Unleashing Creativity: Generalizing Semantic Control for Text-to-Image Diffusion Models

Manuel Brack[1,2,5] **Marlon May**[1,2] **Linoy Tsaban**[4] **Felix Friedrich**[2,3]
Patrick Schramowski[1,2,3] **Apolinário Passos**[4] **Ajinkya Kale**[5] **Kristian Kersting**[1,2,3,6]
[1]German Research Center for Artificial Intelligence (DFKI),
[2]Computer Science Department, TU Darmstadt [3]Hessian.AI,
[4]Huggingface, [5]Adobe Applied Research, [6]Centre for Cognitive Science, TU Darmstadt
brack@cs.tu-darmstadt.de

Figure 1: SEGA facilitates versatile semantic image manipulation. Our method is compatible with any text-to-image diffusion architecture.

Abstract

The recent surge in popularity of text-to-image diffusion models (DMs) can largely be attributed to the versatile, expressive, and intuitive user interfaces provided through textual prompts. These models enable inexperienced people to explore artistic ventures easily and provide exciting new opportunities to experienced artists. However, the semantic control offered through text prompts alone is limited and rather fragile, and overall lacks the fine granularity necessary for creative applications. The majority of methods addressing this issue are restricted to specific DM architectures, severely limiting the creative workflow instead of generalizing it to arbitrary models. In contrast, we demonstrate that semantic guidance (SEGA) generalizes to any DM architecture. Importantly, SEGA is natively compatible with state-of-the-art diffusion transformers. Our empirical results show strong model-agnostic performance, and we highlight new creative possibilities enabled by SEGA, such as enhanced typographic manipulations. This work underscores SEGA's potential to provide consistent, high-quality semantic guidance in a rapidly evolving generative model landscape.

1 Introduction

A key aspect of the popularity of text-to-image diffusion models (DMs) [18, 15, 16] is the versatility, expressiveness, and—most importantly—the intuitive interface they provide to users. Naturally, DMs have been adopted for a variety of creative applications [2]. Many artists utilize generative models in their creative workflow [6, 8], but DMs have also lowered the barrier of entry for users without

Preprint. Under review.

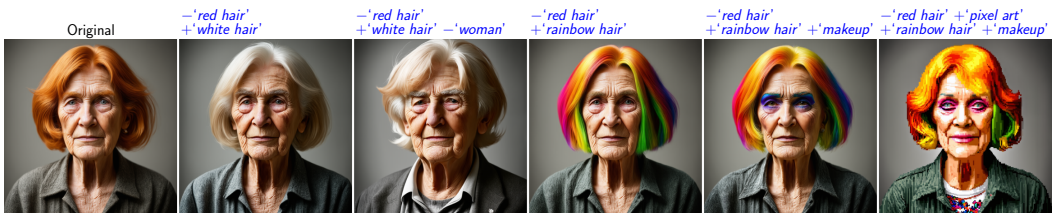


Figure 2: SEGA facilitates creative workflows through iterative instruction refinement. Users can separately manipulate various concepts until the desired output is achieved. Importantly, the results always remain grounded by the original image.

previous artistic expertise. The ability to express a generation's intent can easily be articulated in natural language, making it accessible to more people.

However, the models' intuitive text interface is also one of their largest weaknesses. Especially creative applications of DMs are decidedly iterative, where users make successive adjustments to the generated image until a satisfactory result is achieved. Unfortunately, the diffusion process is rather fragile, with even small changes to the input prompt resulting in completely different images. Consequently, artists and other users alike require more precise and fine-grained control over the iterative generation process.

Recently, various research works have tackled this exact problem to offer more semantic control in text-to-image generation [9, 19, 5, 12, 14, 3]. However, they mostly rely on the unique characteristics of specific DMs and are thus ill-suited for the fast-paced ecosystem of generative image models. Specifically, they are not applicable to current SOTA models like Stable Diffusion 3 [7]. In contrast, we require an approach that generalizes to any type of diffusion model and can be easily adapted for new releases. Such a workflow allows users to familiarize themselves with one creative pipeline that can be applied to any DM. That requirement is particularly relevant in the rapidly evolving landscape of generative models, which sees significant changes over mere weeks.

Recently introduced DMs particularly well highlight the need for such an approach. Many prominent techniques rely on cross-attention layers [9, 5] or U-Net architecture [12] of past models. However, current SOTA DMs [7, 17, 13] replace the U-Net with a diffusion transformer (DiT) and often have no cross-attention layers at all, rendering all of these methods inapplicable. Conversely, we argue that Semantic Guidance (SEGA) [3] has no inherent restrictions regarding input encoding, model architecture, or modalities. For the first time, we demonstrate the application of SEGA for DiTs, highlighting the generalizability of the approach. Specifically, we make the following contributions

- We supply implementations of SEGA for three recent DiT models[1]: SD-3, AurafLow & HunyuanDiT
- We showcase creative use cases with SEGA on these models. We demonstrate the inherent benefits of stronger base models, allowing for novel types of creative image manipulations such as typography.
- We show that semantic vectors' robustness, uniqueness, monotonicity, and isolation apply to DiT.
- We provide strong empirical results confirming the model-agnosticism of SEGA.

2 Creative Workflow

Let us consider what a typical creative workflow with a generative text-to-image model would look like. The user might create a prompt roughly describing the desired output. We would often consider different prompt samples from various seeds and choose the most suitable. That selection may be based on overall aesthetics, specific image composition, or other subjective factors. Importantly, it is unlikely that all details of this initial selection will satisfy a user completely. Instead, we would now like to manipulate certain aspects of the image while maintaining the overall look and composition, as demonstrated in Fig. 2. Unfortunately, the text-to-image models themselves do not support this

[1]Implementations available at `https://github.com/ml-research/diffusers`

Figure 3: Typography manipulation enabled by combining SEGA with recent DiT models. Users can change text and independently manipulate other aspects of the image and even correct spelling mistakes in the original output.

fine-grained level of semantic control. The diffusion process is sensitive to input variations, as minor adjustments to the input can result in drastically different visual outputs. These shortcomings could be tackled with techniques requiring segmentation masks, extensions to the architecture, model fine-tuning, or embedding optimization [1, 9, 11, 19]. However, they disrupt the fast, exploratory workflow that is the strong suit of diffusion models in the first place and sometimes require specialized knowledge, raising the barrier of entry for new users. Further, many of the techniques that do allow for easy, textual-driven semantic control are no longer applicable to current SOTA diffusion transformers [9, 19, 5, 12, 14].

In contrast, SEGA meets all of these requirements. It allows for flexible and intuitive semantic manipulations for any text-to-image architecture, including DiTs. Coming back to Fig. 2, we demonstrate how to use SEGA to quickly explore ideas and iteratively carve out the desired result. Users can easily target specific attributes of the generated image and change them at will. An unsatisfying change can easily be dropped in favor of a different one. Importantly, the entire process remains grounded on the originally generated image, eliminating random changes and leaving the user in control.

This level of control is enabled through the wide range of semantic manipulations supported by SEGA. We depict examples in Fig. 1 and this paper. The types of manipulations include object/attribute removal and addition, replacements, and global changes such as style. Importantly, SEGA's native multi-conditioning allows for arbitrary combinations of these edits, thus facilitating complex, multi-faceted manipulations. Additionally, the sophisticated control offered by SEGA's hyper-parameters enables the users to control the magnitude of any potential edits easily.

Further, SEGA inherits novel capabilities by using it with stronger base models. Specifically, recent DiTs can now reliably perform typography tasks, allowing for the generation of custom text in the output image. Consequently, SEGA can now manipulate text in images as shown in Fig. 1 and in more detail in Fig. 3. Typography manipulations are not only restricted to replacing similar texts, but SEGA also allows for complete rewriting while keeping the rest of the image. Additionally, text editing remains isolated from other aspects of the image and thus can easily be combined with other edit instructions. Further, we observed that SEGA may even be used to correct spelling mistakes that the model might make in the original generation (cf. Fig. 3 right). These new manipulation capabilities highlight one of the great advantages of SEGA's model-agnosticism. The method can be easily applied to any new DM and thus benefit from fast-paced improvements in image generation.

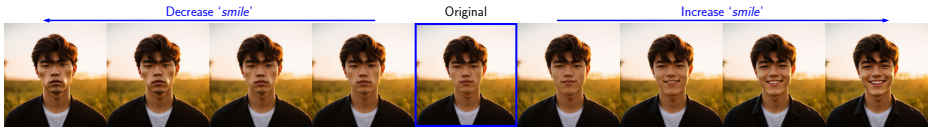
3 Properties of SEGA

Four key properties of semantic guidance are at the core of SEGA's versatile applications in image manipulation and other creative tasks. In the following, we demonstrate that all properties originally demonstrated for U-Net architectures [3] natively translate to DiT models. These results provide more structured evidence of SEGA's architecture-agnosticism.

Robustness. SEGA behaves robustly for incorporating arbitrary concepts into the original image. In Fig. 4a, we applied guidance for the concept '*glasses*' to images from different domains. Notably, this prompt does not provide any context on how to incorporate the glasses into the given image and thus leaves room for interpretation. The depicted examples showcase how SEGA extracts best-effort integration of the target concept into the original image that is semantically grounded. Consequently, SEGA is easy to use and allows for quick explorations of different manipulations without the need for user-drawn masks or other inputs.

(a) Robustness of guidance vectors. Results for guiding toward '*glasses*' in various domains without specifying how the concept should be incorporated.

(b) Uniqueness of guidance vectors. The vector for '*glasses*' is calculated **once** on the blue-marked image and subsequently applied to other prompts (without colored border).



(c) Monotonicity of guidance vectors. The guidance scale for '*smile*' is semantically reflected in the images.

Figure 4: Robustness, uniqueness and monotonicity of SEGA guidance vectors. In a) and b), the top row depicts the original image, and the bottom row depicts the ones guided towards '*glasses*'. (Best viewed in color)

Uniqueness. Guidance vectors γ of one concept are unique and can thus be calculated once and subsequently applied to other images. Fig. 4b shows an example for which we computed the semantic guidance for '*glasses*' on the left-most image and simply added the vector in the diffusion process of other prompts. All faces are generated wearing glasses without a respective ϵ-estimate required. Even significant domain shifts are covered, such as switching from photo-realism to drawings.

However, the transfer is limited to the same initial seed, as ϵ-estimates change significantly with diverging initial noise latents. Furthermore, more extensive changes to the image composition, such as the one from human faces to animals or inanimate objects, require a separate calculation of the guidance vector. Nonetheless, SEGA introduces no visible artifacts to the resulting images.

Monotonicity. The magnitude of a semantic concept in an image scales monotonically with the strength of the semantic guidance vector. In Fig. 4c, we can observe the effect of increasing the strength of semantic guidance s_e. Both for positive and negative guidance, the change in scale correlates with the strength of the smile or frown. Consequently, any changes to a generated image can be steered intuitively using only the semantic guidance scale s_e and warm-up period δ. This level of control over the generation process is also applicable to multiple concepts with arbitrary combinations of the desired strength of the edit per concept.

Isolation. Different concepts are largely isolated because each concept vector requires only a fraction of the total noise estimate. Meaning that different vectors do not interfere with each other. Thus, multiple concepts can be applied to the same image simultaneously, as shown in Fig. 6. For example, we can see that the glasses added first remain unchanged with subsequently added edits. We can utilize this behavior to perform more complex changes, which are best expressed using multiple concepts. Fig. 5 provides additional details on how SEGA achieves this isolation level. The methodology ensures that any manipulation restricts changes to the semantically relevant regions of the image.

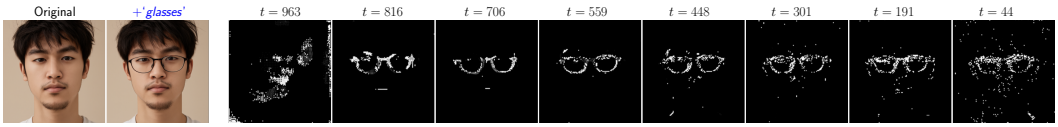


Figure 5: SEGA simplifies creativity for digital artworks as it implicitly identifies image regions relevant to certain concepts. Implicit masks are shown over diffusion timesteps t of SD-3 for edit concept '*glasses*'.

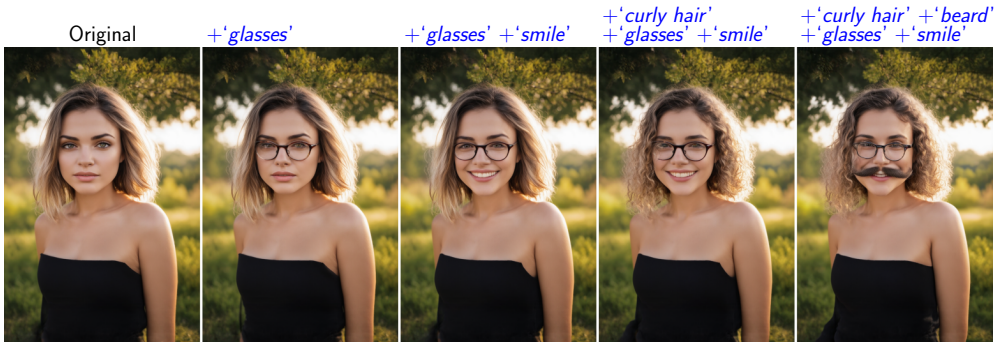
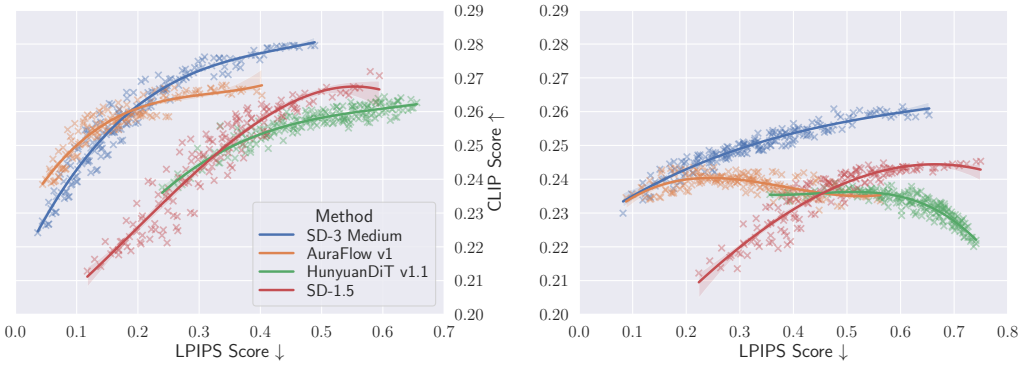


Figure 6: Successive combination of concepts. We progressively manipulate an additional concept in each image. Concepts do not interfere with each other and only change the relevant portion of the image.

4 Empirical Evaluation

Next, we empirically confirm that SEGA offers similar levels of generation steerability for DiT architectures. To that end, we compare three recent text-to-image DiTs [7, 17, 13] against Stable Diffusion 1.5 [16], on which SEGA was evaluated originally. We rely on a well-established setup for semantic image manipulation to evaluate attribute manipulation in facial images [3, 4]. Specifically, we generate portrait images and subsequently manipulate five different attributes like '*smile*' or '*earrings*'. In the first evaluation setting, we only manipulate one attribute at a time, whereas in the second setting, we consider multi-conditioning for three simultaneous edits for a total of 10 combinations.

To accurately reflect the trade-off between the versatility of edits and the precision of those manipulations, we perform an extensive hyperparameter sweep for each model. We perform each edit across 16 different seeds, resulting in over 130k generated and evaluated images in total. As measures for comparison, we employ CLIP [10] and LPIPS [20] scores. CLIP measures the text-to-image similarity of the edit instruction to the edited image, and LPIPS measures the image-to-image similarity of the original to the edited image. Consequently, CLIP assesses the versatility and LPIPS the precision,



(a) Single-concept editing

(b) Multi-concept editing

Figure 7: Comparison of instruction-alignment vs. image similarity trade-off of SEGA for different models. Results were reported for manipulating five facial attributes, with each point representing one hyper-parameter configuration. We plot CLIP scores (higher is better) of the target attributes against LPIPS similarity (lower is better). Steerability with SEGA clearly extends to DiT models while benefiting from stronger base capabilities.

allowing us to plot the trade-off between them. We base the SEGA implementation of all models on their respective pipelines in `diffusers`[2].

The results of both experiments are depicted in Fig. 7. We can observe a similar trade-off curve for SD-1.5 and SD-3 in both scenarios, suggesting a comparable level of semantic control offered by SEGA. Further, we can observe that SD-3 strongly improves on the general capabilities of its precursor, delivering better scores across the board. SEGA as a model-agnostic approach can thus always provide a consistent user experience for the strongest models available. Additionally, we see some general differences between the single and multi-edit experiments. First, the maximum CLIP scores are lower for the latter, which can mainly be attributed to the fact that the multi-edits do not lead to a single over-pronounced feature. Second, both AuraFlow and HunyuanDiT significantly underperform in the multi-conditioning setting, hinting at some underlying limitations in these models.

5 Conclusion

In this work, we demonstrated that Semantic Guidance (SEGA) generalizes to different DM architectures, including state-of-the-art diffusion transformers (DiTs). Our findings highlight the robustness, uniqueness, monotonicity, and isolation properties of SEGA, confirming its architecture-agnostic nature. By extending SEGA to recent DiT models, we have shown that it maintains and enhances the creative control available to users, enabling new possibilities such as sophisticated typographic manipulations.

Our empirical evaluations, conducted on multiple DiT models, indicate strong performance in image manipulations, providing a consistent user experience across different model architectures. Furthermore, SEGA's ability to facilitate iterative creative workflows and complex semantic manipulations without the need for user-drawn masks or other inputs underscores its practical utility in various creative applications. As generative models continue to evolve rapidly, the ability to apply a consistent, high-quality semantic manipulation method like SEGA across different architectures will become increasingly valuable.

Acknowledgments

We acknowledge support of the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation), the hessian.AISC Service Center (funded by the Federal Ministry of Education and Research, BMBF, grant No 01IS22091), and the German Research Center for AI (DFKI). Further, this work benefited from the ICT-48 Network of AI Research Excellence Center "TAILOR" (EU Horizon 2020, GA No 952215), the Hessian research priority program LOEWE within the project WhiteBox, the HMWK cluster projects "Adaptive Mind" and "Third Wave of AI", and from the NHR4CES.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph*, 42, 2023.

[2] Victor Boutin, Thomas Fel, Lakshya Singhal, Rishav Mukherji, Akash Nagaraj, Julien Colin, and Thomas Serre. Diffusion models as artists: Are we closing the gap between humans and machines? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[3] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[4] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. LEDITS++: limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[2] `https://github.com/huggingface/diffusers`

[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.*, 42, 2023.

[6] Sofia Crespo. Sofia crespo - ai artists. `https://aiartists.org/sofia-crespo`, 2024. Accessed: 2024-08-16.

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[8] Michael Filimowicz. The ten most influential works of ai art. `https://medium.com/higher-neurons/the-ten-most-influential-works-of-ai-art-820c596b8840`, 2020. Accessed: 2024-08-16.

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[11] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have A semantic latent space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[13] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv:2405.08748*, 2024.

[14] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022.

[15] 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.

[16] 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 (CVPR)*, 2022.

[17] Simo Ryu. Introducing auraflow v0.1, an open exploration of large rectified flow models. 2024. Accessed: 2024-08-16.

[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[19] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.