Prompting Hard or Hardly Prompting: Prompt Inversion for Text-to-Image Diffusion Models

Shweta Mahajan[1,2] Tanzila Rahman[1,2] Kwang Moo Yi[1] Leonid Sigal[1,2]
[1]University of British Columbia
[2]Vector Institute for AI
{s.mahajan, trahman8, kmyi, lsigal}@cs.ubc.ca

Abstract

The quality of the prompts provided to text-to-image diffusion models determines how faithful the generated content is to the user's intent, often requiring 'prompt engineering'. To harness visual concepts from target images without prompt engineering, current approaches largely rely on embedding inversion by optimizing and then mapping them to pseudo-tokens. However, working with such high-dimensional vector representations is challenging because they lack semantics and interpretability, and only allow simple vector operations when using them. Instead, this work focuses on inverting the diffusion model to obtain interpretable language prompts directly. The challenge of doing this lies in the fact that the resulting optimization problem is fundamentally discrete and the space of prompts is exponentially large; this makes using standard optimization techniques, such as stochastic gradient descent, difficult. To this end, we utilize a delayed projection scheme to optimize for prompts representative of the vocabulary space in the model. Further, we leverage the findings that different timesteps of the diffusion process cater to different levels of detail in an image. The later, noisy, timesteps of the forward diffusion process correspond to the semantic information, and therefore, prompt inversion in this range provides tokens representative of the image semantics. We show that our approach can identify semantically interpretable and meaningful prompts for a target image which can be used to synthesize diverse images with similar content. We further illustrate the application of the optimized prompts in evolutionary image generation and concept removal.

1. Introduction

Text-to-image conditional diffusion models [31, 38, 41] are trained on an enormous amount of image-text data and have transformed the domain of generative learning in computer vision. These models demonstrate exceptional generative capabilities and provide an opportunity for creative image generation with customized concepts [22, 32, 34, 35, 43]. The quality and the regularity of the content produced by diffusion models are, however, subject to the quality of the input prompts; which in turn depends on the training data [16, 57]. For pre-trained diffusion models, identifying and formulating the prompts that produce the intended visual content is challenging in the large-scale data regime.

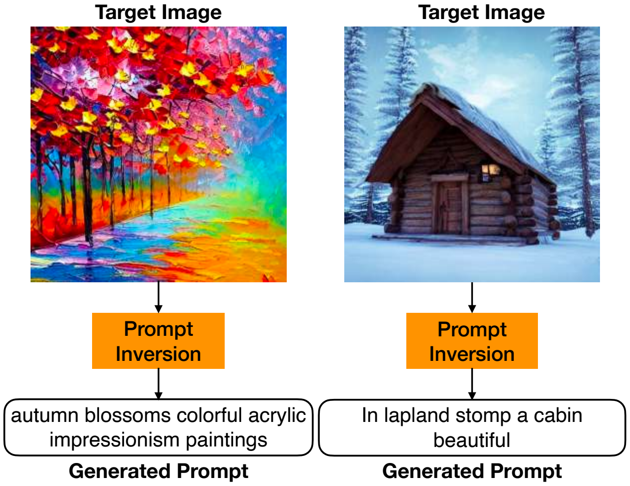

Target Image
Target Image
Prompt Inversion
Prompt Inversion
autumn blossoms colorful acrylic impressionism paintings
In lapland stomp a cabin beautiful
Generated Prompt
Generated Prompt

Figure 1. **Illustration of Hard Prompt Inversion.** Target images (top) along with inverted prompts (bottom) obtained using proposed *PH2P* approach; note the exhibited prompt semantics.

For example, to generate the painting shown in Fig. 1, one would require an understanding of art styles such as *acrylic* or *impressionism*. These concepts, though encoded in the diffusion model, might not be accessible without specialized domain knowledge or without access to the model's training vocabulary. Additionally, in many downstream applications of diffusion models such as image composition [24, 27, 49] or segmentation [50] target concepts are available solely as images and require the discovery of the appropriate prompt as an intermediate step. This drives a relevant research question: *What is a likely prompt that would generate visual contents of the target image?* Discovering, or optimizing, such prompts would greatly simplify creative and editing processes; particularly for novice users.

Prompts for the visual content of interest are predominantly hand-crafted through laborious trial and error requiring human intervention and expertise [17, 20, 30]. Inspired by the inversion techniques that search semantics of the target images in the latent space of generative adversarial networks (GANs) [2, 8, 16, 58], recent work has instead resorted to the automated discovery of the target visual concepts through inversion of diffusion models [14, 24, 40, 48]. With text-guided synthesis at its core, text-conditioned diffusion models apply inversion to the conditioning representations. Specifically, new pseudo-tokens are introduced in the model vocabulary and the corresponding embeddings are optimized for the target image(s). These embeddings, however, have been found to display less or no correlation to the original model vocabulary [40]. Moreover, a single vector representation encodes varied concepts for a collection of images, thereby restricting the readability and generalization capabilities of the learned concepts.

The aforementioned limitations can be addressed by inverting the diffusion model to yield the text tokens *i.e*. hard prompts within the model's vocabulary.[1] Optimization of the hard prompts involves a discrete optimization procedure with re-projection of the learned embedding onto the nearest neighbor in the pre-specified vocabulary [49]. Hard prompt inversion in diffusion models is particularly cumbersome owing to the complexity of the generative model with a two-stage pipeline consisting of the conditioning network (CLIP text encoder; [33]) and the generative component (a U-Net [39]) making the flow of gradients to the input layer difficult (vanishing gradients). Secondly, the diffusion trajectory has high variance given the stochastic components in the generative model [44]. As a result, it can be observed (see Sec. 4), that standard SGD-based optimization with projection leads to poor performance in practice.

We overcome these challenges for hard prompt inversion and make the following significant contributions:

- We study the effect of prompt conditioning at different timesteps of the diffusion process (Fig. 2b). We observe that noisy, later steps, have greater sensitivity to prompt conditioning.
- Based on our findings, we propose a *Prompting Hard or Hardly Prompting (PH2P)*[2] inversion procedure where the diffusion loss is applied to the sub-range of the timesteps, thereby reducing the high variance of the optimization process. Prompt inversion in the conditioning-sensitive range also provides with better flow of gradients. We employ quasi-newton L-BFGS [42] based reprojection techniques to learn discrete tokens. This provides a refined framework for sensitive multi-token optimization.
- We empirically validate that our approach yields semantically meaningful prompts that can synthesize accurate yet diverse images for a target visual concept. The prompts obtained with our proposed inversion technique are interpretable and applicable to a broad set of tasks, such as evolutionary image generation and concept removal in images through negative visual concepts.

2. Related Work

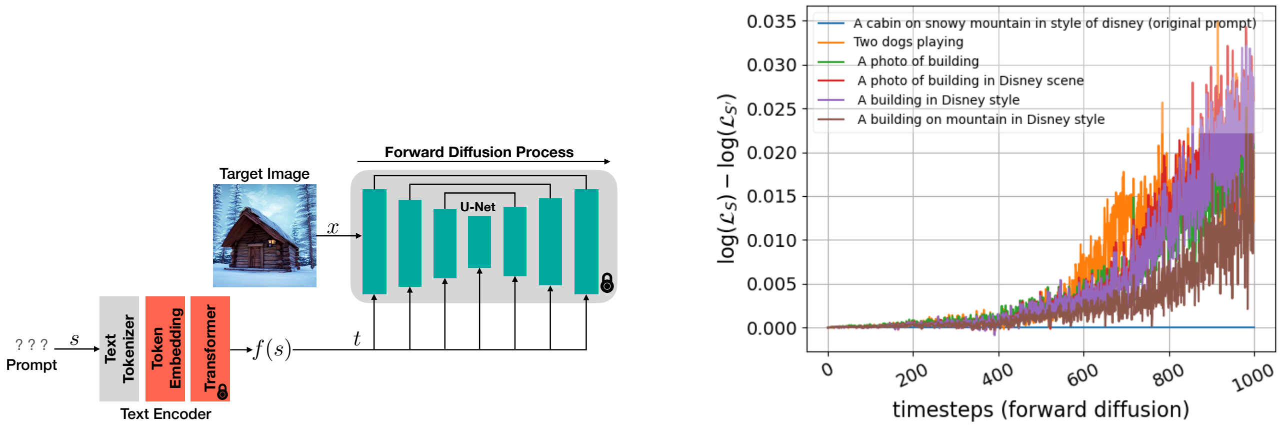
Text-guided image synthesis. Extension of deep generative models to conditional, controllable generation is well-founded in generative adversarial networks (GANs) [15, 37, 46, 52], variational autoencoders (VAEs) [23, 53, 59] and normalizing flows [1, 12, 13]. Early work for text-to-image synthesis built upon GANs [37] and more recently on the normalizing flow-based priors [28] to align the image-text distributions. High quality and realistic image synthesis is now possible with models such as DALL-E [35] and Cogview [11] that build upon the more expressive neural architectures such as transformers [47] and discrete variational autoencoders (VQ-VAE) [36].

High-fidelity image generation has been revolutionized with diffusion models [21] and has gained traction for controllable generation [18, 31, 38, 54]. Nichol *et al*. [31] use CLIP [33] guidance to replace class labels and classifier-free guidance for conditional generation. Saharia *et al*. instead of training the text encoder on paired image-text data, utilize a pre-trained language model as a text-encoder [9]. Recent approaches [18, 38] apply diffusion models to the low-dimensional representations of the input thereby reducing the complexity of the generative component. These low-dimensional diffusion models trained on large-scale datasets have become a standard backbone for the current generative modeling tasks such as visual art generation [22, 32] or multi-frame generation in videos or stories [34, 43]. The quality of the synthesized content in these models is influenced by the guiding text. It is important to find the correct text that yields desired visual concepts of interest and therefore, in this work, we develop our approach with latent diffusion models as working backbone.

Prompting in diffusion models. The natural language descriptions are referred to as discrete prompts or hard prompts and have a significant impact on image generation in diffusion models [17]. To this end, various approaches have explored diverse prompting techniques based on retrieval and captioning to enhance the quality of the prompts [19, 30, 50] and consequently image synthesis. Wu *et al*. [50] introduce sub-classes in the prompts based on the main class in images. Ni *et al*. [30] specify the name of the target object to be generated as input to a general purpose transformer (GPT) [4]. Hertz *et al*. [20] modify words in the

[1] 'Hard' here also refers to the fact that these tokens are discrete choices, and not 'soft' weightings of the training vocabulary, which result in continuous vectors in the text embedding space.

[2] Code available at https://github.com/ubc-vision/Prompting-Hard-Hardly-Prompting



(a) Overview of prompt optimization with Latent Diffusion Model.

(b) Effect of the conditioning on diffusion objective at different timesteps.

Figure 2. **Prompt Inversion in Diffusion Models.** *Left*: Overview of the text-to-image diffusion model with a CLIP text-encoder. *Right*: Analysis of the sensitivity of different timesteps to conditioning information for a given target image illustrated on the left.

text and inject the cross-attention maps of the pixels corresponding to target styles. Moreover, Brack *et al*. [3] relies on user-defined prompts for image editing. Our work, on the other hand, aims to generate the prompts directly from the text-prior within the diffusion model without any auxiliary task or model and without human intervention.

Inversion in diffusion models. Inspired by the optimization methods for inversion in GANs [2, 8, 16, 58], optimization-based approaches have been developed for diffusion models. In order to synthesize concepts with novel scenes, in diffusion models, the few approaches directly perform inversion in the image space [6, 10, 29, 44]. Song *et al*. and Mokady *et al*. [29] approaches search for the initial noise that reconstructs the image [44]. Instead of working directly with the image features, recent approaches benefit from the controllability in the textual embedding space [14, 40, 48]. At a high-level, these methods invert visual concepts in the embedding space of the text and encode new features as new tokens in the vocabulary. The DreamBooth approach of [40] fine-tunes the diffusion model to learn a new visual featues and Kumari *et al*. update the parameters of the cross-attention layers. Textual inversion [14] inverts the concepts from a set of reference images in the textual embedding space by introducing a new "pseudo-word" and adding a new representation to the model vocabulary. Voynov *et al*. [48] introduce a new token for each of the layers of the U-Net in the diffusion model. The approaches provide better editing freedom to the users compared to approaches that modify the image space. However, the inverted features are not interpretable in the space of text and therefore, have limited generalization. Moreover, the recovered embeddings can be redundant and the image-level features may be present in the model vocabulary. Recent, unpublished as of this submission, work of Wen *et al*. [49] attempts to optimize hard prompts for CLIP [33] with image-text similarity matching. Similar in spirit, to avoid biases, we propose hard prompt optimization that leverages text-conditioned diffusion model and its loss directly.

3. Prompt Inversion for Diffusion

While prior work for textual inversion [14, 48] is confined to the embedding space of the conditioning variable, in this work, we explicitly aim to find the text tokens that under the parameterized diffusion model would yield the target image, *i.e*., *what prompt would likely generate the desired image?* We formalize our approach under the constraint that we do not have access to any additional auxiliary image-text similarity model and have access only to the pre-trained conditional diffusion model along with the text encoder used for conditioning.

3.1. Conditional Diffusion Model Backbone

We first provide an overview of the model components: the diffusion model with U-Net and the conditioning network, that we consider in the prompt inversion.

Latent diffusion model. We consider the Latent Diffusion Model (LDM) [38] as the class of generative model for which we recover the prompt tokens likely to generate the target image. In LDM, the diffusion process is applied to a lower-dimensional spatial representation $\mathbf{x}$ of the input image $\mathbf{I}$. An encoder $E(\cdot)$ maps the input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ to a latent representation $\mathbf{x} \in \mathbb{R}^{h\times w\times c}$, downsampling the image to a lower spatial dimension. The decoder $D(\cdot)$ maps $\mathbf{x}$ back to the original image space. In a conditional setting, the diffusion model is applied to the representation $\mathbf{x}$, where time-conditioned U-Net [39] $\epsilon_\theta(\mathbf{x}_t, t)$ is employed to model the diffusion process. The diffusion process is applied over T timesteps where noise is gradually added to $\mathbf{x}$ to generate a series of noisy representations $\mathbf{x}_{1:T}$.

For a conditional generative model, in addition to the timesteps, the U-Net is conditioned on $f(\mathbf{S})$, where conditioning input $\mathbf{S}$ is encoded with the mapping $f(\cdot)$. The

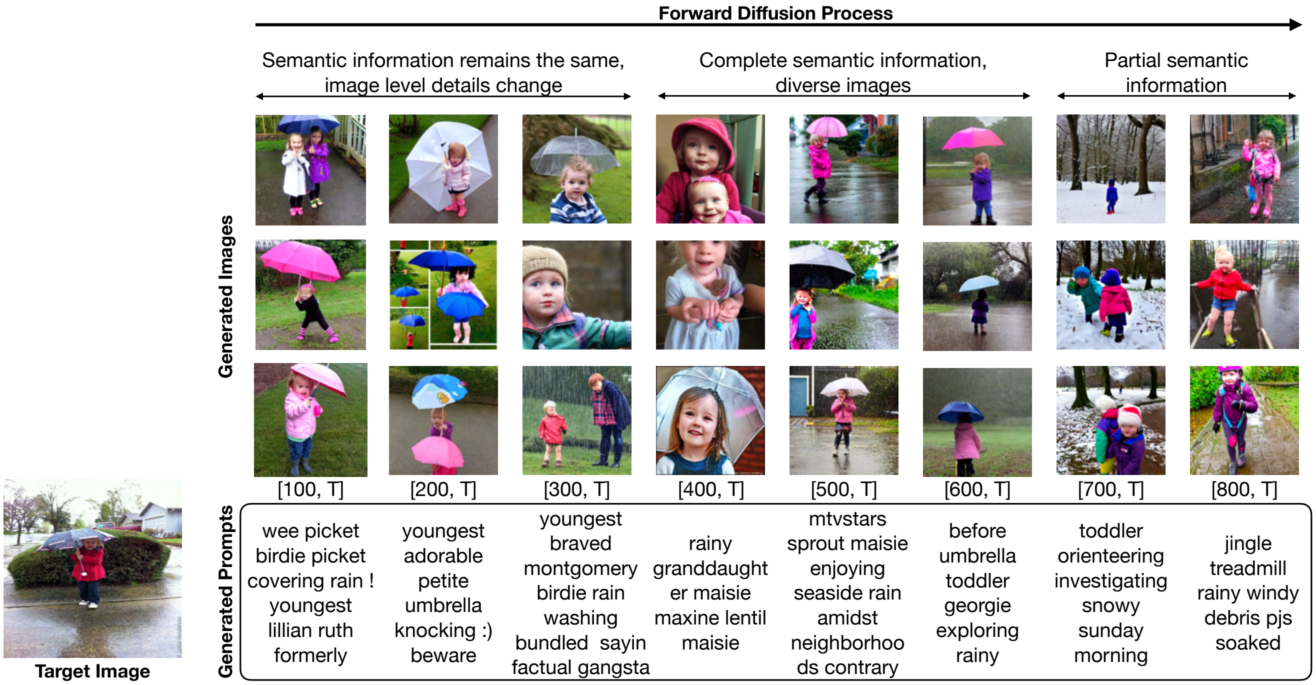

Forward Diffusion Process
Semantic information remains the same, image level details change
Complete semantic information, diverse images
Partial semantic information
Generated Images
[100, T]
[200, T]
[300, T]
[400, T]
[500, T]
[600, T]
[700, T]
[800, T]
Target Image
Generated Prompts
wee picket birdie picket covering rain ! youngest lillian ruth formerly
youngest adorable petite umbrella knocking :) beware
youngest braved montgomery birdie rain washing bundled sayin factual gangsta
rainy granddaughter maisie maxine lentil maisie
mtvstars sprout maisie enjoying seaside rain amidst neighborhoods contrary
before umbrella toddler georgie exploring rainy
toddler orienteering investigating snowy sunday morning
jingle treadmill rainy windy debris pjs soaked

Figure 3. **Effects of Timesteps on Prompt Inversion.** Prompts generated for the different sections of timesteps in text-to-image diffusion models with maximum noise at timestep T. The images generated show that the semantic information is completely recovered for the range starting from the middle timesteps. Once the semantic information appears in the images, only image-level details change when considering the larger range starting from the early steps of the forward diffusion process.

objective of the conditional latent diffusion model is,

$$\mathcal{L}_{LDM} := \mathbb{E}_{t,\mathbf{x},\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, f(\mathbf{S}))\|_2^2\right]. \tag{1}$$

During training, given the image $\mathbf{x}$ and the conditioning input *e.g.* a prompt $\mathbf{S}$, a timestep is randomly chosen from $\{1, \ldots, T\}$ at each learning step and the model parameters θ that minimize Eq. (1) are estimated.

Embedding space for optimization. The text encoder $f(.)$ takes the input prompt $\mathbf{S}$, which is input to the latent diffusion model. As shown in Fig. 2a, the text tokenizer converts the prompt strings into tokens, the indices in a pre-specified vocabulary of size $|V|$. Each token is then mapped to its corresponding embedding vector $\mathbf{e}_i \in \mathbf{E}$ where $i \in \{1, \ldots, |V|\}$ in the embedding space $\mathcal{E}$. The embedding vectors of the tokens in an input prompt are combined with the positional embeddings which are input to the transformer layer to get the encoded representation $f(\mathbf{S})$.

Prior work on textual inversion in the embedding space $\mathcal{E}$ [14, 48] aims to learn new concepts for a set of images by introducing placeholder strings that are associated with a newly learned embedding vector. This entails extending the existing vocabulary with new concepts. These new concepts have, however, been found to be un-interpretable when read as natural language because the learned embeddings are typically far away from the embeddings of the original vocabulary. In our approach, on the other hand, given the target image, we aim to optimize the embedding vectors $\mathbf{e}_i$, $i \in \{1, \ldots, L\}$ from the existing vocabulary for a prompt with maximum length L. This yields prompts that are human-readable. Concurrent work in this direction optimizes the hard prompts with CLIP similarity [49] of the image and text encoding from the image and text CLIP encoders [33]. In contrast, our approach is agnostic to the choice of encoders and depends entirely on the latent diffusion backbone utilized to perform prompt inversion.

3.2. Prompting Hard or Hardly Prompting

Given a target image $\mathbf{I}$ and the latent diffusion model parameterized by ϵ_θ, we optimize for the tokens in the existing vocabulary of the conditioning text encoder (CLIP text encoding) that best represent the visual content of the image. To learn the optimal tokens $\mathbf{S}^*$ that are likely to generate the target image, we use the following formulation:

$$\mathbf{S}^* = \arg\min_{\mathbf{S}} \mathbb{E}_{t,\mathbf{x},\epsilon}[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, f(\mathbf{S}))\|_2^2]. \tag{2}$$

We aim to recover the tokens $\mathbf{S}^*$ which would satisfy the vocabulary of the conditional diffusion model. Note that we keep the parameters of the text encoder and the diffusion model frozen. Our approach, therefore, takes into account the language prior introduced in the generative model.

A naive approach to solving Eq. (2) would be to optimize for all diffusion time steps, with the typical SGD optimizer. This, however, does not work in practice as we will show in Sec. 4. We thus introduce two key insights: (1) we focus on time steps that actually matter; (2) we keep our solutions *strictly* within the vocabulary space that the model was trained with through projected gradient descent.

Timesteps for semantic information. To enable efficient and effective optimization, we investigate the impact conditioning has within a pre-trained diffusion model. Figure 2b shows the diffusion loss of different prompts for a given image at different timesteps of the forward diffusion process. We observe that the conditioning signal for the diffusion model is stronger at the later, more noisy timesteps of the diffusion process. Conversely, the values of the diffusion loss are similar in the initial timesteps irrespective of the prompt used for the given image.

More specifically, consider the image (Fig. 2a), which was generated with stable diffusion using the prompt "A cabin on a snowy mountain in the style of Disney". In Fig. 2b a random prompt "Two dogs playing" has an objective value very close to that of the original prompt in the initial timesteps up to ~ 400. The values of the objective diverge only for the later timesteps. Moreover, when we provide random prompts with more information representative of the content of the image as well as the original prompt, the difference between the objective values of the random and the original prompts decreases. This hints that the high-level semantic information is encoded in the later timesteps while the low-level details are present in the initial timesteps of the forward diffusion process.

Furthermore, in Fig. 3, we show the effect of optimizing a prompt from different timestep ranges. We observe that as we increase the range of timesteps to be optimized from more noise $t = T$ to less noise $T = 0$, the semantic information gradually increases up to the middle range of timesteps ~ 500. For the larger ranges including the initial timesteps of the forward diffusion, no new semantic details are recovered from the target image.

We use these key observations in our optimization refined *prompting hard or hardly prompting (PH2P)* algorithm (Algorithm 1) where we limit the range of t to the later timesteps ≥ 500. Optimizing only for the later timesteps has two-fold advantage. First, since we are optimizing for the initial layers of a very deep neural network, the gradients are very small. By optimizing for larger values of loss, we get better gradients and therefore, a better direction for descent in the high-dimensional space. Secondly, we empirically observe that the tokens in the generated prompts, when optimized for the earlier timesteps, do not add any new conditioning information for the prompt optimization and may contain special characters; see Fig. 3. Thus, optimizing in the noisy range of the diffusion process yields tokens that are more representative of the content.

Projected gradient descent for meaningful prompts. Following our observations, we choose the starting timestep $t_a < T$ and optimize Eq. (2) with $t \in [t_a, T]$ using projected gradient descent with delayed projections [7, 25, 45] in our PH2P prompt inversion for text-to-image diffusion models as shown in Algorithm 1. Starting with a random prompt with embedding $\hat{\mathbf{e}}$, the loss is computed with respect to the embeddings in the feasible set *i.e.* the embeddings representing the vocabulary. During optimization, the variable $\hat{\mathbf{e}}$ is updated without the projection. This improves the efficiency of the descent. Concurrent work builds upon a similar theory of delayed projected descent and applies the projection for CLIP similarity matching [49] with standard gradient descent. However, when using a diffusion model, image encoders are not accessible and hence one cannot use projection with the image-text similarity loss. In our work, at each iteration, the update is computed with respect to the projections of the updated embeddings using the L-BFGS [42] algorithm. The variable updates performed in this manner provide for better convergence compared to projected descent with Adam optimizer; see Tab. 4.

Algorithm 1: PH2P Prompt Inversion

```
Input: Diffusion model parameters: $\theta$, Target image:
  $\mathbf{x} = E(\mathbf{I})$, Initial prompt: $\mathbf{S}$, Prompt embedding:
  $\hat{\mathbf{e}}$, Timesteps: $[t_a, T]$; Learning rate: $\lambda$,
  Optimization steps: $N$
for $i \leftarrow 1$ to $N$ do
    /* Projection on feasible set */
    $\tilde{\mathbf{e}} = \mathrm{Proj}_{\mathbf{E}}(\hat{\mathbf{e}})$
    /* Select diffusion timestep */
    $t = \mathrm{random}([t_a, T])$
    /* Apply L-BFGS to Eq. (2) */
    $g = \mathrm{LBFGS}_{\tilde{\mathbf{e}}}(\mathcal{L}_{LDM}(\mathbf{x}_t, \theta, t, f(\tilde{\mathbf{e}}))$
    $\hat{\mathbf{e}} = \hat{\mathbf{e}} - \lambda g$
end
/* Delayed projection */
return $\mathrm{Proj}_{\mathbf{E}}(\hat{\mathbf{e}})$
```

4. Experiments

We empirically demonstrate our prompt inversion procedure on the COCO [26] and the SUN [51] datasets. We validate our results on 500 images from the validation sets across 5 seeds and use Stable Diffusion v1.5 as the pretrained diffusion model.

Evaluation metrics. We consider the following three aspects when evaluating the quality of inverted prompts:

- Accuracy of the inverted prompts: we measure the CLIP [33] and the LPIPS [55] image similarity scores to measure the semantic similarity between the target and the generated images.
- Diversity of the generated images: we measure the diversity of the generated images using LPIPS between differ-

Method	COCO			LSUN		
	CLIP Similarity(↑)	LPIPS Similarity (↓)	LPIPS Diversity (↑)	CLIP Similarity(↑)	LPIPS Similarity (↓)	LPIPS Diversity (↑)
PEZ [49]	0.72	0.477	0.417	0.70	0.480	0.420
PH2P(Ours)	**0.77**	**0.462**	**0.435**	**0.77**	**0.463**	**0.422**

Table 1. **Quantitative Evaluation.** Evaluation of the quality of the images generated with the prompts from inversion.

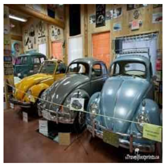
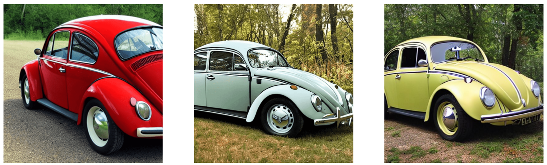
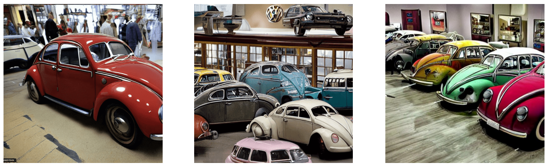
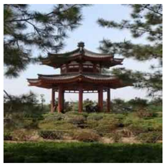
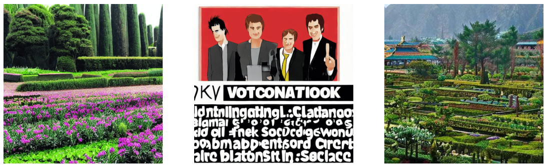
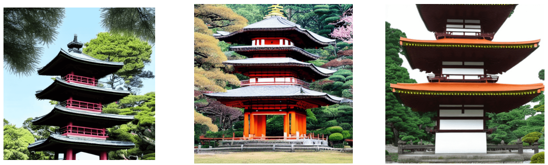
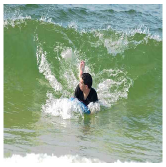
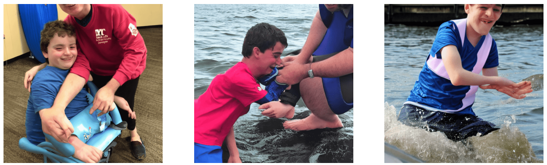
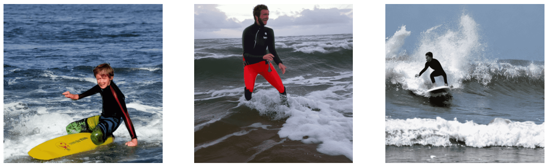
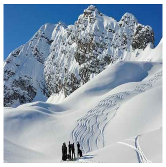
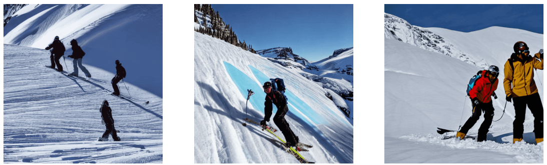
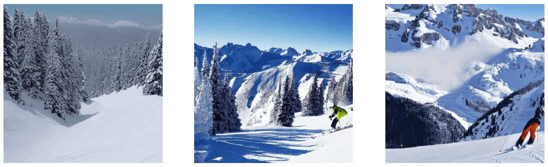
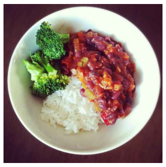
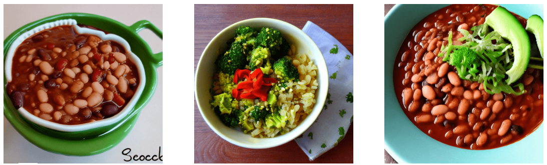
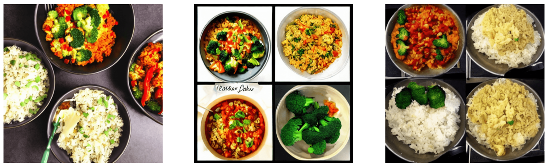
Target Image	PEZ [49]	PH2P (Ours)
	 visited retro vw ptic unpopular beetle isn ldnont evie ”@ oudiscovery-destinations dyk ote avalley	 pressed shops converted volkswagen rat headlights seized rescued danube smithsonian museum dose tycoon outdated
	 xian horticulitinerary seis ,@ #@ yu :-rillagamification recommend blancvisubrook	 pavilion pagoda depicted japanese resources bearing hilltop bearing mainland tang
	 rts disability featured ,@ nyc autism / kidpatrick sinai kf enjoys grasp hugging otw waves	 toby enjoying rough navigate surfing surf wave implications
	 bbloventureaktail vows skipped glacier corrie you ¡¡¡ pow tour guided vinci agronhikes	 alps tux beautiful skiing else S after resort
	 ♡blogged clegg pulls paleo jaredbroccoli bres ’? ” beans protein omo chili の	 homemade cauliflower chilli meals broccoli ashi rice

Table 2. **Qualitative Comparison**. Target image, inverted prompts (from PEZ [49] and PH2P), and corresponding generated images.

ent generated samples.

- Interpretability of the inverted prompts: we use the BertScore [56], which correlates well with human judgment. We measure the semantic equivalence between the embeddings of two sentences by taking into account the context of the tokens.

4.1. Evaluation of the Inverted Prompts

Image generation with inverted prompts. We assess the effectiveness of the prompts generated using our PH2P inversion for image generation. Given a target image, we first perform inversion to retrieve the prompt and evaluate the relevance of the images generated with the recovered prompt using the diffusion model. We compare our approach to the prompts generated using (concurrent/unpublished) PEZ [49] as a baseline. As shown in Tab. 1, the images generated with our PH2P prompts outperform the PEZ prompts in terms of CLIP similarity by an absolute value of 5 percentage points and 7 percentage points on the COCO and SUN datasets, respectively. Furthermore, our PH2P prompts display better performance in terms of the LPIPS similarity. This demonstrates the accuracy and relevance of the PH2P prompts for generating

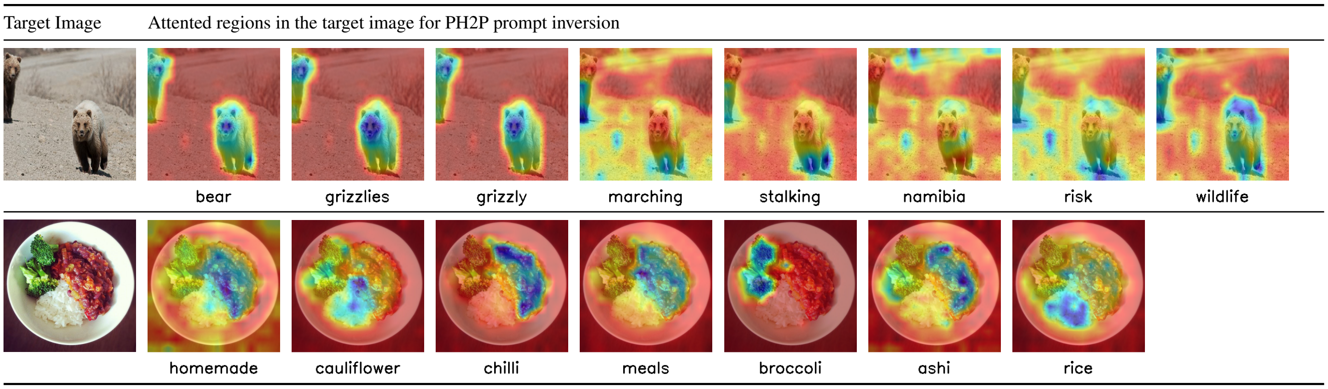


Figure 4. **Application of Unsupervised Segmentation.** Illustration of the tokens and corresponding regions (that can be used for unsupervised segmentation; see [50]) obtained for the target images. Note the accuracy of both prompts and corresponding attention.

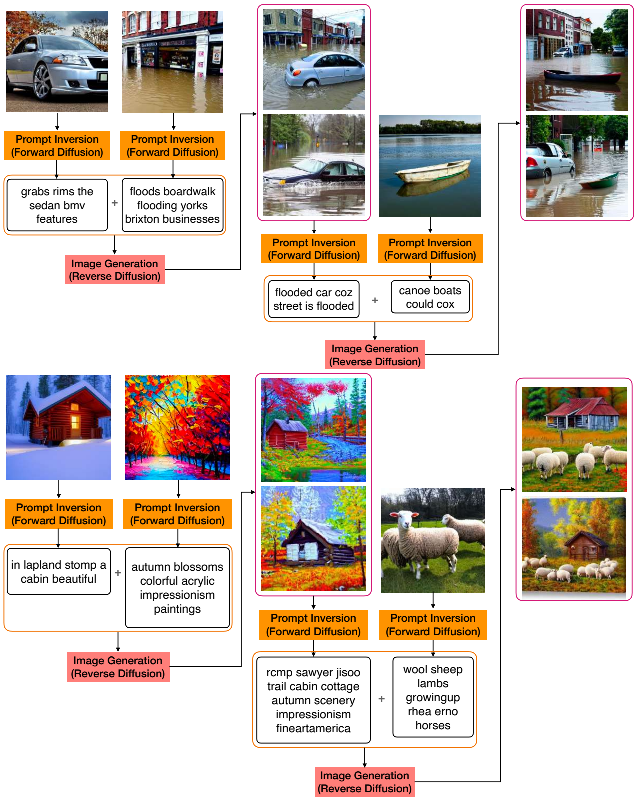

Prompt Inversion (Forward Diffusion)
Prompt Inversion (Forward Diffusion)
grabs rims the sedan bmv features
+
floods boardwalk flooding yorks brixton businesses
Image Generation (Reverse Diffusion)
Prompt Inversion (Forward Diffusion)
Prompt Inversion (Forward Diffusion)
flooded car coz street is flooded
+
canoe boats could cox
Image Generation (Reverse Diffusion)
Prompt Inversion (Forward Diffusion)
Prompt Inversion (Forward Diffusion)
in lapland stomp a cabin beautiful
+
autumn blossoms colorful acrylic impressionism paintings
Image Generation (Reverse Diffusion)
Prompt Inversion (Forward Diffusion)
Prompt Inversion (Forward Diffusion)
rcmp sawyer jisoo trail cabin cottage autumn scenery impressionism fineartamerica
+
wool sheep lambs growingup rhea erno horses
Image Generation (Reverse Diffusion)

Figure 5. **Application of Evolutionary Multi-concept Generation with Proposed PH2P.** Images generated with the composed prompts are inverted with PH2P to create prompts which are further combined with prompts recovered from the new target image.

Method	Precision(↑)	Recall (↑)	F1 (↑)
PEZ [49]	0.772	0.835	0.802
PH2P (Ours)	**0.803**	**0.838**	**0.820**

Table 3. **Quantitative Evaluation of Prompt Quality.** Evaluation of quality of prompts generated on the COCO dataset as captured bu the BertScore [56]. We compare the similarity between the inverted prompts and the ground-truth captions.

Method	CLIP Similarity(↑)	LPIPS Similarity (↓)	LPIPS Diversity (↑)
PH2P	**0.77**	**0.462**	**0.435**
LDM+adam	0.65	0.501	0.400
LDM+all t	0.72	0.479	0.423

Table 4. **Ablation.** Ablations study the significance of optimization choices for inversion in PH2P.

images with visual concepts of intent.

When comparing the diversity of the images generated with our PH2P versus PEZ [49], our PH2P approach yields higher diversity while maintaining the image semantics. Our results highlight that our approach recovers the prompts in the space of the model's vocabulary; these prompts can be used to sample diverse yet relevant images with visual concepts aligned with those of the target image.

Extensive qualitative results in Tab. 2 show that our approach generates prompts representative of the visual content in the target images. Unlike PEZ [49], where the generated images, at times, cannot capture the intended target concepts (*e.g.*, row 2), our approach consistently generates images with accurate semantics.

Quality of the prompts. We compare the contextual similarity between the prompts inverted for the COCO dataset and the ground-truth annotations (captions) of the corresponding images, with the BertScore [56]. Results in Tab. 3 show that our approach outperforms the prompts generated by PEZ [49] in terms of BertScore, especially with respect to the precision evaluation validating that our PH2P prompts have greater semantic similarity to the human captions. Our PH2P provides relatively precise prompts compared to those from PEZ and has better contextual similarity to the ground-truth captions. We observe from Tab. 2 that the prompts from our PH2P approach are more crisp and clear compared to those from PEZ [49]. The prompts generated with PEZ tend to have a high frequency of special or uninterpretable characters; see rows 2, 4, 6 in Tab. 2.

Ablations. To justify the optimization choices for PH2P prompt inversion in diffusion models, we show in Tab. 4 the performance of prompt inversion with Adam optimizer and with optimization for all timesteps (as opposed to the selected range). When using Adam optimizer, the approach yields prompts that generate images with relatively low

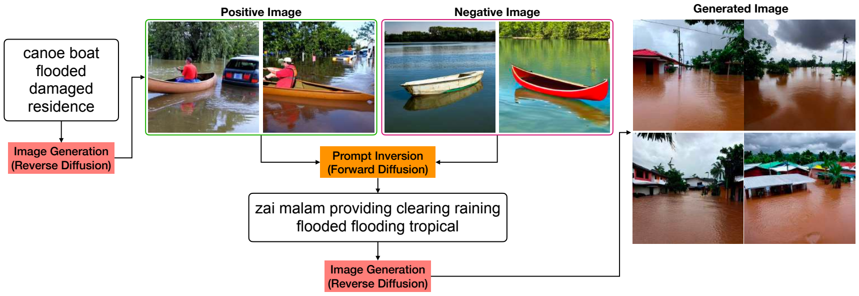

Positive Image
Negative Image
Generated Image
canoe boat flooded damaged residence
Image Generation (Reverse Diffusion)
Prompt Inversion (Forward Diffusion)
zai malam providing clearing raining flooded flooding tropical
Image Generation (Reverse Diffusion)

Figure 6. **Application of Concept Removal with Negative Target Images.** PH2P yields a prompt that removes the visual concept given in the negative image from the positive target image. The PH2P prompts can be used to generate diverse images with removed concepts.

CLIP and LPIPS similarity. The Adam optimizer with a fixed learning rate often does not converge to a good solution. Similarly, when optimizing with all the timesteps even though, the CLIP similarity of the generated images with recovered prompts is comparable to that of PEZ, due to high variance in the trajectory the results lag behind our proposed PH2P procedure.

4.2. Applications of Prompt Inversion

In this section, we show the benefits of prompt inversion in evolutionary multi-concept synthesis and in concept removal via negative image prompting.

Evolutionary multi-concept image synthesis. Consider an application where the goal of the user is to synthesize a mentally constructed image (*e.g*., *an image of a flooded city*). A convenient way to do so, would be to select some sample images (*e.g*., obtained from Google image search) that contain concepts he/she wants to see (*e.g*., *car*, *flooded street*). These images can be used to prompt the text-to-image generative model to generate desired illustration. Once the target images are generated, a user may realize that inserting a boat may give a more striking impact and would want to incorporate that concept. Such creative process is fundamentally enabled by our prompt inversion. Consider the two examples in Fig. 5. In the first step the concepts of two (or more) images can be composed by performing a prompt inversion of each image (set) and subsequently concatenating the prompts of the two (sets) of images. In step two, the newly composed images generated with the concatenated prompt can further be inverted to get a precise prompt that takes into account the text prior within the diffusion models. The inverted prompt in step 2 can further be combined with the inverted prompt of a new image depicting an additional concept. This allows convenient multi-concept composition (also enabling user prompt editing).

Negative image prompting. Our PH2P algorithm allows for the removal of concepts from images through negative image prompting; see Fig. 6. Given an input prompt, a set of positive images is generated representing the semantics of the input prompt. To remove a visual concept from the positive images, a negative target image with the negative visual concept is presented. To do this, we adapt our PH2P procedure and include negative gradients for the negative target image, generating prompts that drive towards the concepts in positive images and at the same away from the concepts in the negative target image; see supplemental for details.

Unsupervised Segmentation. Another application of prompt inversion is to generate results for unsupervised semantic segmentation (see [50]) where the cross-attention maps are used to generate segmentation masks. With our approach, one can generate prompts for the representative concepts directly from the diffusion model without any external information. For a given target image, we first perform PH2P inversion to generate such prompts. We visualize in Fig. 4 the cross-attention maps between the target image and the generated prompts [5]. Clearly, the tokens reflect the concepts relevant to the image, accurately attend to corresponding image, and can generate segmentations.

5. Conclusion

In this paper, we designed a simple but effective procedure for prompt inversion in text-to-image diffusion models. The prompts generated with our approach are readable, crisp, and importantly, representative of the content of the target image. We demonstrate the effectiveness of our inverted prompts for diverse image generation, evolutionary concept generation, and even concept removal. We believe that our inversion procedure would make prompting diffusion models much easier by allowing users to construct (and edit) effective textual descriptions of concepts they desire.

Acknowledgement. This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, NSERC Collaborative Research and Development Grant. Resources used in this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, the Digital Research Alliance of Canada alliance.can.ca, companies sponsoring the Vector Institute, and Advanced Research Computing at the University of British Columbia. Additional hardware support was provided by John R. Evans Leaders Fund CFI grant and Compute Canada under the Resource Allocation Competition award.

References

[1] Apratim Bhattacharyya, Shweta Mahajan, Mario Fritz, Bernt Schiele, and Stefan Roth. Normalizing flows with multi-scale autoregressive priors. In *CVPR*, 2020. 2

[2] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *ICML*, 2018. 2, 3

[3] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *NeurIPS*, 2023. 3

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 2023. 8

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 3

[7] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, 2015. 5

[8] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE Trans. Neural Networks Learn. Syst.*, 2019. 2, 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2

[10] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[11] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021. 2

[12] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In *ICLR Workshop*, 2015. 2

[13] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 2

[14] 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 *ICLR*, 2023. 2, 3, 4

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[16] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code GAN prior. In *CVPR*, 2020. 1, 2, 3

[17] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip H. S. Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 2

[18] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2

[19] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip H. S. Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 2

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[22] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *CVPR*, 2023. 1, 2

[23] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014. 2

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 1, 2

[25] Xiang Li and Zhihua Zhang. Delayed projection techniques for linearly constrained problems: Convergence rates, acceleration, and applications. *arXiv preprint arXiv:2101.01505*, 2021. 5

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 5

[27] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 1

[28] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *ICLR*, 2020. 2

[29] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 3

[30] Minheng Ni, Zitong Huang, Kailai Feng, and Wangmeng Zuo. Imaginarynet: Learning object detectors without real images and annotations. In *ICLR*, 2023. 2

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 1, 2

[32] Zhihong Pan, Xin Zhou, and Hao Tian. Arbitrary style guidance for enhanced diffusion-based text-to-image generation. In *WACV*, 2023. 1, 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4, 5

[34] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023. 1, 2

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1, 2

[36] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 2

[37] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2, 3

[40] 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 *CVPR*, 2023. 2, 3

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[42] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24 (111), 1970. 2, 5

[43] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3

[45] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates. *JMLR*, 2020. 5

[46] Hung-Yu Tseng, Hsin-Ying Lee, Lu Jiang, Ming-Hsuan Yang, and Weilong Yang. Retrievegan: Image synthesis via differentiable patch retrieval. In *ECCV*, 2020. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[48] 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. 2, 3, 4

[49] 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. *arXiv preprint arXiv:2302.03668*, 2023. 1, 2, 3, 4, 5, 6, 7

[50] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 1, 2, 7, 8

[51] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5

[52] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. In *CVPR*, 2019. 2

[53] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 2

[54] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[55] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[56] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020. 6, 7

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 1

[58] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 2, 3

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019. 2