Conditional Gradient-based Textual Inversion

Xi Wang, Vicky Kalogeiton

Conditional Gradient-based Textual Inversion

Xi Wang and Vicky Kalogeiton

LIX, CNRS, Ecole Polytechnique, IP Paris
{xi.wang,vicky.kalogeiton}@polytechnique.edu

Abstract. Generative models excel in image generation but often require trail-and-errors for specific concepts. Textual inversion offers a solution; yet, is computationally costly. We propose using conditional gradient data to select or sample informative timesteps for textual inversion. Our methods improve computational cost and generation quality.

1 Introduction

Generative models have shown remarkable capabilities for synthesizing diverse and high-quality images, typically adherent to complex conditioning. However, users often require laborious test-and-trial processes to generate specific concepts, especially if these are hard to describe by words, such as a particular reference object or a personalized abstract concept. Textual inversion offers a solution by optimizing the textual embedding (condition latent) across all timesteps. [3] was among the first to address this through an optimization approach, where the unknown embedding in the text space is learnt from a set of examples. This inspired several follow-up works, such as editing, customization, or style transfer [8,9,18]. While these methods address personalized generation, their primary drawback is the significant amount of time required for optimization.

To address this, we propose a method that tackles personalized generation while simultaneously reducing the optimization time. Our intuition is that not all timesteps in the generating process contribute equally. Recently, [1, 17] show that the initial timesteps are primarily involved in creating an image from noise, the middle stages handle content creation, while the final stages focus on denoising, which is often less announced by the conditioning. This suggests that each timestep may have a different contribution to the inversion performance. In this work, we first measure the behaviour of different timesteps using the gradient norm of the condition. Then, we use this information to either pre-select some highly influential timestep intervals for inversion or to sample timesteps adaptively according to the grad norm during the inversion. Our experiments show that our proposed gradient-based textual inversion is more effective, achieving competitive or superior inversion performance with fewer optimization steps.

Our contributions are: (1) We propose a novel method to measure the conditioned behaviour at different timesteps by collecting the grad norm of guidance and show that each timestep corresponds to the different inversion results; (2) We show that using the timesteps with the high grad norm (via adaptive weighted-sampling or predefined ranges, even one-timestep) leads to a high-quality generation while providing a smoother optimisation with lower computation.

2 Background and Related Work

Diffusion models aim to convert noise into a target data distribution [4, 15, 16]. Following DDPM [4], diffusion consists in training a network ϵ_θ to denoise a noisy input to recover the original data at different noise levels, driven by a noise scheduler. The goal is to recover x_0, the original datapoint from $x_t{=}\sqrt{\gamma(t)}x_0{+}\sqrt{1{-}\gamma(t)}\epsilon$, where $\gamma(t){\in}[0,1]$ is a noise scheduler function of the timestep t and applied to Gaussian noise $\epsilon{\sim}\mathcal{N}(0,1)$. ϵ_θ is then trained with:

$$L_{\text{simple}} = \mathbb{E}_{x_0\sim p_{\text{data}},\epsilon\sim\mathcal{N}(0,1),t\sim\mathcal{U}[0,1]}\left[\|\epsilon_\theta(x_t) - \epsilon\|\right] \quad . \tag{1}$$

Once the network is trained, we can sample from p_{data} by setting $x_T{=}\epsilon \sim \mathcal{N}(0,1)$, and gradually denoising to reach the data point $x_0{\sim}p_{\text{data}}$.

Conditional Generation consists of conditioning the diffusion model. Classifier-free guidance (CFG) [5] applies conditioning without a pre-trained classifier [2] but by exploiting an implicit classifier. This is achieved by replacing $\epsilon_\theta(x_t)$ with $\epsilon_\theta(x_t, c)$ in Eq. 1 and omitting the label during training with a certain probability, resulting a single network and the mix of conditional and unconditional response to guide the generation process, controlled by the guidance weight ω:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t) + (\omega + 1) \cdot (\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t)) \quad . \tag{2}$$

In text-to-image generation, c is typically encoded by text encoders, e.g. CLIP [11]. Moreover, instead of executing the denoising on pixel space such as DDPM [4], recent works [10, 12] diffuse in a VAE latent space, a paradigm followed by the Stable Diffusion series and Midjourney.

Textual Inversion (TI) optimizes textual embeddings: conditioning parameter c_θ from prompt y for specific image pairs x. Similar to the loss in Eq 1, during this optimization process, timesteps and noise levels are sampled to compute the loss against the text embedding v. Typically, this embedding uses a placeholder or *abused* token to represent the target concept, such as "A photo of $<S^*>$".

$$v_* = \arg\min_v \mathbb{E}_{x\sim\epsilon(x),y\sim\mathcal{N}(0,1),t}\left[\|\epsilon_\theta(x_t, t, c_\theta(y)) - \epsilon\|\right] \quad , \tag{3}$$

with textual encoder $c_\theta(y)$ and diffusion network ϵ_θ frozen during optimisation. TI has inspired several customization, editing and style transfer methods [8, 18]. [9] optimize for editing transformations instead of specific instances. Beyond TI, Dreambooth [13] introduces fine-tuning with a class-specific prior preservation loss. Null-inversion [7] employs inversion for prompt-driven editing.

3 Our Few-step Adaptive Textual Inversion Method

Recent works [1, 17] suggest that diffusion timesteps serve distinct functions at different timesteps. For example, the initial stages ($t \sim 1000$) focus more on creating the image content from the sampled Gaussian noise, while the final stages ($t \sim 0$) primarily focus on denoising and completing high-frequency textures, which are often less relevant to the conditional information. To quantify

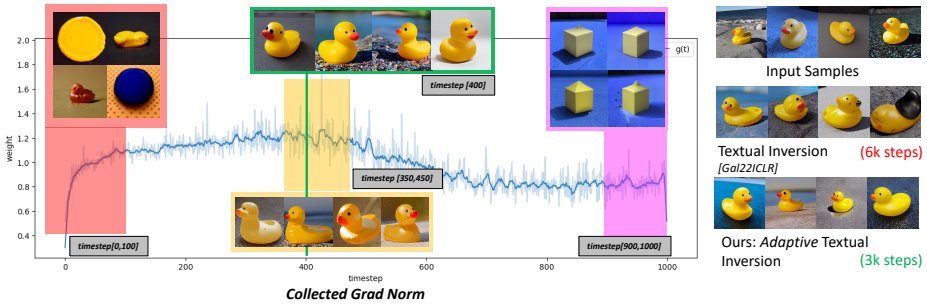


Fig. 1: (left) Post-optimised gradient norm ($g(t)$) over diffusion timesteps t of the tested T2I model. Different colored areas mark the inversion results from different timestep ranges: the red images on the left have been generated only using the $t = [0, 100]$. Interestingly, even a single timestep at a high $g(t)$ area ($t = 400$) can yield good results, while lower $g(t)$ areas, e.g. red and purple regions, result in irrelevant outcomes, showing that not all timesteps are equally critical for conditioning, thus leading to: (right) The proposed adaptive sampling of timesteps from the on-the-fly computed $g(t)$ data. Our adaptive method, compared to default textual inversion, shows faster convergence and better performance in terms of the image details and the similarity to the input samples.

and represent this relevance to the condition embedding, we propose to use the gradient norm $g(t) = \|\nabla_c \epsilon_\theta(x_t, c)\|$, as it quantifies the *influence* of the condition c on the generation process of $\epsilon(x_t, c)$ at timestep t. If the value of $\epsilon(x_t, c)$ remains consistent across different c values (*i.e.*, the gradient is zero), it may indicate that the condition does not significantly influence the generation at this timestep, thus suggesting a less reliable classifier $p(c \mid x_t)$.

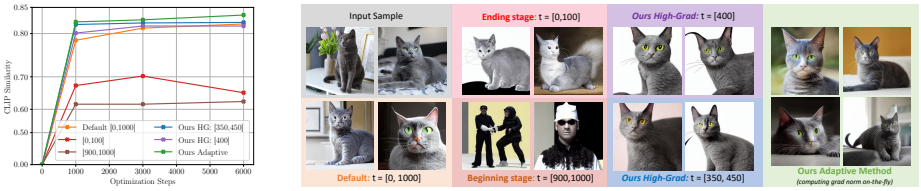
$$g(t) = \mathbb{E}_{t,x}[\|\nabla_c \|\epsilon(x_t, c) - \epsilon\|\|] \quad . \tag{4}$$

Importantly, $g(t)$ can be derived directly from the training loss through its expectation. In this work, we propose two strategies for computing $g(t)$.

(1) High-grad. Given a model trained with various images (e.g. COCO [6] test set), we first compute its grad norm during post-training optimization and then we empirically choose the timestep intervals (or even one timestep) with the highest grad norm values (yellow and green regions in Figure 1).

(2) Adaptive. We collect the gradient norm $g(t)$ *on-the-fly* while simultaneously performing textual inversion on input samples X^*. Timesteps are adaptively sampled according to $g(t)$, treated as a probability distribution where timesteps with higher $g(t)$ have a greater likelihood of being selected, as described in Eq. 5 where the PDF is the probability density function of the normalised $g(t)$:

$$t \sim PDF(g(t)) = PDF(\mathbb{E}_{t,x \in X^*}[\|\nabla_c \|\epsilon(x_t, c) - \epsilon\|\|]) \quad . \tag{5}$$



(a) CLIP similarity Results **(b)** Qualitative results

Fig. 2: (a) CLIP Similarity vs. Optimization Steps. Both our high-grad methods (blue, purple) and especially our adaptive one (green) outperform all other strategies in terms of CLIP similarity and convergence speed. **(b) Qualitative Results.** Our adaptive method produces images (green) with better details and higher similarity to the input samples (gray, top left corner) than other methods (brown, red).

4 Experiments

Setup. We use Dreambooth [13] dataset, comprising 30 subjects of animals and still objects. Each subject contains 4 to 6 images with ground truth prompts, captured under varying conditions. Following the original textual inversion [3], we use the LDM [2] pre-trained on LAION-400M [14] as our diffusion backbone.
Metrics. We measure the CLIP-based similarity between input samples and the regenerated images with inverted text embedding S^*, e.g.: "An image of S^*".
Baselines. We compare images obtained via our proposed adaptive sampling method ($g(t)$ is computed *on-the-fly*, as shown in Eq. 5) to images obtained through the default textual inversion [3] (i.e. from the standard interval $[0, 1000]$) and to images obtained via five baselines: (i) the beginning stage $[900, 1000]$, (ii) the ending stage $[0, 100]$, (iii) the critical grad norm interval $[350, 450]$ (see Figure 1 for collected grad norm curve), and (iv) a single timestep $[400]$.
Results. Figure 2a shows the CLIP similarity over optimisation steps for various methods. It confirms that both our *high-grad* (blue, purple) and *adaptive* (green) approaches converge faster compared to the baseline scheme $[0, 1000]$ (orange), confirming the selection of more informative timestep ranges by our proposed $g(t)$. Surprisingly, even a single timestep ($t = 400$, purple) can effectively achieve the inversion, further validating the informativeness of $g(t)$. Conversely, the beginning and ending intervals result in poor performances (red, brown), suggesting the conditioning is weak at these ranges, corroborating our argument. Adaptive (green) achieves the best performance and fastest convergence, benefiting from its overfitting to the input images and its adaptive sampling on more informative timesteps. Figures 1 and 2b depict examples obtained via different strategies.
Conclusion. We proposed exploiting conditional gradient norms to select timestep intervals or adaptively sample timesteps for textual inversion, significantly reducing optimization time while maintaining high-quality personalized generation. Future work includes generalizing to other domains, i.e., motion, and further confirming the informativeness and semantics of the single-step inversion for different concepts to help applications such as compositional generation.

5 Acknowledgement

This work was supported by ANR APATE ANR-22-CE39-0016, Hi!Paris grant and fellowship, and was granted access to the High-Performance Computing (HPC) resources of IDRIS under the allocations 2024-AD011014300R1 made by GENCI. We would like to thank Nicolas Dufour, David Picard, and the anonymous reviewers for their insightful comments and suggestions.

References

1. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 11472–11481 (2022)
2. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural Inform. Process. Syst. (NeurIPS) (2021)
3. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Int. Conf. Learn. Represent. (2022)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inform. Process. Syst. (NeurIPS) (2020)
5. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
6. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. (ECCV) (2014)
7. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2023)
8. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In: ACM Int. Conf. Multimedia (2023)
9. Nguyen, T., Li, Y., Ojha, U., Lee, Y.J.: Visual instruction inversion: image editing via visual prompting. In: Adv. Neural Inform. Process. Syst. (NeurIPS). pp. 9598–9613 (2023)
10. Nichol, A., Dhariwal, P.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
13. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (2023)
14. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)

15. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML). pp. 2256–2265. PMLR (2015)
16. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
17. Wang, X., Dufour, N., Andreou, N., Cani, M.P., Abrevaya, V.F., Picard, D., Kalogeiton, V.: Analysis of classifier-free guidance weight schedulers. arXiv preprint arXiv:2404.13040 (2024)
18. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR). pp. 10146–10156 (2023)