

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

Shweta Mahajan^{1,2} Tanzila Rahman^{1,2} Kwang Moo Yi¹ Leonid Sigal^{1,2}

¹University of British Columbia

²Vector Institute for AI

{s.mahajan, trahman8, kmyi, lsigal}@cs.ubc.ca

Overview

In the following, we outline the algorithm for the negative image prompting for prompt inversion and also provide additional qualitative results to show that the prompts generated with our PH2P approach are precise, readable, and can be used for various downstream applications.

1. Negative Prompting

In Algorithm 2, we show the adaptation of our PH2P Algorithm 1 to get prompts that contain the concepts from a target image and at the same time do not contain concepts specified in the negative image. The images thus generated with the optimized prompt contain concepts in \mathbf{I} excluding or removing the concepts present in \mathbf{I}^{neg} . We find the algorithm to perform well when removing concepts in composed images (Figs. 6 and 9).

Algorithm 2: PH2P Prompt Inversion with Negative Image Prompting

```
1 Input: Diffusion model parameters:  $\theta$ , Target image:  $\mathbf{x} = E(\mathbf{I})$ , Negative
  image:  $\mathbf{x}^{\text{neg}} = E(\mathbf{I}^{\text{neg}})$ , Initial prompt:  $\mathbf{S}$ , Prompt embedding:  $\hat{\mathbf{e}}$ ,
  Timesteps:  $[t_a, T]$ ; Learning rate:  $\lambda$ , Optimization steps:  $N$ 
2 for  $i \leftarrow 1$  to  $N$  do
3   /* Projection on feasible set */
4    $\hat{\mathbf{e}} = \text{Proj}_{\mathbf{E}}(\hat{\mathbf{e}})$ 
5   /* Select diffusion timestep */
6    $t = \text{random}([t_a, T])$ 
7   /* Apply L-BFGS */
8    $g =$ 
9      $\text{LBFGS}_{\hat{\mathbf{e}}}(\mathcal{L}_{LDM}(\mathbf{x}_t, \theta, t, f(\hat{\mathbf{e}})) - \mathcal{L}_{LDM}(\mathbf{x}_t^{\text{neg}}, \theta, t, f(\hat{\mathbf{e}})))$ 
10   $\hat{\mathbf{e}} = \hat{\mathbf{e}} - \lambda g$ 
11 end
12 /* Delayed projection */
13 return  $\text{Proj}_{\mathbf{E}}(\hat{\mathbf{e}})$ 
```

2. Additional Qualitative Results

In Tab. 6 we provide additional qualitative results comparing the quality of the prompts from our PH2P approach with

the PEZ approach based on CLIP similarity. Here, we observe consistently better prompts. Images generated with our inverted prompts reflect the concepts of target image.

Additionally, we show in Tab. 7 the prompts generated with the standard Adam optimization (LDM+Adam) and also consider the setting with all the timesteps for inversion (LDM+ all t). We observe that the prompts generated with our PH2P procedure are consistent with the concepts in the target image. The prompts with LDM+Adam are short but do not always contain the target concepts. This shows the poor convergence of the standard SGD methods for hard prompt inversion. The prompts obtained with LDM+ all steps are longer than our PH2P procedure and tend to include words that do not provide any information on the concepts in an image. Our PH2P prompt inversion, on the other hand, provides prompts that are readable, precise, and consistent with the content of the image.

Similarly, in Fig. 7 we show the regions in the target image attended to by the optimized prompts which can be leveraged for downstream tasks such as unsupervised semantic segmentation as proposed in Wu *et al.* [49].

We also include additional examples for the application of the PH2P procedure to multi-concept generation Fig. 8. Our approach can faithfully compose images with diverse concepts. Figure 9 shows example cases where concepts can be removed from a set of target images using our PH2P procedure. Given the negative images of the concepts to be removed, PH2P with Algorithm 2 can be used to yield a prompt that removes concepts from target (positive) images. These observations demonstrate that the quality of the prompts generated with the PH2P procedure is representative of the vocabulary of the LDM [37] backbone.

3. Different text conditioning modules

The application of PEZ [48] with its CLIP-based loss is limited to the CLIP-based text encoder. Our PH2P approach generalizes to any text conditioning module within



Target Image	Conditioning in text-to-image diffusion model		
	CLIP [37]	BERT [37]	T5 [40]
	In lapland stomp a cabin beautiful	apartment resort house rent condo	alpine cute pine forests with chalet
	homemade cauliflower chilli meals broccoli ashi rice	learning make mixing salad	rice tonight broccoli ham in lovely bento

Table 5. **PH2P inversion with different text-to-image diffusion models.** Prompt inversion with different conditioning modules yields prompts with different levels of semantic information and fluency.

the framework of text-to-image diffusion models. As shown in Tab. 5, with our PH2P approach, prompts can be generated from the BERT-based latent diffusion model and T5-based Deepfloyd [40]¹. Notably, different text encoders yield prompts with varying fluency and semantic information. While T5 generates fluent, high-quality prompts, prompts from LDM with BERT conditioning module have limited vocabulary.

¹Implementation from <https://github.com/deep-floyd/IF.git>

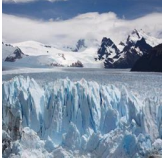













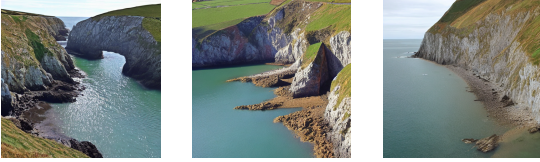

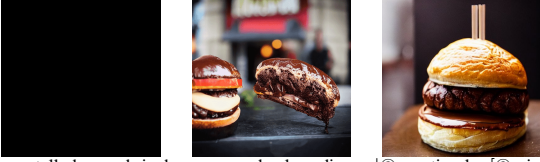

Target Image	PEZ [48]	PH2P (Ours)
	 <p>collaborate dreamliner shared ingrakyleargentina lake ices @ { } amadiscover tos — carmenono augmente</p>	 <p>glacier flows forgot milford idyllic glaciers milky forests</p>
	 <p>permitted facilities cats dryer drexyulhometems veterinlished gyn ext- ensive specializes scrubs mnt treated mnt</p>	 <p>tokyo ips seals dentistry simulation surgeries</p>
	 <p>suffolkworkforce shared editor , agron app quotehomegrown ginger biscuits recipe lovely</p>	 <p>length bread wheat cookies round moist worthless perhaps</p>
	 <p>renovated penthouse clubhouse — welch grand wisconsinindianfoot- ball renovated — blueentv occupoccup</p>	 <p>plantation clubhouse proposed room wanting chamber renovated</p>
	 <p>discoverkeywords temps monasterdiscover caverirrational apostles chose marqtours gaubois september coastline</p>	 <p>fitz caves conwy coast showing rt margaret vigilant margaret gilli cliffs</p>
	 <p>nutella burger brioche consumed solves dis @ wasting kx @ vin johnny wimpoaching imagery</p>	 <p>sometimes pancakes itis :) ” petr soho</p>

Table 6. **Qualitative Comparison.** Target image, inverted prompts (from PEZ [48] and PH2P), and corresponding generated images.


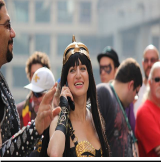


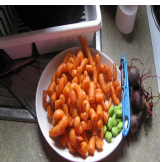
Target Image	LDM+Adam	LDM+ all t	PH2P (Ours)
	watched flying supremacy baptiste scho mueller at	airplanes sean moulin chuck claw peak intimidation intimidation	airplane photography skies myrtle striking farmer sometimes
	smokers concern criminal snippets geneva	rinking lexa cybermonday pilgrimage vendor apparently glimpse pend ctar iterate	visitors eyes hacked ' protesters reflects metoo
	A : coventry motorcycles vina championship	sbk rbs triumph coventry sprints slalom bhar classiccar ingu	badgers motorcycle driving racers championship
	skater ffi actions fitt osc looking *	rotation " reversed skateboarding sitter skateboarding skill handshake upside	michal illusion precision skate sitter skateboarding routine
	:) dinner knows dayz examples vegetables dunno	shaped lower diameter saves tying carrots use	carrots picked leftover wid mara lls munchies

Table 7. **Qualitative Comparison to Ablations.** Qualitative comparison of the prompts inverted and the diverse images generated with our PH2P to the baselines for ablation.

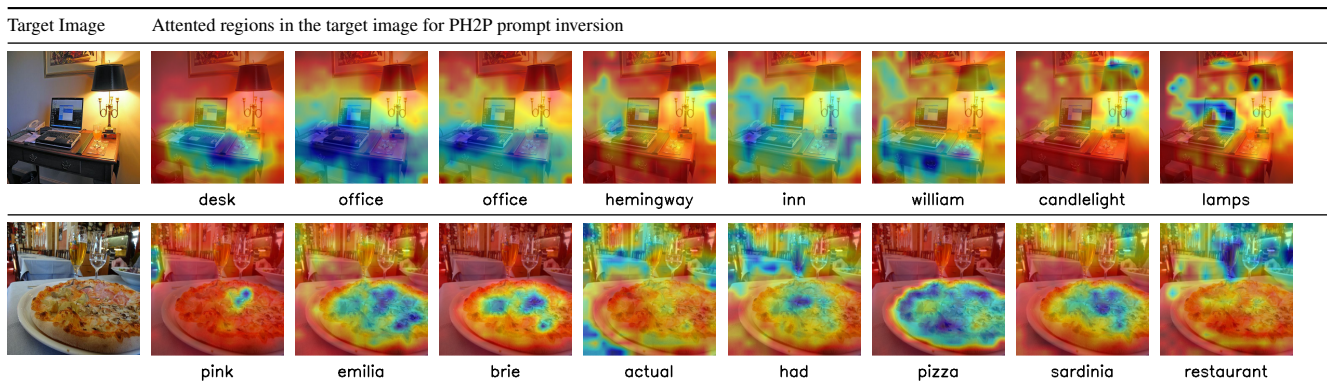


Figure 7. **Application of Unsupervised Segmentation.** Additional results on illustration of the tokens and corresponding regions (that can be used for unsupervised segmentation; see [49]) obtained for the target images. Note the accuracy of both prompts and corresponding attention.



Figure 8. **Application of Evolutionary Multi-concept Generation with Proposed PH2P.** Additional results on 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.

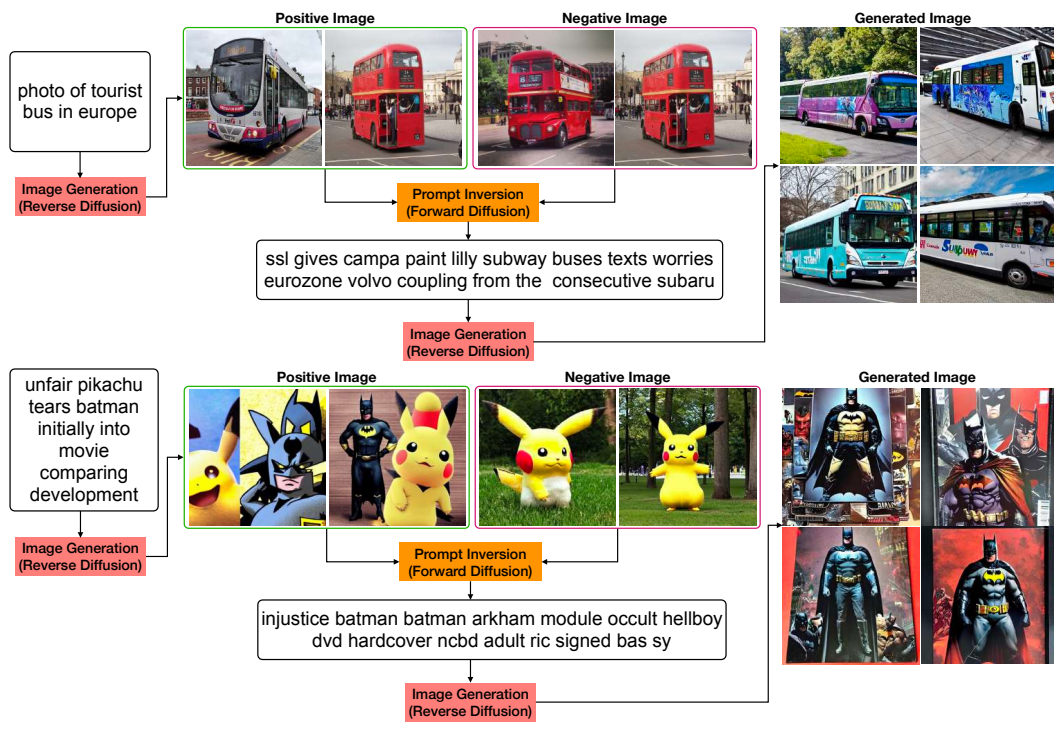


Figure 9. **Application of Concept Removal with Negative Target Images.** Additional results showing that PH2P yields a prompt that removes the visual concept given in the negative image from the positive target image. The PH2P prompts can be used to generate diverse images with removed concepts.