance on both the CLIP score and BERTScore.

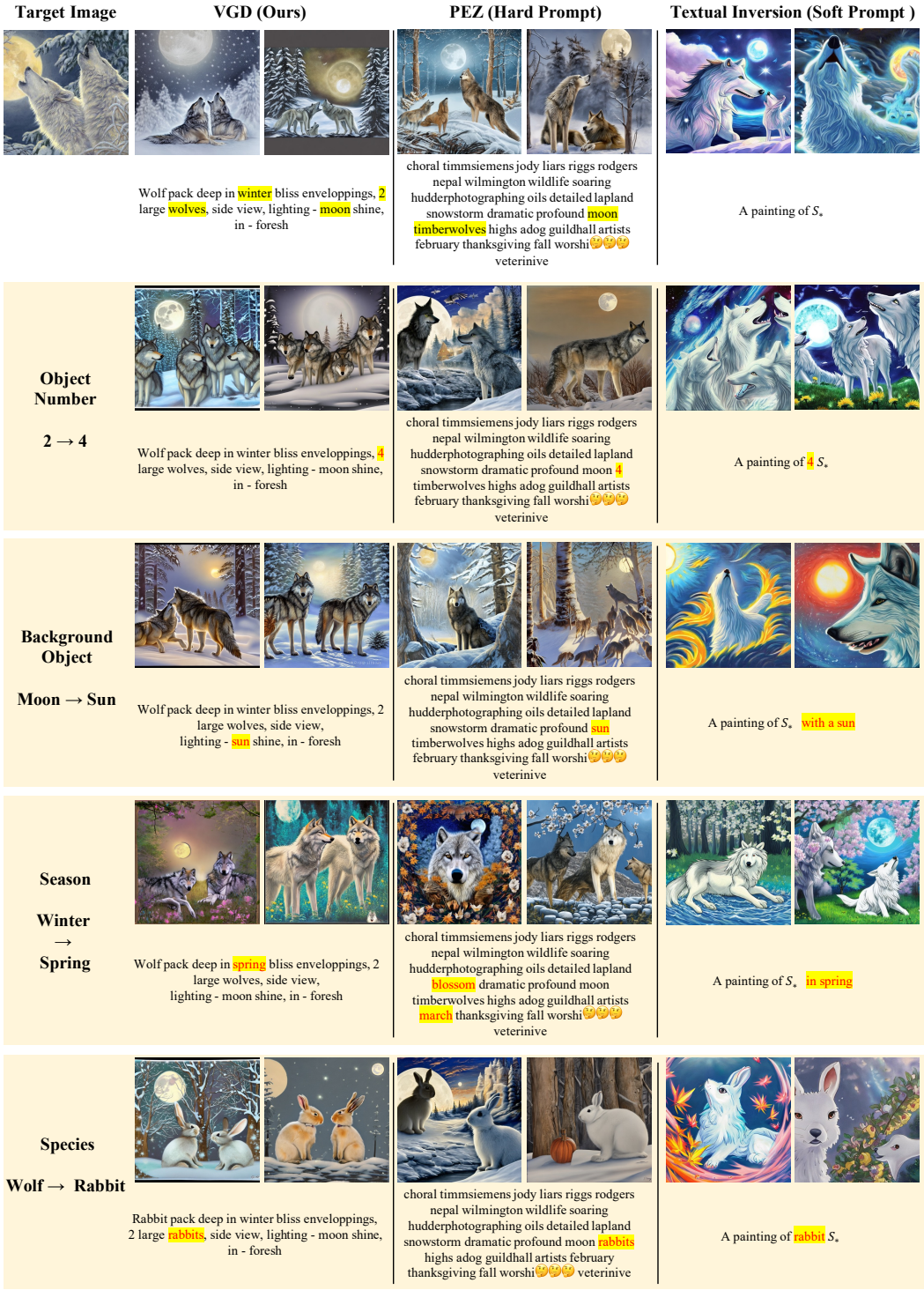


Figure 9: Prompt interpretability. With the good interpretability of the prompt, we can modify them to edit the original image to the desired direction.

CLIP Model Ablation VGD leverages the same CLIP model that is used in the text-to-image model, Stable Diffusion 2.1. To assess the importance of this alignment, we perform an ablation study using two alternative CLIP models: `openai/ViT-B-32` and `openai/ViT-L-14`. Table 3 shows that using a misaligned CLIP model leads to significant performance degradation. This highlights the importance of the proper approximation $P(T|I) \simeq P_{\text{CLIP}}(T|I)$ in the decoding pro-

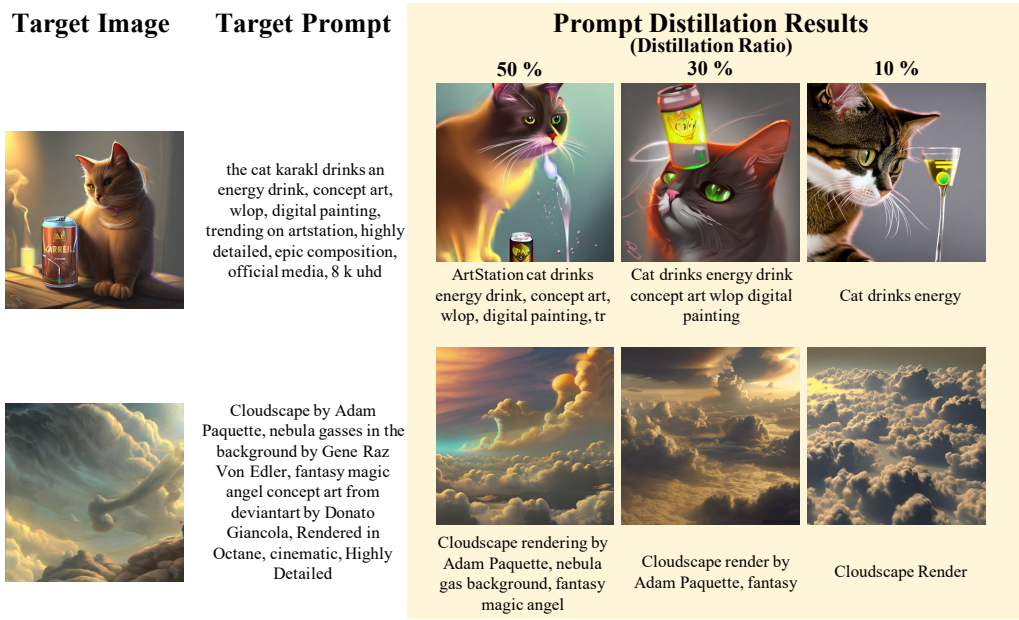


Figure 10: Prompt distillation results generated by the proposed VGD. With VGD, we can distillate the hard prompt that can still generate images similar in concept to the original.

Table 4: Ablation on language models.

LLM	#Tokens	LAION	MS COCO	Celeb-A	Lexica.art	MS COCO			Lexica.art		
		CLIP-I Score				Precision	Recall	F1	Precision	Recall	F1
Mistral-7B	16	0.443	0.615	0.422	0.706	0.808	0.850	0.828	0.815	0.780	0.797
	32	0.463	0.656	0.466	0.729	0.803	0.858	0.829	0.808	0.787	0.797
	64	0.497	0.680	0.476	0.751	0.783	0.858	0.818	0.791	0.789	0.790
LLaMA2-7B	16	0.484	0.650	0.482	0.700	0.833	0.862	0.847	0.827	0.779	0.802
	32	0.493	0.670	0.506	0.735	0.818	0.868	0.842	0.816	0.786	0.801
	64	0.511	0.678	0.514	0.753	0.787	0.863	0.823	0.799	0.789	0.794
LLaMA3-8B	16	0.480	0.648	0.521	0.731	0.820	0.861	0.840	0.829	0.784	0.806
	32	0.506	0.661	0.540	0.752	0.805	0.864	0.833	0.817	0.790	0.803
	64	0.515	0.685	0.543	0.769	0.769	0.858	0.811	0.796	0.794	0.795

cess. When a different CLIP model is used, the approximation no longer holds, leading to a drop in performance.

Language Model Ablation We conduct an ablation study on the language model using three different LLMs: Mistral-7B (Jiang et al., 2023), LLaMA2-7B (Touvron et al., 2023), and LLaMA3-8B (Dubey et al., 2024). As shown in Table 4, although LLaMA3-8B exhibits the best performance, the overall differences among the models are marginal. This indicates that VGD is not dependent on the specific language model used, allowing users to choose an LLM that best suits their system.

5 Conclusion

This paper tackles the difficulty of creating effective textual prompts for advanced text-to-image models like DALL-E and Stable Diffusion. Traditional prompt inversion methods often lack interpretability and coherence, limiting their usefulness. We introduce VGD, a gradient-free technique that combines LLMs with CLIP-based guidance to generate clear and semantically accurate prompts. VGD enhances prompt interpretability, generalization, and flexibility without requiring additional training. Our experiments show that VGD outperforms existing methods, producing more understandable and contextually relevant prompts. By integrating seamlessly with various LLMs such as LLaMA2, LLaMA3, and Mistral, VGD provides an efficient and versatile solution for improving user interactions with text-to-image generative models.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 936–945, 2017.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2): 263–311, 1993.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Vetim: Expanding the vocabulary of text-to-image models only with text. In *The 34th British Machine Vision Conference (BMVC 2023)*. BMVA, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7545–7556, 2023.

Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6986–6996, 2024.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–850, 2019.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Daniel Jurafsky. Speech and language processing, 2000.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 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.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, November 2021.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6808–6817, 2024.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5571–5584, 2023.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2024.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13215–13224, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

A DETAILS

A.1 PROMPT SETUP FOR LARGE LANGUAGE MODELS

We use different system and user prompts for different image generation tasks. Tables 5 and 6 show the prompts used for the experiments.

Table 5: LLM prompt setting for different image generation tasks.

System Prompt
You are a respectful and honest visual description generator for Stable Diffusion text prompt. Answer in 1 sentence and do not mention anything other than the prompt. Do not mention 'description'.
User Prompt (Prompt Inversion)
Please generate the diffusion prompt on the given condition containing the objects, people, background, and the style of the image:
User Prompt (Style Transfer)
Please generate the diffusion prompt of the image style based on the given condition containing the painting style, color, and shapes of the image:
User Prompt (Prompt Distillation)
Please generate the diffusion prompt within ${max_length} tokens so that you can generate same images with a given prompt: ${target_prompt}
Model prompt
Answer: Sure, here is a prompt for stable diffusion within ${model.max_length} tokens:

Table 6: LLaVA prompt setting for different image generation tasks.

Prompt (Image Captioning)
USER: < image > Describe the scene in this image with one sentence. ASSISTANT:
Prompt (VGD)
USER: < image > Please generate the diffusion prompt containing the objects, people, background and the style of the image. ASSISTANT: Sure, here is a prompt for stable diffusion within ${model.max_length} tokens:

A.2 HYPERPARAMETERS AND MODELS

The beam width K is set to 10. The balancing hyperparameter α is set to 0.67. For VGD generation, we used `laion/CLIP-ViT-H-14-laion2B-s32B-b79K`[3]. For CLIP-I score evaluation, we used `laion/CLIP-ViT-g-14-laion2B-s12B-b42K`[4]. We used Stable Diffusion `stabilityai/stable-diffusion-2-1`[5] for text-to-image model.

A.3 DATASET CHARACTERISTICS

LAION-400M (Schuhmann et al., 2021; 2022) comprises 400 million diverse CLIP-filtered images. MS COCO (Lin et al., 2014) mainly contains real-life photographs with multiple common objects,

[3] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K
[4] https://huggingface.co/laion/CLIP-ViT-g-14-laion2B-s12B-b42K
[5] https://huggingface.co/stabilityai/stable-diffusion-2-1

whereas Celeb-A (Liu et al., 2015) consists of celebrity portraits. Lexica.art contains AI-generated paintings with their prompts. While different datasets exhibit different characteristics, VGD works effectively regardless of the data type.

A.4 CLIP-Interrogator's Approach

CLIP-Interrogator supplements missing details in generated image captions by extensive trial-and-error approach. Specifically, CLIP-Interrogator augments the caption with additional tokens selected from a pre-collected bank of 100K keywords (*e.g.*, artist names, styles, mediums, movements), and compare each with the target image through CLIP model to filter out image-irrelevant keywords. While effective in generating the hard prompt, CLIP-Interrogator has several drawbacks: 1) an inability to generate images for styles, artists, or objects that are not registered in the keyword bank, and 2) the computational burden of processing 100K phrases through a CLIP encoder. These limitations highlight the challenges of relying solely on image captions for image synthesis, and underscore the need for efficient and flexible hard prompt generation techniques.

B Additional Related Works

B.1 Prompt Editing for Text-to-Image Models

To fully harness the capabilities of text-to-image models, various approaches have been proposed to tailor text prompts to the user's specific intentions (Hertz et al., 2023; Brooks et al., 2023; Wang et al., 2024) Notably, a rich text editor has been introduced, enabling users to design complex textual guidance to image generation models (Ge et al., 2023). Similarly, Prompt Highlighter, an interactive text-to-image interface utilizing multi-modal LLMs, has been proposed (Zhang et al., 2024). These studies underscore the necessity for user-friendly and controllable text editing methods. Our VGD can be seamlessly integrated into LLM-based user interfaces, as illustrated in Figure 1. VGD leverages the same token generation process as LLMs, benefiting from the highly optimized inference process without requiring additional training.

B.2 Interpretability of Text-to-Image Models

As text-to-image applications become standard in the industry, interpretability studies for these models have gained significant attention. For instance, research has focused on the cross-attention mechanism to identify the model's attention for each word (Tang et al., 2023). In advance, the diffusion objective has been investigated to understand the internal behavior of text-to-image models (Kong et al., 2024). Building on these findings, recent works modified the model's architecture to more faithfully reflect user's instructions (Orgad et al., 2023; Guo & Lin, 2024). We emphasize that VGD aims to enhance the human interpretability without architectural modifications. We believe that previous works can also be also applied to VGD to further improve usability.

C Additional Experimental Results

C.1 Generalization of Generated Prompts to Different T2I Models

We demonstrate how prompts generated by our method and other baseline methods behave when applied to different text-to-image models (see Fig. 13, 14, and 15). In this experiment, we did not generate any new prompt for the newly tested T2I model (*i.e.*, MidJourney, DALL-E 2), meaning that we simply reused the prompts generated for SD in evaluating the effectiveness of producing high-quality images with various T2I models. In doing so, we can ensure a fair comparison and also highlight the generalizability of the prompts across models. Finally, we would like to point out that this type of generalization cannot be done with soft prompt methods like Textual Inversion, as they are inherently tied to the specific model they are trained on. When compared to hard prompt inversion methods (such as PH2P and PEZ), VGD generates consistently better results in various text-to-image models, which underscores the robustness and flexibility of the proposed VGD.

C.2 Prompt Generation Time

We further investigate the efficiency of VGD in comparison with other baseline methods, measured on a single A100 80GB GPU. As shown in Table 7, the proposed VGD needs significantly less time to generate each image; 7x faster than Textual Inversion, 12x faster than PEZ, and more than 2200x faster than PH2P. This is because VGD does not require any training steps and only the decoding process is slightly modified.

Table 7: Prompt generation time of VGD, PEZ, PH2P, Textual Inversion, and CLIP Interrogator.

Method	PEZ	PH2P	Textual Inversion	CLIP Interrogator	**VGD (ours)**
Time (sec)	193.14	34264.74	103.45	24.85	**15.33**

C.3 More Qualitative Comparison

As shown in Fig. 16 and 17, we conduct a qualitative comparison of image variation and style transfer performance across different methods, including PEZ, Textual Inversion, and CLIP Interrogator. The results demonstrate that, similar to the findings in Table 1, our method outperforms prior works, with CLIP Interrogator coming second.

While CLIP Interrogator captures objects effectively by generating captions using LLaVA (Liu et al., 2024) and predefined rich keyword bank, there are some critical weaknesses in prompts generated by the CLIP Interrogator (see Fig. 16). First, the subsequent keywords extracted from the caption's keyword bank using CLIP similarity often suffer from redundancy and lack of interpretability. Second, CLIP Interrogator tends to generate prompts of full length (77 words), which limits its applicability to diverse tasks like multi-concept image generation and style transfer. Finally, as seen in the prompt example for the tiger image in Fig. 16, the quality of the generated images often decreases when using the extracted keywords, compared to images generated without attaching the keywords. In some cases, this leads to all extracted keywords being rejected entirely.

C.4 Human Evaluation

To evaluate whether the images generated by VGD are semantically aligned with the original image and to assess image quality, we conduct human preference evaluation. For this experiment, participants were shown images generated for image variation and style transfer by five different methods (VGD, Textual Inversion, PEZ, PH2P, and CLIP Interrogator) and were asked to select the image most similar to the target image and style (Kumari et al., 2023) (see Fig. 12). The evaluation was conducted using 65 images randomly sampled from Lexica.art dataset. Fig. 11 demonstrates that our method consistently performed best in both image variation and style transfer tasks according to user preferences. Specifically, 60.5% and 52.6% of the participants selected VGD's results as the most similarly generated ones for image variation and most instruction-following generations for style transfer tasks, respectively.

C.5 Bad Cases

While VGD generally shows decent results, our method sometimes struggles to capture fine-grained details of regional or background objects in complex images. These cases occur when the image contains multiple objects (see Fig. 18). This is because VGD has difficulty in generating prompts for regional or local objects in the background. We believe this phenomenon is mainly due to the CLIP vision encoder's tendency to prioritize main objects in an image (Jing et al.; Paiss et al., 2023; Ranasinghe et al., 2023; Yuksekgonul et al., 2022; Zhong et al., 2022).

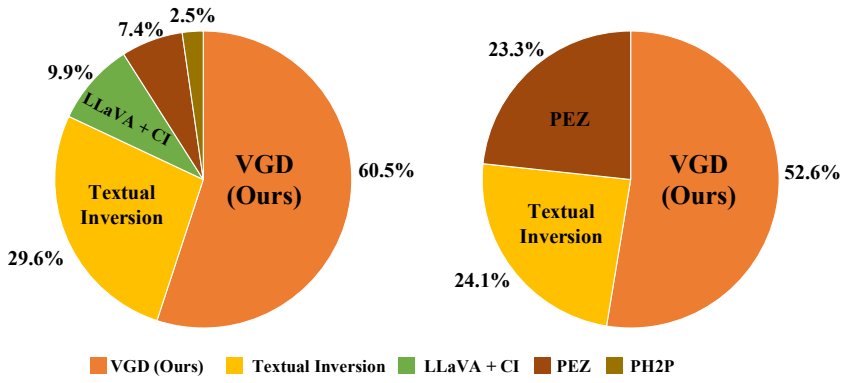
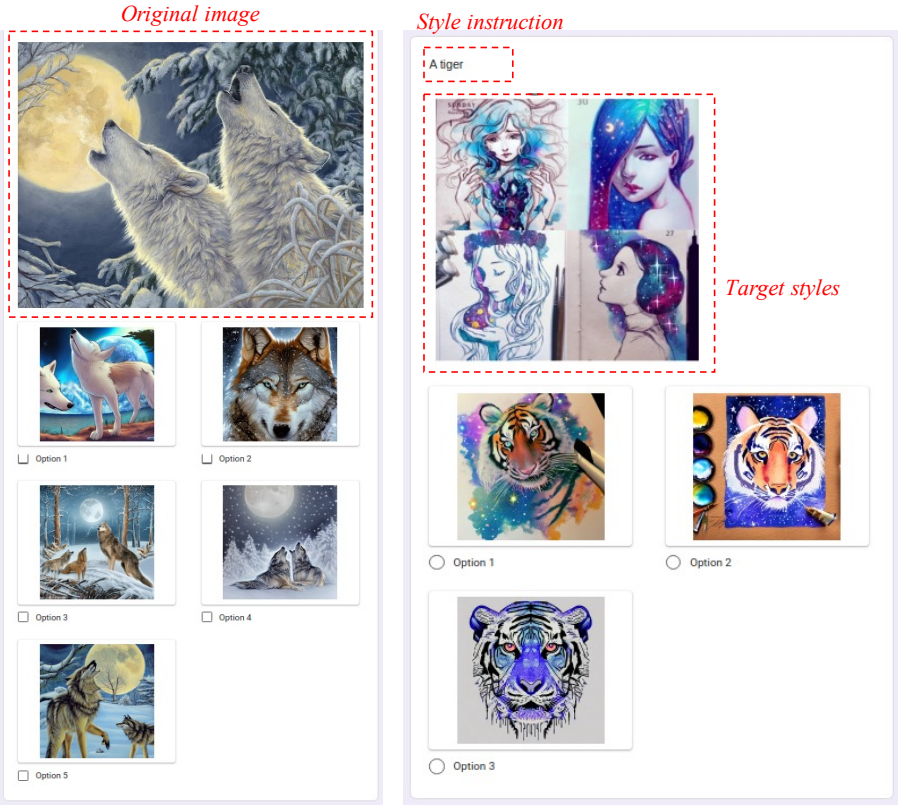


Figure 11: Human preference evaluation for image variation (left) and style transfer (right) tasks using different prompt generation methods. CI indicates CLIP Interrogator. Best viewed in color.



(a) Image Variation Task (b) Style Transfer Task

Figure 12: Webpage used for human preference evaluation.

C.6 Multi-concept Generation

We further explore the potential of VGD by combining five different concepts into a single image. Figure 19 demonstrates that the prompts generated by VGD can be concatenated to generate a complex image without missing concepts. However, generated prompts from PEZ suffers omits important details, such as the parliamentary, wolves, and the style of The Starry Night. This shows that VGD generates hard prompts that are compact and meaningful.

D LIMITATION AND ETHICAL STATEMENT

D.1 LIMITATIONS

Although we showcase the generalizability of VGD-generated prompts across multiple text-to-image models, our evaluations primarily focus on popular models such as Stable Diffusion and MidJourney. However, since many other models are trained on similar datasets and follow comparable methods and architectures, we anticipate that the results would be consistent across these models as well.

D.2 ETHICS STATEMENT

The development of VGD raises important ethical considerations, particularly regarding potential misuse. While our method improves the interpretability and generalizability of text-to-image generation, it could also be exploited to generate harmful or inappropriate content, such as deepfakes or offensive imagery. We strongly discourage such uses and emphasize that this method is intended for responsible applications in creative, educational, and professional contexts.

Figure 13: Qualitative examples generated by Stable Diffusion, Midjourney and Dall-E 2 using prompts generated from VGD, PEZ, and LLaVA+CLIP-Interrogator.

Figure 14: More examples generated by MidJourney.

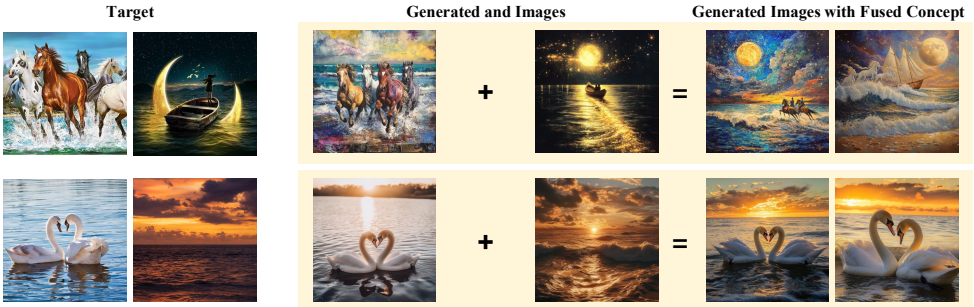


Figure 15: Multi-concept image generation by MidJourney.

Figure 16: Image variation comparison between VGD and baselines.

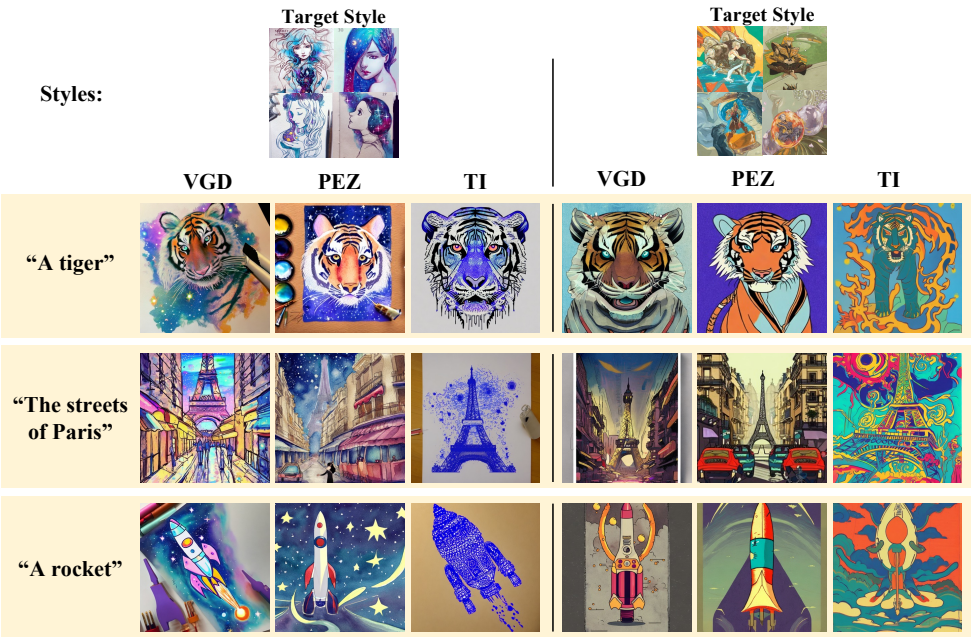


Figure 17: Style transfer comparison between VGD and baselines.

Figure 18: Bad case examples of VGD.

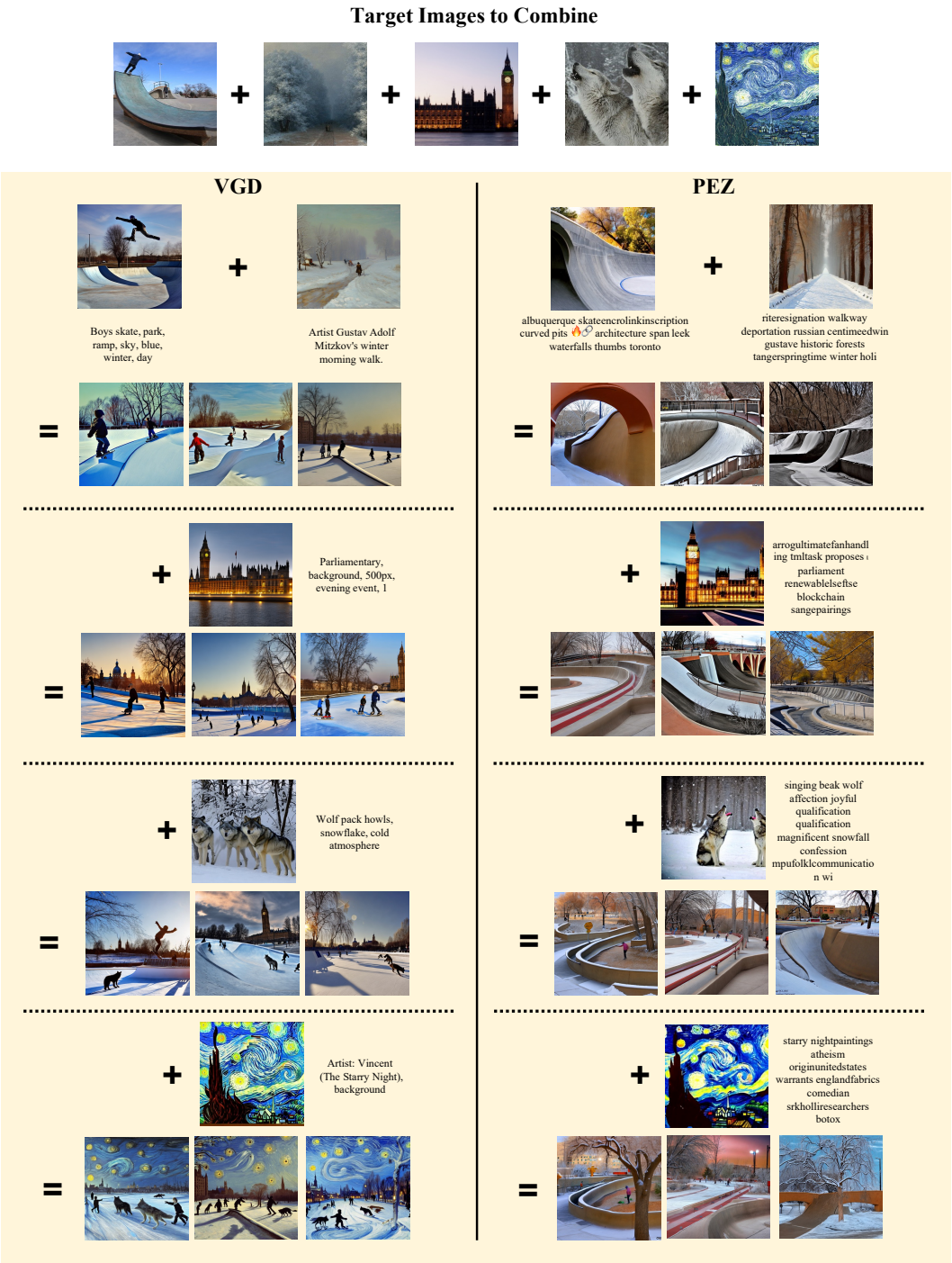


Figure 19: Five-concept image generation comparison between VGD and PEZ.