SEGA: Instructing Diffusion using Semantic Dimensions

Manuel Brack [1] **Felix Friedrich** [1 2] **Dominik Hintersdorf** [1] **Lukas Struppek** [1] **Patrick Schramowski** [1 2 3 4]
Kristian Kersting [1 2 3 5]

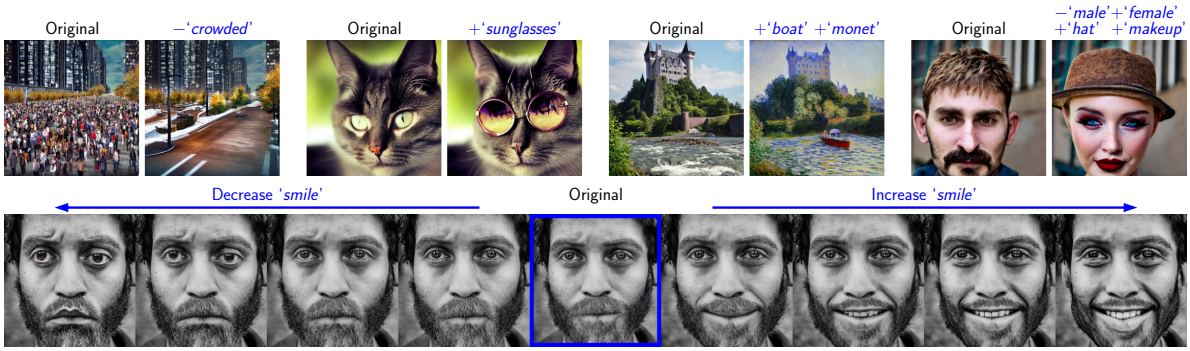


Figure 1: Semantic control over image generation with SEGA. It allows intuitive control over arbitrary concepts and their combinations using simple textual edits. SEGA an be executed ad-hoc during inference without any fine-tuning or optimization. The bottom row demonstrates that SEGA guidance vectors scale monotonically. (Best viewed in color)

Abstract

Text-to-image diffusion models have recently received a lot of interest for their astonishing ability to produce high-fidelity images from text only. However, achieving one-shot generation that aligns with the user's intent is nearly impossible, yet small changes to the input prompt often result in very different images. This leaves the user with little semantic control. To put the user in control, we show how to interact with the diffusion process to flexibly steer it along semantic directions. This semantic guidance (SEGA) allows for subtle and extensive edits, changes in composition and style, as well as optimizing the overall artistic conception. We demonstrate SEGA's effectiveness on a variety of tasks and provide evidence for its versatility and flexibility.

1. Introduction

The recent popularity of text-to-image diffusion models (Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022) can largely be attributed to their versatility, expressiveness, and—most importantly—the intuitive interface they provide to users. The generation's intent can easily be expressed in natural language, with the model producing faithful interpretations of a text prompt. Despite the impressive capabilities of these models, the initially generated images are rarely of high quality. Accordingly, a human user will likely be unsatisfied with certain aspects of the initial image, which they will attempt to improve over multiple iterations. Unfortunately, the diffusion process is rather fragile as small changes to the input prompt lead to entirely different images. Consequently, fine-grained semantic control over the generation process is necessary, which should be as easy and versatile to use as the initial generation.

Previous attempts to influence dedicated concepts during the generation process require additional segmentation masks, extensions to the architecture, model fine-tuning, embedding optimization (Avrahami et al., 2022; Hertz et al., 2022; Kawar et al., 2022; Wu et al., 2022). While these techniques produce satisfactory results, they disrupt the fast, exploratory workflow that is the strong suit of diffusion models in the first place. We propose Semantic Guidance (SEGA) to uncover and interact with semantic directions inherent to the model. SEGA requires no additional training, no extensions to the architecture, nor external guidance and is calculated within a single forward pass. We demonstrate that this semantic control can be inferred from simple textual descriptions using the model's noise estimate alone. With this, we also refute previous research claiming these estimates to be unsuitable for semantic control (Kwon et al.,

[1]Technical University Darmstadt [2]Hessian Center for AI (hessian.AI) [3]German Research Center for Artificial Intelligence (DFKI) [4]LAION [5]Centre for Cognitive Science Darmstadt. Correspondence to: Manuel Brack <brack@cs.tu-darmstadt.de>, Patrick Schramowski <schramowski@cs.tu-darmstadt.de>.

2022). The guidance vectors uncovered with SEGA are robust, scale monotonically, and are largely isolated. This enables simultaneous applications of subtle edits to images, changes in composition and style, as well as optimizing the artistic conception. Furthermore, SEGA allows for probing the latent space of diffusion models to gain insights into how abstract concepts are represented by the model and how their interpretation reflects on the generated image.

In this paper, we establish the methodical benefits of SEGA and demonstrate that this intuitive, lightweight approach offers sophisticated semantic control over image generations. Specifically, we contribute by (i) devising a formal definition of Semantic Guidance and discussing the numerical intuition of the corresponding semantic space, (ii) demonstrating the robustness, uniqueness, monotonicity, and isolation of semantic vectors, and (iii) providing an exhaustive empirical evaluation of SEGA's semantic control.

2. Background

Semantic Dimensions. Research on expressive semantic vectors that allow for meaningful interpolation and arithmetic pre-date generative diffusion models. Addition and subtraction on text embeddings such as word2vec (Mikolov et al., 2013a;b) have been shown to reflect semantic and linguistic relationships in natural language (Mikolov et al., 2013c; Vylomova et al., 2016). One of the most prominent examples is that the vector representation of 'King - male + female' is very close to 'Queen'. SEGA enables similar arithmetic for image generation with diffusion (cf. Fig. 2b). Up to now, StyleGANs (Karras et al., 2019; 2020) also contain semantic dimensions that can be utilized during generation. For example, Patashnik et al. (2021) combined these models with CLIP (Radford et al., 2021) to offer limited textual control over generated attributes. However, training StyleGANs at scale with subsequent fine-tuning is notoriously fragile due to the challenging balance between reconstruction and adversarial loss. Yet, large-scale pre-training is the base of flexible and capable generative models (Petroni et al., 2019).

Image Diffusion. Recently, large-scale, text-guided diffusion models have enabled a more versatile approach for image generation (Saharia et al., 2022; Ramesh et al., 2022; Balaji et al., 2022). Especially latent diffusion models (Rombach et al., 2022) have been gaining much attention. These models perform the diffusion process on a compressed space perceptually equivalent to the image space. For one, this approach reduces computational requirements. Additionally, the latent representations can be utilized for other downstream applications (Frans et al., 2021; Liu et al., 2015).

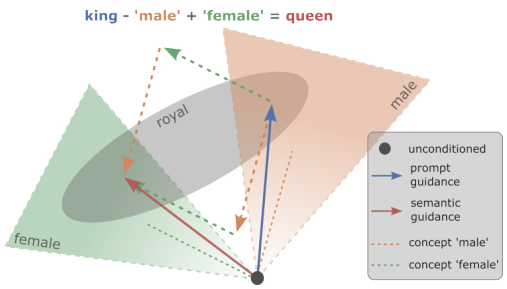
Image Editing. While these models produce astonishing, high-quality images, fine-grained control over this process remains challenging. Minor changes to the text prompt often lead to entirely different images. One approach to tackle this issue is inpainting, where the user provides additionally semantic masks to restrict changes to certain areas of the image (Avrahami et al., 2022; Nichol et al., 2022). Other methods involve computationally expensive fine-tuning of the model to condition it on the source image before applying edits (Kawar et al., 2022; Valevski et al., 2022). In contrast, SEGA performs edits on the relevant image regions through text descriptions alone and requires no tuning.

Semantic Control. Other works have explored more semantically grounded approaches for interacting with image generation. Prompt-to-Prompt utilizes the semantic information of the model's cross-attention layers that attribute pixels to tokens from the text prompt (Hertz et al., 2022). Dedicated operations on the cross-attention maps enable various changes to the generated image. On the other hand, SEGA does not require token-based conditioning and allows for combinations of multiple semantic changes. Wu et al. (2022) studied the disentanglement of concepts for diffusion models using linear combinations of text embeddings. However, for each text prompt and target concept, a dedicated combination must be inferred through optimization. Moreover, the approach only works for more substantial changes to an image and fails for small edits. SEGA, in contrast, is capable of performing such edits without optimization.

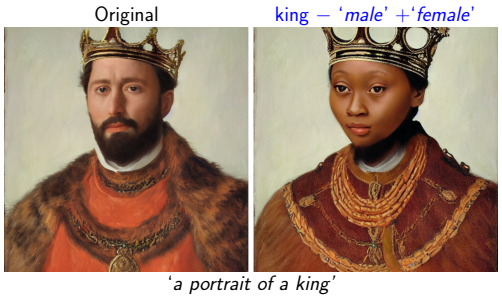
Noise-Estimate Manipulation. Our work is closely related to previous research working directly on the noise estimates of diffusion models. Liu et al. (2022) combine multiple estimates to facilitate changes in image composition. However, more subtle semantic changes to an image remain unfeasible with this method. In fact, Kwon et al. (2022) argue that the noise-estimate space of diffusion models is unsuited for semantic manipulation of the image. Instead, they use a learned mapping function on changes to the bottleneck of the underlying U-Net. This approach enables various manipulations that preserve the original image quality. However, it does not allow for arbitrary spontaneous edits of the image, as each editing concept requires minutes of training. SEGA, in comparison, requires no extension to the architecture and produces semantic vectors ad-hoc for any textual prompt. Lastly, Safe Latent Diffusion (SLD) uses targeted manipulation of the noise estimate to suppress inappropriate content during image generation (Schramowski et al., 2022). Instead of arbitrary changes to an image, SLD prevents one dedicated concept from being generated. Additionally, SLD is complex, and the hyperparameter formulation can be improved through a deeper understanding of the numerical properties of diffusion models' noise estimate space.

3. Semantic Guidance

Let us now devise Semantic Guidance for diffusion models.



(a) A (latent) diffusion process inherently organizes concepts and learns implicitly relationships between them, although there is no supervision.



(b) Guidance arithmetic: Guiding the image 'a portrait of a king' (left) using 'king'−'male'+'female' results in an image of a 'queen' (right).

Figure 2: Semantic guidance (SEGA) applied to the image 'a portrait of a king' using 'king'−'male'+'female'. (Best viewed in color)

3.1. Guided Diffusion

The first step towards SEGA is guided diffusion. Specifically, diffusion models (DM) iteratively denoise a Gaussian distributed variable to produce samples of a learned data distribution. For text-to-image generation, the model is conditioned on a text prompt p and guided toward an image faithful to that prompt. The training objective of a diffusion model $\hat{x}_\theta$ can be written as

$$\mathbb{E}_{\mathbf{x},\mathbf{c}_p,\epsilon,t}\left[w_t||\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\omega_t\epsilon,\mathbf{c}_p)-\mathbf{x}||_2^2\right] \quad (1)$$

where $(\mathbf{x},\mathbf{c}_p)$ is conditioned on text prompt p, t is drawn from a uniform distribution $t\sim\mathcal{U}([0,1])$, ϵ sampled from a Gaussian $\epsilon\sim\mathcal{N}(0,\mathbf{I})$, and w_t,ω_t,α_t influence the image fidelity depending on t. Consequently, the DM is trained to denoise $\mathbf{z}_t:=\mathbf{x}+\epsilon$ yielding $\mathbf{x}$ with the squared error loss. At inference, the DM is sampled using the model's prediction of $\mathbf{x}=(\mathbf{z}_t-\tilde{\epsilon_\theta})$, with $\tilde{\epsilon_\theta}$ as described below.

Classifier-free guidance (Ho & Salimans, 2022) is a conditioning method using a purely generative diffusion model, eliminating the need for an additional pre-trained classifier. During training, the text conditioning $\mathbf{c}_p$ drops randomly with a fixed probability, resulting in a joint model for unconditional and conditional objectives. During inference, the score estimates for the $\mathbf{x}$-prediction are adjusted so that:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t,\mathbf{c}_p):=\epsilon_\theta(\mathbf{z}_t)+s_g(\epsilon_\theta(\mathbf{z}_t,\mathbf{c}_p)-\epsilon_\theta(\mathbf{z}_t)) \quad (2)$$

with guidance scale s_g and ϵ_θ defining the noise estimate with parameters θ. Intuitively, the unconditioned ϵ-prediction is pushed in the direction of the conditioned one, with s_g determining the extent of the adjustment.

3.2. Semantic Guidance on Concepts

We introduce SEGA to influence the diffusion process along several directions. To this end, we substantially extend the principles introduced in classifier-free guidance by solely interacting with the concepts already present in the model's latent space. Therefore, SEGA requires no additional training, no extensions to the architecture, and no external guidance. Instead, it is calculated during the existing diffusion iteration. More specifically, SEGA uses multiple textual descriptions e_i, representing the given target concepts of the generated image, in addition to the text prompt p.

Intuition. The overall idea of SEGA is best explained using a 2D abstraction of the high dimensional ϵ-space, as shown in Fig. 2. Intuitively, we can understand the space as a composition of arbitrary sub-spaces representing semantic concepts. Let us consider the example of generating an image of a king. The unconditioned noise estimate (black dot) starts at some random point in the ϵ-space without semantic grounding. The guidance corresponding to the prompt "a portrait of a king" represents a vector (blue vector) moving us into a portion of ϵ-space where the concepts 'male' and royal overlap, resulting in an image of a king. We can now further manipulate the generation process using SEGA. From the unconditioned starting point, we get the directions of 'male' and 'female' (orange/green lines) using estimates conditioned on the respective prompts. If we subtract this inferred 'male' direction from our prompt guidance and add the 'female' one, we now reach a point in the ϵ-space at the intersection of the 'royal' and 'female' sub-spaces, i.e., a queen. This vector represents the final direction (red vector) resulting from semantic guidance.

Isolating Semantics in Diffusion. Next, we investigate the actual noise-estimate space of Stable Diffusion (SD). This enables extracting semantic concepts from within that space and applying them during image generation.

Numerical values of ϵ-estimates are generally Gaussian distributed. While the value in each dimension of the latent vector can differ significantly between seeds, text prompts, and diffusion steps, the overall distribution always remains similar to a Gaussian distribution (cf. App. B). Using the arithmetic principles of classifier-free guidance, we can now identify those dimensions of a latent vector encoding an

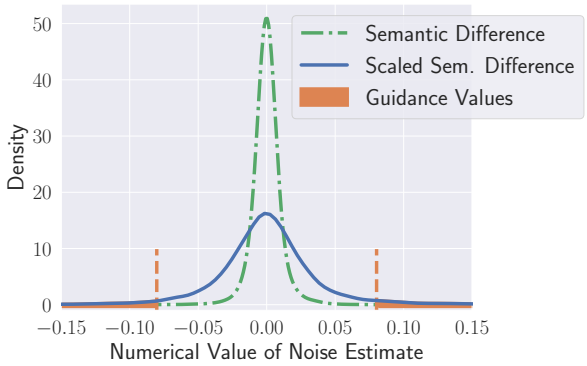


Figure 3: Numerical intuition of semantic guidance. The difference between the concept-conditioned and unconditioned estimates is first scaled. Subsequently, the values in the upper and lower tail are used as the dimensions representing the specified concept. Distribution plots calculated using kernel-density estimates with Gaussian smoothing.

arbitrary semantic concept. To that end, we calculate the noise estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$, which is conditioned on a concept description e. We then take the difference between $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$ and the unconditioned estimate $\epsilon_\theta(\mathbf{z}_t)$ and scale it. Again, the numerical values of the resulting latent vector are Gaussian distributed, as shown in Fig. 3. We will demonstrate that those latent dimensions falling into the upper and lower tail of the distribution alone encode the target concept. We empirically determined that using only 1-5% of the ϵ-estimate's dimensions is sufficient to apply the desired changes to an image. Consequently, the resulting concept vectors are largely isolated; thus, multiple ones can be applied simultaneously without interference (cf. Sec. 4). We subsequently refer to the space of these sparse noise-estimate vectors as *semantic space*.

One Direction. Let us formally define the previous intuition for SEGA by starting with a single direction, i.e., editing prompt. Again, we use three ϵ-predictions to move the unconditioned score estimate $\epsilon_\theta(\mathbf{z}_t)$ towards the prompt conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$ and simultaneously away/towards the concept conditioned estimate $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$, depending on the editing direction. Formally, we compute

$$\epsilon_\theta(\mathbf{z}_t) + s_g\big(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p) - \epsilon_\theta(\mathbf{z}_t)\big) + \gamma(\mathbf{z}_t, \mathbf{c}_e) \tag{3}$$

with the semantic guidance term γ

$$\gamma(\mathbf{z}_t, \mathbf{c}_e) = \mu(\psi; s_e, \lambda)\psi(\mathbf{z}_t, \mathbf{c}_e) \tag{4}$$

where μ applies an editing guidance scale s_e element-wise, and ψ depends on the editing direction:

$$\psi(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_e) = \begin{cases} \epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e) - \epsilon_\theta(\mathbf{z}_t) & \text{if pos. guidance} \\ -\big(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e) - \epsilon_\theta(\mathbf{z}_t)\big) & \text{if neg. guidance} \end{cases} \tag{5}$$

Consequently, changing the guidance direction is reflected by the direction of the vector between $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$ and $\epsilon_\theta(\mathbf{z}_t)$.

The term μ (Eq. 4) considers those dimensions of the prompt conditioned estimate relevant to the defined editing prompt e. To this end, μ takes the largest absolute values of the difference between the unconditioned and concept-conditioned estimates. This corresponds to the upper and lower tail of the numerical distribution as defined by percentile threshold λ. All values in the tails are scaled by an edit scaling factor s_e, with everything else being set to 0, such that

$$\mu(\psi; s_e, \lambda) = \begin{cases} s_e & \text{where } |\psi| \geq \eta_\lambda(|\psi|) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\eta_\lambda(\psi)$ is the λ-th percentile of ψ. Consequently, larger values of s_e increase the effect of semantic guidance.

To offer even more control over the diffusion process, we make two adjustments to the methodology presented above. We add a warm-up parameter δ that will only apply guidance γ after an initial warm-up period in the diffusion process, i.e., $\gamma(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_S) := \mathbf{0}$ if $t < \delta$. Naturally, higher values for δ lead to less significant adjustments of the generated image. If we aim to keep the overall composition of the image unchanged, selecting a sufficiently high δ ensures only altering fine-grained details of the output.

Furthermore, we add a momentum term ν_t to the semantic guidance γ in order to accelerate guidance over time steps for dimensions that are continuously guided in the same direction. Hence, γ_t is defined as:

$$\gamma(\mathbf{z}_t, \mathbf{c}_e) = \mu(\psi; s_e, \lambda)\psi(\mathbf{z}_t, \mathbf{c}_e) + s_m\nu_t \tag{7}$$

with momentum scale $s_m \in [0, 1]$ and ν being updated as

$$\nu_{t+1} = \beta_m\nu_t + (1 - \beta_m)\gamma_t \tag{8}$$

where $\nu_0 = \mathbf{0}$ and $\beta_m \in [0, 1)$. Thus, larger β_m lead to less volatile changes in momentum. Momentum is already built up during the warm-up period, even though γ_t is not applied during these steps.

Beyond One Direction. Now, we are ready to move beyond using just one direction towards multiple concepts e_i and, in turn, combining multiple calculations of γ_t.

For all e_i, we calculate γ_t^i as described above with each defining their own hyperparameter values λ^i, s_e^i. The weighted sum of all γ_t^i results in

$$\hat{\gamma}_t(\mathbf{z}_t, \mathbf{c}_p; \mathbf{e}) = \sum\nolimits_{i \in I} g_i\gamma_t^i(\mathbf{z}_t, \mathbf{c}_p, \mathbf{c}_{e_i}) \tag{9}$$

In order to account for different warm-up periods, g_i is defined as $g_i = 0$ if $t < \delta_i$. However, momentum is built

(a) Robustness of guidance vectors. Results for guiding towards '*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).

Figure 4: Showcasing robustness and uniqueness of the semantic guidance vectors inferred with SEGA. Top row depicts the unchanged image, while the bottom row depicts the ones guided towards '*glasses*'. (Best viewed in color)

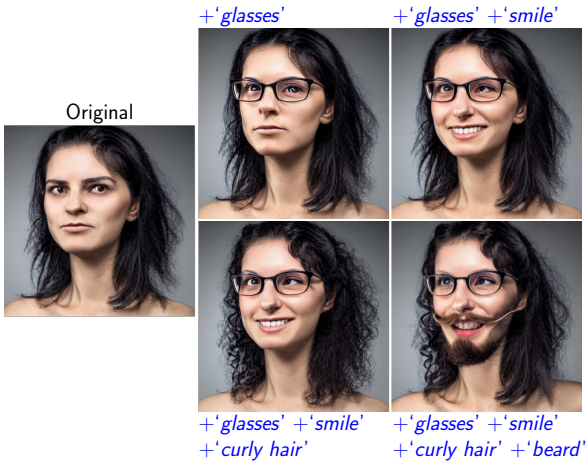


Figure 5: Successive combination of concepts. From top left to bottom right an additional concept is added in each image. The concepts do not interfere with each other and only change the relevant portion of the image. (Best viewed in color)

up using all editing prompts and applied once all warm-up periods are completed, i.e., $\forall \delta_i : \delta_i \geq t$. We provide a pseudo-code implementation of SEGA in App. A.

SEGA's underlying methodology is architecture-agnostic and applicable to any model employing classifier-free guidance. For subsequent experiments, we base our implementation on SD v1.5[1] and make our code available online. We note that SEGA can easily be applied to real images using reconstruction techniques for diffusion models. However, this is out of scope for this work, wherefore we limit the depicted examples and evaluation to generated images.

4. Properties of Semantic Space

With the fundamentals of semantic guidance established, we next investigate the properties of SEGA's semantic space.

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

In addition to the following discussion, we present further examples in the Appendix.

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. This makes SEGA's use easy and provides the same exploratory nature as the initial image generation.

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. This even covers significant domain shifts, as seen in the one 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. 1, 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

Figure 6: Examples from our empirical evaluation benchmark. Showcases the 10 attributes edited with SEGA. Original and edited images are evaluated through a user study on whether they exhibit a certain feature. (Best viewed in color)

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. 5. We can see, for example, that the glasses which were added first remain unchanged with subsequently added edits. We can utilize this behavior to perform more complex changes, best expressed using multiple concepts. One example is the change of gender by simultaneously removing the 'male' concept and adding the 'female' one (cf. Figs. 1 and 6).

5. Experimental Evaluation

Next, we present an exhaustive evaluation of semantic guidance on an empirical benchmark, as well as on a variety of qualitative tasks. The intention with SEGA is not to outperform existing methods. The performance and fidelity of outputs always depends on the underlying capabilities of the model. Our main focus are the inherent semantic capabilities of diffusion models by examining the level of control available in interacting with the model. Refuting claims of previous research, we demonstrate the suitability of noise estimates for semantic control (Kwon et al., 2022) and its capability for small, subtle changes (Wu et al., 2022).

5.1. Empirical

We performed an extensive empirical evaluation on human faces and respective attributes. This setting is inspired by the CelebA dataset (Liu et al., 2015) and marks a well-established benchmark for semantic changes in image generation. We generated 250 images with unique seeds using the prompt 'an image of the face of a random person' and manipulated ten facial attributes. These attributes are a subset of the CelebA labels. All attributes and respective examples of the corresponding manipulation using SEGA are depicted in Fig. 6. In addition to additive image edits, we evaluated negative guidance using three of these attributes as well as two combinations of four simultaneous edits, as shown in Fig. 5. Therefore, our empirical evaluation spans 15 attribute changes and combinations in total.

We evaluated the generated images with a user study and provide more details on its implementation in App. D. The results are shown in Tab. 1. For positive guidance, SEGA faithfully adds the target concept to the image on average in 95% of the cases, with the majority of attributes exceeding 97%. We further manually investigated the two outliers, '*Bald*' and '*Bangs*'. We assume that many of the non-native English-speaking annotators were not familiar with the term 'bangs' itself. This assumption is based on correspondence with some of the workers, and the conspicuously low rate of annotator consensus. Consequently, the numbers for '*bangs*' should be taken with a grain of salt. For baldness, we found that long hair often makes up a large portion of a portrait and thus requires more substantial changes to the image. Consequently, such edits require stronger hyperparameters than those chosen for this study. We observe negative guidance to remove existing attributes from an image to work similarly well. It is worth pointing out that the guidance away from '*beard*' usually resulted in a substantial reduction in facial hair, but failed to remove it entirely for $\sim$10% of the images. Again, this suggests that the hyperparameters were probably not strong enough.

Lastly, we look into the simultaneous guidance of multiple concepts at once. The results in Tab. 2 empirically demonstrate the isolation of semantic guidance vectors. The per-attribute success rate remains similar for four instead of one distinct edit concept, suggesting no interference of guidance vectors. Consequently, the success of multiple edits only depends on the joint probability of the individual concepts. In comparison, if only two out of four applied concepts were interfering with each other to be mutually exclusive, the success rate of such a combination would always be 0%. Contrary to that, we successfully apply concurrent concepts in up to 91% of generated images.

Table 1: Empirical results of our user study conducted on face attributes. Sample sizes result from the portion of the 250 original images that did not contain the target attribute. Annotator consensus refers to the percentage of images for which the majority of annotators agreed on a label. Success rate is reported on those images with annotator consensus.

	Attribute	Samples	Consensus (%)	Success (%)
Pos. Guidance	Gender	241	99.2	100.0
	Glasses	243	99.6	100.0
	Smile	146	100.0	99.3
	Bald	220	91.2	82.1
	Beard	135	97.8	97.0
	Hat	210	99.0	99.0
	Curls	173	95.4	97.0
	Makeup	197	99.5	99.0
	Gray hair	165	97.6	91.2
	Bangs	192	86.2	82.7
	Overall	**1922**	**96.5**	**95.0**
Neg. Guidance	No Glasses	6	98.9	100.0
	No Smile	93	100.0	94.4
	No Beard	111	100.0	89.9
	Overall	**210**	**99.5**	**92.1**

Table 2: Results of user study on simultaneous combination of face attributes. Sample sizes result from the portion of the 250 original images that did not contain any target attribute. Annotator consensus refers to the percentage of images for which the annotator's majority agreed on a label. Success rates for combinations on any combination of x attributes with per-attribute scores reflecting the isolated success of that edit. Nonetheless, all scores are shown for images with all 4 edit concepts applied simultaneously.

	Attribute	Samples	Consensus (%)	Success (%)
Combination 1	$\geq$ 1 Attr.	55	98.2	100.0
	$\geq$ 2 Attr.			100.0
	$\geq$ 3 Attr.			98.2
	All 4 Attr.			**90.7**
	Glasses	55	96.4	100.0
	Smile		100.0	96.4
	Curls		100.0	98.2
	Beard		100.0	100.0
Combination 2	$\geq$ 1 Attr.	45	81.8	100.0
	$\geq$ 2 Attr.			100.0
	$\geq$ 3 Attr.			100.0
	All 4 Attr.			**75.6**
	No Smile	45	100.0	97.8
	Makeup		100.0	97.6
	Hat		100.0	88.6
	Female		100.0	100.0

5.2. Qualitative

In addition to the empirical evaluation, we present qualitative examples on other domains and tasks. We show further examples in higher resolution in the Appendix. Overall, this highlights the versatility of SEGA since it allows interaction with any of the abundant number of concepts diffusion models are capable of generating in the first place. In Fig. 7a, we can see various edits being performed on an image of a car. This showcases the range of potential changes feasible with SEGA, which include color changes, altering the vehicle type or surrounding scenery. In each scenario, the guidance vector inferred using SEGA accurately targets the respective image region and performs changes faithful to the editing prompt. All while making next to no changes to the irrelevant image portions.

Furthermore, we performed a diverse set of style transfers, as shown in Fig. 7b. SEGA faithfully applies the styles of famous artists, as well as artistic epochs and drawing techniques. In this case, the entirety of the image has to be changed while keeping the image composition the same. Consequently, we observed that alterations to the entire output—as in style transfer—require a slightly lower threshold of $\lambda \approx 0.9$. Nonetheless, this still means that 10% of the ϵ-space is sufficient to change the entire style of an image. Fig. 7b also includes a comparison between outputs produced by SEGA with those from simple extensions to the prompt text. Changing the prompt also significantly alters the image composition. These results further highlight the advantages of semantic control, which allows versatile and yet robust changes.

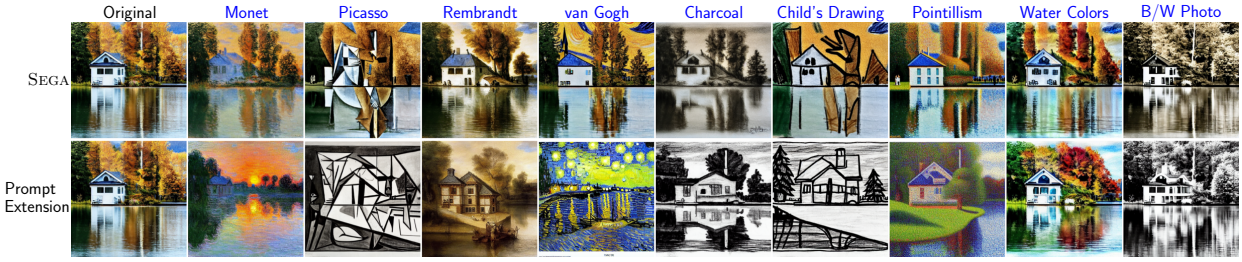
An additional benefit of semantic guidance directly on noise estimates is its independence on the modality of the concept description. In the case of SD, the modality happens to be text in natural language, but SEGA can be applied to any conditional diffusion model. We explore this idea further with the goal of optimizing the overall artistic conception of generated images. Instead of defining the target concept with one text prompt, we directly use an abstract embedding, representing a particular type of imagery. To that end, we collected prompts known to produce high-quality results[2] for five different types of images in *portrait photography, animation, concept art, character design*, and *modern architecture*. The conditioning embedding for one style is calculated as the average over the embeddings of all collected prompts. Exemplary outputs are depicted in Fig. 7c. The results are of high quality and stay close to the original image but accurately reflect the targeted artistic direction beyond a single text prompt. Since concept vectors are isolated (cf. Sec. 4), we can also apply various types of changes (e.g. style transfer + composition) to the image simultaneously.

6. Broader Impact on Society

Recent developments in text-to-image models (Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022) have the potential for a far-reaching impact on society, both positive and negative, when deployed in applications such as image generation, image editing, or search engines. Previous research (Bianchi et al., 2022; Schramowski et al., 2022) described many potential negative societal implications that may arise due to the careless use of such large-scale gener-

[2]Prompts taken from https://mpost.io/best-100-prompts

(a) Examples of image editing with prompt 'an image of a car'. All images are generated from the same noise latent and text prompts. SEGA performs high quality local and global edits with minimal changes to other aspects of the image.



(b) Style transfer using SEGA. All images are generated from the same noise latent using the text prompt 'a house at a lake'. SEGA easily applies the characteristics of a specific artist, epoch or drawing/imaging technique to the original image while preserving the overall composition. In contrast, appending the prompt with a style instruction results in images that significantly change the composition.

(c) Optimization of artistic conception. Guidance conditioned on average embedding of prompt targeting one type of imagery. The resulting images are of high quality, reflect the targeted artistic style, while staying true to the original composition.

Figure 7: Qualitative Examples of Semantic Guidance on various tasks and domains. (Best viewed in color)

ative models. Many of these problems can be attributed to the noisy, large-scale datasets these models rely on. Since recent text-to-image models, such as SD, are trained on web-crawled data containing inappropriate content (Schuhmann et al., 2022), they are no exception to this issue. Specifically, current versions of SD show signs of inappropriate degeneration (Schramowski et al., 2022). While Schramowski et al. (2022) utilize the model's notion of inappropriateness to steer the model away from generating related content, it is noteworthy that we introduce an approach that could also be used to guide image generation toward inappropriate material. However, on the positive side, SEGA has the potential to mitigate bias. As demonstrated by Nichol et al. (2022), removing data from the training set has adverse effects, e.g., on a model's generalization ability. In contrast, SEGA works at inference promoting fairness in the outcome. Therefore, we advocate for further research in this direction.

Another frequently voiced point of criticism is the notion that generative models like SD are replacing human artists and illustrators. At first glance, the great results produced by these models might warrant this impression which is further fueled by unbalanced media coverage. However, the generative process as a whole still requires a substantial amount of iterative human feedback and creative thinking. SEGA further promotes creative artwork by providing intuitive means of interaction that support an exploratory, creative process.

7. Conclusions

We introduced semantic guidance (SEGA) for diffusion models. SEGA facilitates interaction with arbitrary concepts during image generation. The approach requires no additional training, no extensions to the architecture, no external guidance, and is calculated during the existing generation process. The concept vectors identified with SEGA are robust, isolated, can be combined arbitrarily, and scale monotonically. We evaluated SEGA on a variety of tasks and domains, highlighting—among others—sophisticated image composition and editing capabilities.

Our findings are highly relevant to the debate on disentangling models' latent spaces. So far, disentanglement as a property has been actively pursued (Karras et al., 2020). However, it is usually not a necessary quality in itself but a means to an end to easily interact with semantic concepts. We demonstrated that this level of control is feasible without disentanglement and motivate research in this direction.

Additionally, we see several other exciting avenues for future work. For one, it is interesting to investigate further how concepts are represented in the latent space of DM's and how to quantify them. More importantly, automatically detecting concepts could provide novel insights and toolsets to mitigate biases, as well as enacting privacy concerns of real people memorized by the model.

Acknowledgments This research has benefited from the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) cluster projects "The Third Wave of AI" and hessian.AI, from the German Center for Artificial Intelligence (DFKI) project "SAINT", the Federal Ministry of Education and Research (BMBF) project KISTRA (reference no. 13N15343), as well as from the joint ATHENE project of the HMWK and the BMBF "AVSV".

References

Avrahami, O., Fried, O., and Lischinski, D. Blended latent diffusion. *arXiv:2206.02779*, 2022. 1, 2

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., and Liu, M. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv:2211.01324*, 2022. 2

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv:2211.03759*, 2022. 7

Frans, K., Soros, L. B., and Witkowski, O. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv:2106.14843*, 2021. 2

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control, 2022. 1, 2

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 3

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 8

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models, 2022. 1, 2

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. *arXiv:2210.10960*, 2022. 1, 2, 6

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, December 2015. 2, 6

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013a. 2

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2013b. 2

Mikolov, T., Yih, W., and Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2013c. 2

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2022. 2, 7, 8

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P. S. H., Bakhtin, A., Wu, Y., and Miller, A. H. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 2

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents, 2022. 1, 2, 7

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022. 1, 2, 7

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *arXiv:2211.05105*, 2022. 2, 7, 8

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 8

Valevski, D., Kalman, M., Matias, Y., and Leviathan, Y. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv:2210.09477*, 2022. 2

Vylomova, E., Rimell, L., Cohn, T., and Baldwin, T. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL*, 2016. 2

Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., and Chang, S. Uncovering the disentanglement capability in text-to-image diffusion models. *arXiv:1212.08698*, 2022. 1, 2, 6

A. Implementation of SEGA

The implementation of SEGA builds on the Stable Diffusion pipeline from the diffusers library.[3] As describe in Sec. 3 we calculate an additional noise estimate for each semantic concept. Consequently, we make one pass through the models U-Net for each concept description. Our implementation is included in the supplementary material, together with the necessary code and hyper-parameters to generate examples shown in the paper.

Additionally, we also provide the pseudo-code notation of SEGA in Alg 1. Please note that this code is slightly simplified to use one single warm-up period δ for all edit prompts e_i.

B. Numerical Properties of Noise Estimates

As discussed in Sec. 3.2 the numerical values of noise estimates are generally Gaussian distributed. This can be attributed to the fact that they are trained to estimate a Gaussian sampled ϵ. We plotted exemplary numerical distributions of three noise estimates in Fig. 8. These estimates are an unconditioned $\epsilon_\theta(\mathbf{z}_t)$, a text conditioned $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_p)$, and an edit conditioned $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}_e)$ one. All calculated at diffusion step 20. While these estimates lead to substantially different images, it is clear that their overall numerical distribution is almost identical.

[3] https://github.com/huggingface/diffusers

Algorithm 1 Semantic Guidance (SEGA)

Please note that this notation makes the simplifying assumption of one single warm-up period δ for all edit prompts e_i.

Require: model weights θ, text condition $text_p$, edit texts **List**($text_e$) and diffusion steps T
Ensure: $s_m \in [0,1]$, $\nu_{t=0} = 0$, $\beta_m \in [0,1)$, $\lambda^i \in (0,1)$, $s_e^i \in [0,20]$, $s_g \in [0,20]$, $\delta \in [0,20]$, $t = 0$

DM $\leftarrow$ init-diffusion-model(θ)
$c_p \leftarrow$ DM.encode($text_p$)
List(c_e) $\leftarrow$ DM.encode(**List**($text_e$))
$latents \leftarrow$ DM.sample($seed$)
while $t \neq T$ **do**
  $n_\emptyset, n_p \leftarrow$ DM.predict-noise($latents, c_p$)
  List(n_e) $\leftarrow$ DM.predict-noise($latents$, **List**(c_e))
  $g = s_g * (n_p - n_\emptyset)$
  $i \leftarrow 0$
  for all n_e^i in **List**(n_e) **do**
    $\phi_t^i \leftarrow n_e^i - n_\emptyset$
    if negative guidance **then**
      $\phi_t^i \leftarrow -\phi_t^i$
    end if
    $\mu_t^i \leftarrow \mathbf{0}$
    $\mu_t^i \leftarrow$ where($|\phi_t^i| \geq \lambda_e^i, s_e^i$)
    $\gamma_t^i \leftarrow \mu_t^i \cdot \phi_t^i$
    $i \leftarrow i + 1$
  end for
  $\gamma_t \leftarrow \sum_{i \in I} g_i * \gamma_t^i$
  $\gamma_t \leftarrow \gamma_t + s_m * \nu_t$
  $\nu_{t+1} \leftarrow \beta_m * \nu_t (1 - \beta_m) * \gamma_t$
  if $t \geq \delta$ **then**
    $g \leftarrow g + \gamma_t$
  end if
  $pred \leftarrow n_\emptyset + g$
  $latents \leftarrow$ DM.update-latents($pred, latents$)
  $t \leftarrow t + 1$
end while

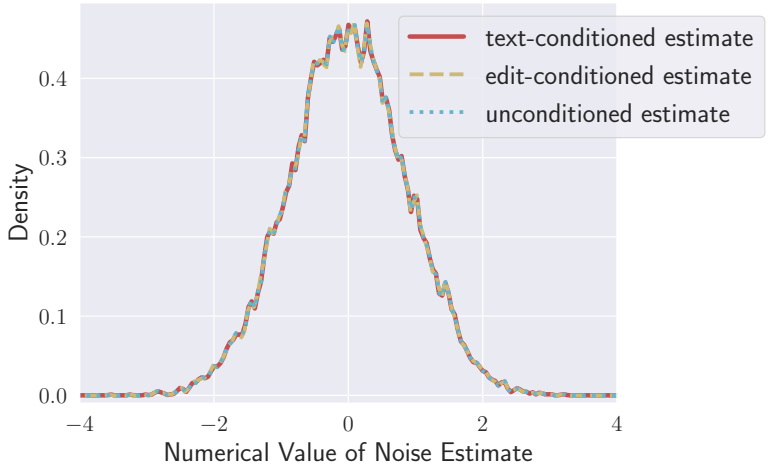


Figure 8: Distribution of numerical values in different noise estimates. All estimates follow a Gaussian distribution disregarding their different conditioning. Distribution plots calculated using kernel-density estimates with Gaussian smoothing. (Best viewed in color)

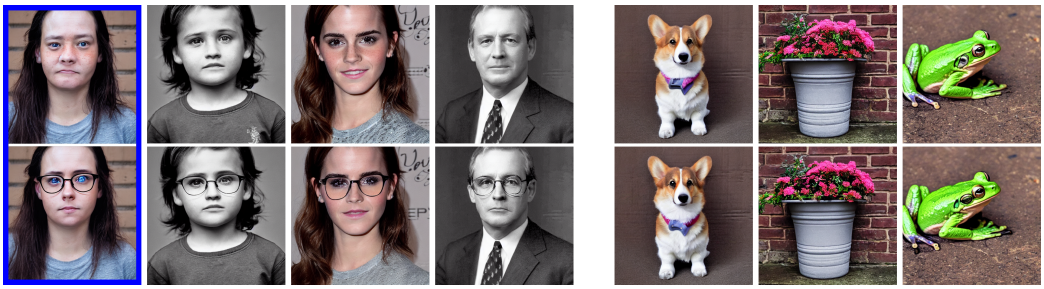


Figure 9: Uniqueness of guidance vectors. The vector for '*glasses*' is calculated **once** on the blue-marked image and subsequently applied to other prompts (without colored border). On images where no glasses are added the guidance vector produces no visual artifacts. (Best viewed in color)

C. Further Examples of Semantic Space's Properties

Here, we provide further examples of the properties of the semantic space created by SEGA (cf. Sec. 4). First, we revisit the uniqueness of guidance vectors. As shown in Fig. 4b these vectors may be calculated once and subsequently applied to other prompts. However, this is limited to the same initial noise latent and requires a similar image composition. We depict this difference in Fig. 9 where the guidance vector for '*glasses*' is successfully applied to three images and fails to add glasses on three other images. However, the unsuccessfully edited images on the right retain the same image fidelity showcasing that SEGA's guidance vectors do not introduce artifacts. Nonetheless, SEGA can easily incorporate glasses into these images by computing a dedicated guidance vector (cf. Sec. 4).

Next, we demonstrate that multiple concept vectors scale monotonically and independent of each other. Meaning that in a multi-conditioning scenario the magnitude of each concept can be targeted independently. We depict an example in Fig. 10. Starting from the original image in the top left corner, we remove one concept from the image following the x- or y-axis. In both cases, SEGA monotonically reduces the number of people or trees until the respective concept is removed entirely. Again, the rest of the image remains fairly consistent, especially the concept that is not targeted. Going even further, a similar level of control is also possible with an arbitrary mixture of applied concepts. Let us consider the second row of Figure 10, for example. The number of trees is kept at a medium level in that the row of trees on the left side of each image is always removed, but the larger tree on the right remains. While keeping the number of trees stable, we can still gradually remove the crowd at the same time.

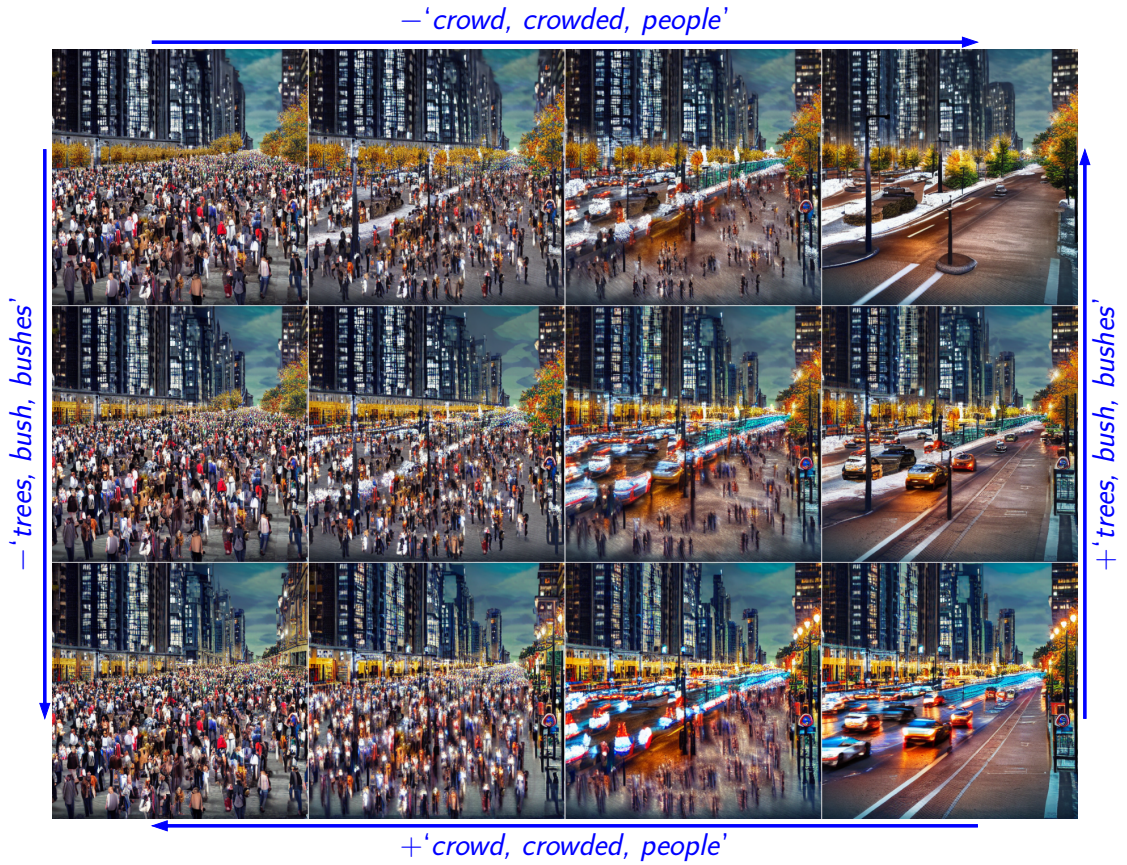


Figure 10: SEGA offers strong control over the semantic space and can gradually perform edits at the desired strength. All images generated from the same initial noise latent using the prompt 'a crowded boulevard'. Editing prompts denoted in blue and are gradually increased in strength from left to right and top to bottom. (Best viewed in color)

D. User Study

We subsequently provide further details on the empirical evaluation presented in Sec. 5.1. We attempted to label the generated images using a classifier trained on CelebA. However, despite the model achieving over 95% accuracy on the CelebA validation set, the accuracy for the synthetic images was for some classes below 20%. Consequently, we opted to rely on human evaluators instead. For each of the 250 generated images we collected labels for all 10 attributes. Questions were posed in the form

> Is the person wearing glasses?

The three answer options were:

- Yes
- No
- Cannot Tell

or in the case of the gender attribute:

- Male
- Female
- Cannot Tell

Each user was tasked with labeling a batch of 28 image/attribute pairs; 25 out of those were randomly sampled from our generated images and each batch contained 3 hand-selected images from CelebA as sanity check. If users labeled the 3 CelebA images incorrectly the batch was discarded and added back to the task pool. Each images/attribute combination was labeled by 3 different annotators resulting in annotator consensus if at least 2 selected the same label.

To conduct our study we relied on Amazon Mechanical Turk where we set the following qualification requirements for our users: HIT Approval Rate over 95% and at least 1000 HITs approved. Annotators were fairly compensated according to Amazon MTurk guidelines. Users were paid $0.70 for a batch of 28 images at an average of roughly 5 minutes needed for the assignment.

E. Further Qualitative Evaluation

Lastly, we provide further qualitative examples further highlighting the capabilities and use cases of SEGA.

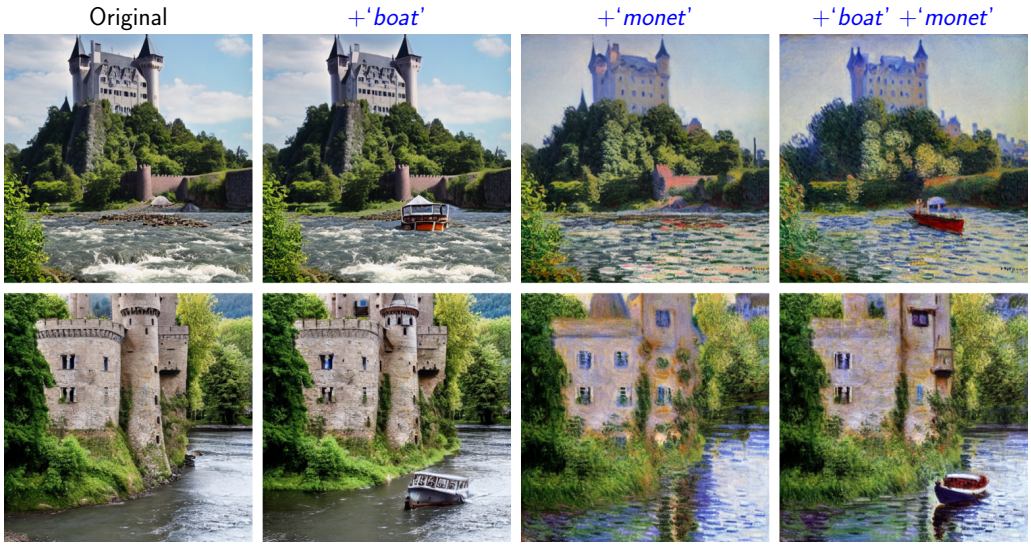


Figure 11: Robustness of SEGA across multiple tasks. Here style transfer and image composition are easily combined and executed simultaneously. The resulting image satisfies both tasks. All images generated using the prompt '*a castle next to a river*'. Top and bottom row use different seeds. (Best viewed in color)

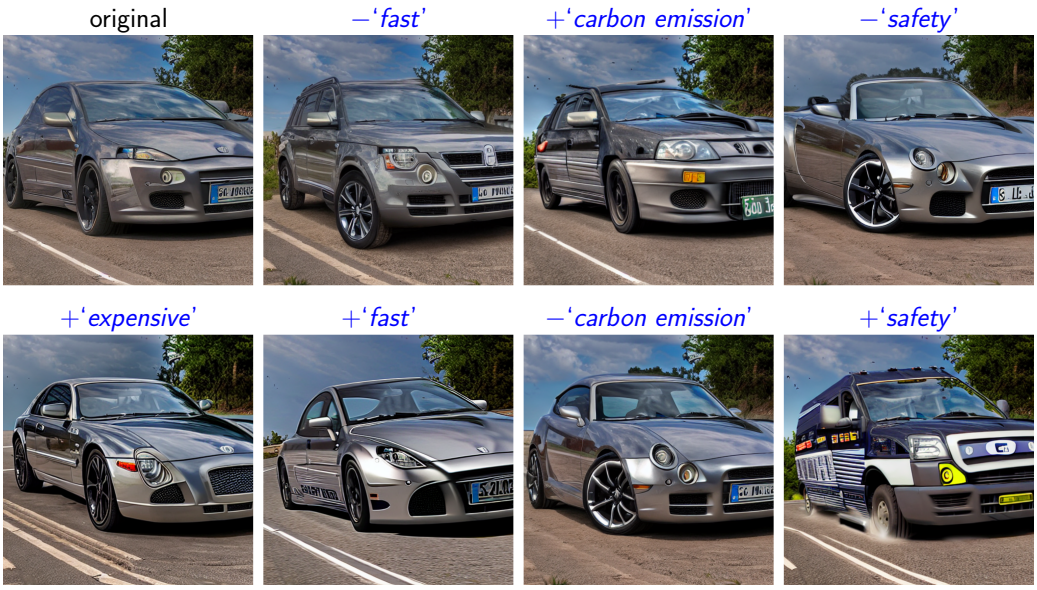


Figure 12: Probing the diffusion model for learned concepts. Guiding towards or away from more complex concepts may reveal the underlying representation of the concept in the model. All images generated using the prompt '*a picture of a car*'. (Best viewed in color)

The isolation of concept vectors (cf. Sec. 4) not only allows for multiple simultaneous edits of the same task but also arbitrary combinations. In Fig. 11 we simultaneously change the image composition and change the style. The resulting image is a faithful interpolation of both changes applied individually. This further highlights both the robustness of SEGA guidance vectors as well as their isolation.

SEGA also contributes in uncovering how the underlying DM "interprets" more complex concepts and gives further insight into learned representations. We perform this kind of probing of the model in Fig. 12. For example, adding the concept '*carbon emissions*' to the generated car produces a seemingly much older vehicle with a presumably larger carbon footprint. Similarly, reducing '*safety*' yields a convertible with no roof and likely increased horsepower. Both these interpretations of the provided concepts with respect to cars are logically sound and provide valuable insights into the learned concepts of DMs. These experiments suggest a deeper natural language and image "understanding" that go beyond descriptive captions of images.