SEMANTIC SELF-FLOW: RELATION LEARNING IN SELF-SUPERVISED FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow models still struggle to capture image semantics, such as spatial relations between objects. Recently, Self-Flow showed that self-distillation across different noise levels for the student and teacher views can reduce reliance on local evidence for reconstruction and improve semantic representations. However, a random student mask does not target any particular semantic dependency; for example, it leaves most of the subject at the lower noise level, revealing its location, and hence the spatial relation, without requiring the relation condition. We introduce Semantic Self-Flow (SSF), which places stronger noise where recovery depends on the semantic constraint of interest. For spatial relations, the subject is noised together with three alternative locations around the reference object, which remains at the lower noise level, making the relation condition informative for reconstruction. On ImageNet, when the relation is flipped under the same noise, both SSF outputs follow their respective relation conditions for 32% of test prompts, compared with 2% for Self-Flow, while class-conditional FID remains nearly unchanged.

1 INTRODUCTION

Diffusion and flow models (Rombach et al., 2022; Lipman et al., 2023) have shown remarkable success in generating high-quality images. However, they often fall short in capturing scene semantics (Schwartz et al., 2023; Zafar et al., 2026). For instance, relation words that describe the relative positions of objects, e.g., *a dog left of a bicycle*, are often ignored (Huang et al., 2023; Ghosh et al., 2023; Yiflach et al., 2026).

One reason is that the denoising objective can be satisfied by relying on local evidence, without capturing the global semantics of the scene. To this end, REPA (Yu et al., 2025) proposed aligning the denoiser's features with a pretrained semantic representation such as DINOv2 (Oquab et al., 2024), which significantly improves generation, and recently, Self-Flow (Chefer et al., 2026) removed the need for a pretrained representation by assigning a different noise level to a random subset of tokens and training the student to match the features of a teacher that sees the whole image at the lower noise level (Figure 1a).

The motivation is information asymmetry, i.e., to match the teacher, the student must infer the noisier tokens from the cleaner context. However, *Self-Flow's random mask does not specify which information the student should recover.* A randomly chosen token can often be predicted from nearby texture, and thus nothing in the objective ties the recovery to a particular semantic property, such as the layout of two objects. This raises the question of whether placing the asymmetry, rather than choosing it at random, can make a chosen piece of information necessary for recovery.

To study this question, we introduce Semantic Self-Flow (SSF), a method that keeps the architecture and objective of Self-Flow but changes where the stronger noise goes, to encourage the denoising to rely on the semantic signal in question, in our case, the spatial relation between a subject and a reference object. For the model to learn the semantics of relations, the noise should be placed such that the relation (i) remains valid, i.e., the reference object, with respect to which the relation is defined, stays at the lower noise level, and (ii) is not trivial to infer, i.e., most of the subject receives the higher noise level. Self-Flow's random mask ensures neither condition, e.g., it noises the reference as often as any other object, and it leaves most of the dog in Figure 1a at the lower noise level, which reveals the location of the subject, and thus the student can infer the relation without

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

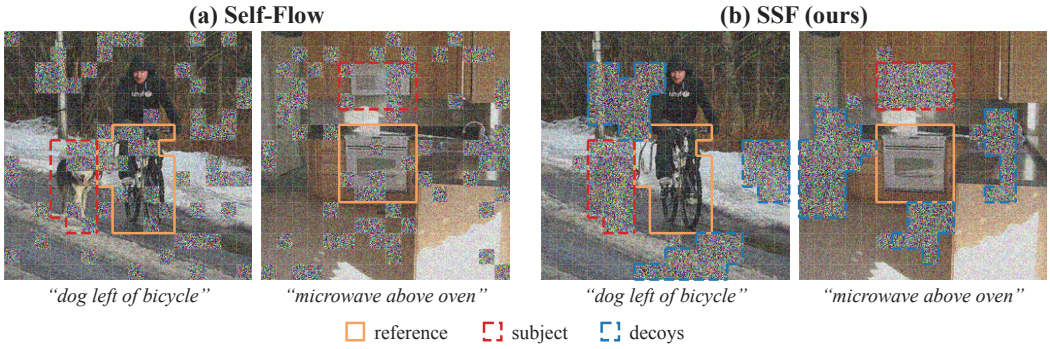


Figure 1: Illustration of Self-Flow's random noise placement and our method for two ImageNet training images. (a) Self-Flow assigns a second noise level to random tokens, including tokens from the reference object. (b) SSF noises the subject together with three decoys, i.e., regions of subjects in other training images, placed according to the three alternative relations, while the reference object remains at the lower noise level. Thus, the mask alone hardly indicates which noisy region corresponds to the subject.

taking the relation condition into account. Noising the entire subject, however, makes the location of the noise itself reveal the relation, and makes the subject hard to reconstruct.

To meet both conditions, SSF noises most of the subject together with decoys at the positions a subject would take under the other three relations, while the reference object stays at the lower noise level. As a result, the mask reveals little about the location of the subject, and also keeps the relation valid, which encourages the model to use the relation to the clean reference to tell which noisy region holds the subject (Figure 1b). We obtain the relation labels from an automatic pipeline that annotates 8.2% of ImageNet (Deng et al., 2009), and apply this mask only to these images, while the remaining images keep the random mask of Self-Flow, from which the model continues to learn the appearance and the natural layout of objects.

We evaluate SSF on synthetic shapes, where mixed sampling proves necessary, and on ImageNet and in text-to-image training, where each test prompt is generated with the original and with the flipped relation from the same noise. On ImageNet, both images follow the relation for 37% of the test prompts with SSF, compared to 3% with Self-Flow, which generates nearly the same layout regardless of the relation condition, and a larger model shows a similar gain (32% vs. 2%) with a nearly unchanged class-conditional FID. In text-to-image training, SSF also follows the relation in the caption more often, i.e., both images are correct for 15.4% of the captions, compared to 9.3%.

Our contributions are as follows.

- We show that random masking in Self-Flow leaves semantic properties such as spatial relations to chance, and thus the model largely ignores the relation condition.
- We propose Semantic Self-Flow (SSF), a simple yet effective method that places the stronger noise where recovery depends on the semantic property of interest.
- Our experiments on ImageNet and text-to-image training show that SSF improves relation learning over Self-Flow with a nearly unchanged class-conditional FID.

2 Related Work

Representation learning for generation Diffusion and flow models learn semantic structure slowly from the denoising objective alone. REPA (Yu et al., 2025) showed that aligning their features with a pretrained encoder, such as DINOv2 (Oquab et al., 2024), significantly accelerates training, and follow-up work refined this alignment (Singh et al., 2026; Wang et al., 2025), extended it to the VAE (Leng et al., 2025; Yao et al., 2025), or generated a semantic latent from the encoder (Zheng et al., 2026; Pan et al., 2026). All these methods depend on an external encoder. SRA (Jiang et al., 2026), LayerSync (Haghighi et al., 2026), and Self-Flow (Chefer et al., 2026) remove this depen-

dency by aligning the model with its own features, where Self-Flow trains a student to match an EMA teacher that sees a cleaner view of the image, following DINO (Caron et al., 2021). SSF builds on Self-Flow, and thus aligns the model with its own features rather than with an external encoder, whereas the noise placement relies on automatically annotated object locations.

Masking and heterogeneous noise Recovering hidden content from the visible context is a classic route to representation learning, from denoising autoencoders (Vincent et al., 2010) to MAE (He et al., 2022) and I-JEPA (Assran et al., 2023). In diffusion training, MaskDiT (Zheng et al., 2024) and MDT (Gao et al., 2023) mask random tokens to reduce the training cost and to learn relations between object parts, respectively, and later work assigned different noise levels to different tokens, e.g., along time in Diffusion Forcing (Chen et al., 2024), across attributes in ComboStoc (Xu et al., 2026), and to a random subset of tokens in Self-Flow. In all these methods, the masked or noisier tokens are chosen independently of the condition. In contrast, SSF chooses them according to the condition, i.e., it noises the subject and counterfactual decoys more strongly than the reference, which encourages the model to use the relation.

Spatial relations and compositional generation Despite their realism, text-to-image models often fail to depict spatial relations (Conwell & Ullman, 2022; Gokhale et al., 2022; Huang et al., 2023; Ghosh et al., 2023), a weakness shared by vision-language models (Kamath et al., 2023). Existing remedies optimize the latents at inference time with attention-based losses, e.g., Attend-and-Excite (Chefer et al., 2023) and Learn-to-Steer (Yiflach et al., 2026), rely on layouts, e.g., GLIGEN (Li et al., 2023b) and LMD (Lian et al., 2024), or fine-tune the model with relation-specific data, modules, or rewards, e.g., SPRIGHT (Chatterjee et al., 2024), CoMPaSS (Zhang et al., 2025), Relation Rectification (Wu et al., 2024), and Flow-GRPO (Liu et al., 2025). In contrast, SSF uses automatically annotated object locations only to decide where the noise is placed during training, and requires no layout or optimization at inference time.

3 METHOD

Our method builds directly on Self-Flow (Chefer et al., 2026). Both methods receive the same condition, e.g., on ImageNet, the class label and a relation prompt such as *dog left of bicycle*.

3.1 PRELIMINARIES: SELF-FLOW

Given a data sample x_0, i.e., N latent tokens x_0^i, and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, flow matching (Lipman et al., 2023) defines

$$x_\tau = (1-\tau)x_0 + \tau\epsilon, \qquad \tau \in [0,1], \tag{1}$$

and trains a network f_θ that predicts the velocity $\epsilon - x_0$ from x_τ, τ, and a condition c, by minimizing

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\|f_\theta(x_\tau, \tau, c) - (\epsilon - x_0)\|_2^2\right]. \tag{2}$$

where the expectation is over the training data, the noise, and the noise level.

Self-Flow augments this objective with self-supervised feature alignment. It samples two noise levels $t_1, t_2 \sim \mathcal{U}[0,1]$ and a random subset $\mathcal{M}$ of about ρN tokens, with a masking ratio $\rho = 0.25$. Each token then receives its own noise level,

$$\tau_i = \begin{cases} t_2, & i \in \mathcal{M}, \\ t_1, & i \notin \mathcal{M}, \end{cases} \qquad x^i = (1-\tau_i)\, x_0^i + \tau_i\, \epsilon^i, \tag{3}$$

i.e., the mask selects which tokens receive a different noise level, rather than removing them. The network then predicts $f_\theta(x, \tau_{1:N}, c)$ with the noise levels $\tau_{1:N} = (\tau_1, \ldots, \tau_N)$, which it receives through a per-token timestep embedding, while the velocity target in Equation (2) remains $\epsilon - x_0$. The student sees this mixed-noise view (Figure 2c), while the teacher $f_{\theta'}$, an exponential moving average (EMA) of the student, sees the entire image with the same Gaussian noise sample ϵ, but at the uniform lower level $\tau_{\min} = \min(t_1, t_2)$. The student is trained to match the teacher's features by minimizing

$$\mathcal{L}_{\text{rep}} = -\frac{1}{N}\sum_{i=1}^{N} \cos\big(h_{\text{s}}^i, \text{sg}(h_{\text{t}}^i)\big), \tag{4}$$

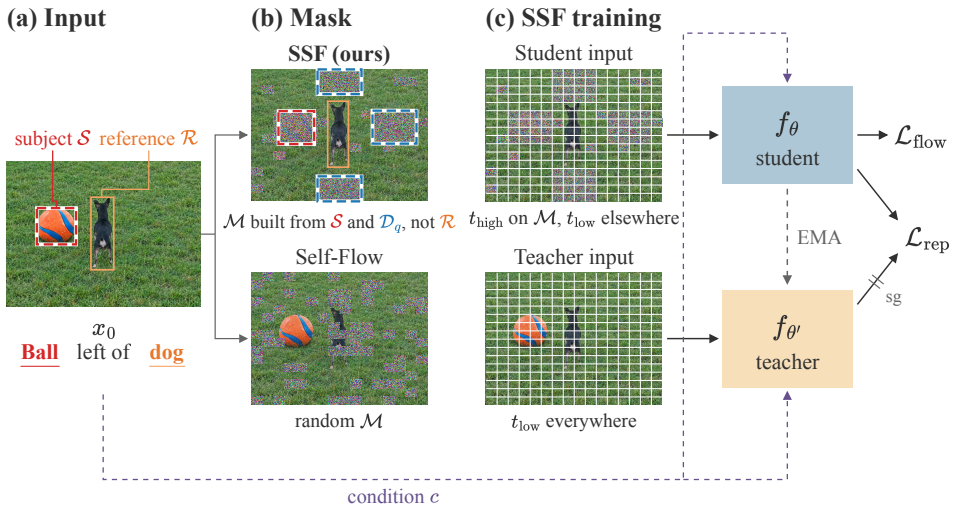


Figure 2: Method illustration. (a) A training image with the relation *ball left of dog*. (b) Self-Flow masks random tokens, whereas SSF builds the mask from the subject and its decoys, with the same budget. (c) Training of SSF, where the student is trained with the Self-Flow objective and the teacher is its EMA. See Section 3 for details.

where h_{s}^i is the projected student feature and h_{t}^i is the teacher feature of token i, both taken from intermediate layers, cos denotes the cosine similarity, and sg denotes stop-gradient. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda \mathcal{L}_{\text{rep}}, \tag{5}$$

where λ weights the representation loss. Note that only the student is trained with this objective, while the teacher is updated through the EMA.

Intuitively, the mixed-noise view encourages the noisier tokens to rely on a broader image context rather than on local cues, which are easier to exploit for reconstruction. However, a random mask does not specify which context the noisier tokens should use.

3.2 Relation Conditioning

Class-conditioned diffusion transformers add a class embedding to the timestep embedding. We condition on a relation prompt in the same way, i.e., we embed its subject s, relation r, and reference object o, e.g., *dog left of bicycle*, with learned linear embedders E_s, E_r, and E_o, and add them to the class embedding $E_y(y)$,

$$\mathbf{c}_i = \phi(\tau_i) + E_y(y) + E_s(s) + E_r(r) + E_o(o), \tag{6}$$

where ϕ is the timestep embedding, and $\mathbf{c}_i$ modulates every block for token i through adaptive layer normalization (adaLN) (Peebles & Xie, 2023). In text-to-image training, the relation is already part of the caption, which the model attends to. Appendix A provides the full architecture details.

3.3 Semantic Self-Flow

Semantic Self-Flow (SSF) chooses the mask $\mathcal{M}$, i.e., where the stronger noise goes, such that recovering the noisier tokens encourages the model to use the *semantic signal* of interest, while limiting shortcuts.

3.3.1 Relation-Aware Masking

For spatial relations, the semantic signal is the relation $r \in \{\text{left}, \text{right}, \text{above}, \text{below}\}$ between a subject and a reference object, e.g., *ball left of dog* (Figure 2(a)). Let $\mathcal{S}$ and $\mathcal{R}$ denote the tokens covered by the subject and by the reference object, respectively. Note that these locations are used only to construct the mask during training, and are not required at inference.

First, the relation must remain valid, and hence the reference stays at the lower noise level, since the relation is defined with respect to it,

$$\mathcal{M} \cap \mathcal{R} = \emptyset. \tag{7}$$

The model can then combine the position of the reference with the relation condition to tell which noisy region holds the subject.

Second, the relation must not be trivial to infer, i.e., most of the subject should receive the higher noise level. However, if only the subject is masked, the location of the mask itself reveals the relation, e.g., on COCO pairs (Lin et al., 2014), a classifier predicts the relation from the reference box and such a mask for 93% of the pairs, compared to 63% from the reference box alone, and 60% when the decoys described next are added (Appendix A.6). SSF therefore adds one decoy $\mathcal{D}_q$ for each of the three alternative relations $q \neq r$, i.e., token locations where a subject could appear under relation q (Figure 2(b) and Appendix A.3), and builds the mask from the candidate tokens

$$\mathcal{C} = \Big(\mathcal{S} \cup \bigcup_{q \neq r} \mathcal{D}_q\Big) \setminus \mathcal{R}. \tag{8}$$

On ImageNet and in text-to-image training, each annotated pair is used in both directions, i.e., *ball left of dog* also appears as *dog right of ball*, and thus every annotated object is also masked as a subject.

Finally, SSF keeps the masking budget k of Self-Flow, i.e., the size of the Self-Flow mask of the same example, and sets

$$\mathcal{M} = \begin{cases} \mathcal{C} \cup \mathcal{F}, & |\mathcal{C}| \leq k, \\ \mathcal{U}_k(\mathcal{C}), & |\mathcal{C}| > k, \end{cases} \tag{9}$$

where $\mathcal{U}_k(\mathcal{C})$ is a uniformly random subset of k tokens of $\mathcal{C}$, and $\mathcal{F}$ fills the remaining $k - |\mathcal{C}|$ tokens with the one-token border around $\mathcal{C}$ and then with random tokens, both outside $\mathcal{R}$. Hence, the subject is not favored over the decoys, and large subjects are noised only partially.

Noise levels SSF keeps the architecture and the objective of Self-Flow (Figure 2(c)). For the noise levels, we replace Equation (3) with

$$\tau_i = \begin{cases} t_{\text{high}}, & i \in \mathcal{M}, \\ t_{\text{low}}, & i \notin \mathcal{M}, \end{cases} \qquad t_{\text{low}} \sim \mathcal{U}[0,1], \quad t_{\text{high}} \sim \mathcal{U}[t_{\text{low}}, 1], \tag{10}$$

and show the teacher the full image at t_{low}. Unlike in Self-Flow, where the masked tokens receive the lower noise level in half of the examples, the mask always receives the higher level, since otherwise the subject and the decoys would be cleaner than the reference, which reverses both conditions. Appendix A provides the remaining details.

Mixed sampling SSF applies the relation-aware mask only to a fraction of the training images, i.e., to the 8.2% of the ImageNet images with annotated relations, to the 25% of the text-to-image examples that are spatial pairs, and to 20% of the images on synthetic shapes, and trains all remaining images exactly as Self-Flow, i.e., with a random mask and the noise levels of Equation (3). This mix is part of the design, since the relation-aware mask teaches the model to use the relation condition, whereas the remaining images keep teaching it the appearance and the natural layout of the objects. Our ablation on synthetic shapes shows that this mix is necessary (Appendix F).

Annotations We annotate the relations automatically with a vision–language model, i.e., Qwen3-VL (Bai et al., 2025), an open-vocabulary detector, i.e., OWLv2 (Minderer et al., 2023), and a segmentation model, i.e., SAM (Kirillov et al., 2023), which yields annotations for 8.2% of the ImageNet training images, while the remaining images are trained as in the baseline (Appendix A.2). In our human evaluation, OWLv2 (Minderer et al., 2023) with the center rule agrees with the majority vote of the raters on whether a generated image shows its caption for 84% of the images (Appendix D).

4 Experiments

In the following, we evaluate SSF in class-conditioned generation on ImageNet (Section 4.1) and in text-to-image generation (Section 4.2).

Table 1: Comparison to Self-Flow on ImageNet. Generation accuracy is the share of test prompts whose image shows the relation, for the original prompt, the flipped prompt, and both. Sensitivity is the change of the prediction when only the relation is flipped. FID and sFID use unguided class-conditional samples.

		Generation acc. (%)			Sensitivity	Image quality	
Model	Method	Original ↑	Flipped ↑	Both ↑	Δv (%) ↑	FID ↓	sFID ↓
SiT-B/2	Self-Flow	43.0	17.0	3.0	0.5	20.10	6.30
	SSF (ours)	**68.0**	**48.0**	**37.0**	**7.7**	**19.92**	**6.25**
SiT-XL/2	Self-Flow	45.0	19.0	2.0	1.1	**7.45**	5.40
	SSF (ours)	**63.0**	**46.0**	**32.0**	**10.0**	7.58	**5.08**

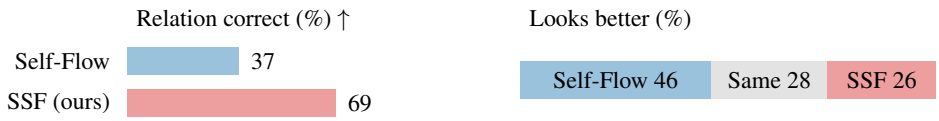


Figure 3: Human evaluation on ImageNet with SiT-XL/2.

In the appendix, we also ablate the share of training images with the relation-aware mask on synthetic shapes, where only the positions of the two objects predict the relation, and find that applying it to all images lowers the accuracy far below that of Self-Flow, whereas applying it to 20% of the images learns the relation faster than Self-Flow, i.e., mixed sampling is necessary (Appendix F). Moreover, we apply SSF to counting, where it reaches 74.5% at 20k steps compared to 64.5% for Self-Flow (Appendix G), and we answer further questions, e.g., whether a mask without decoys reveals the relation (Appendix A.6), whether the model uses the relation condition, and in the right direction, on real images (Appendix B), and whether the gains hold across guidance scales and for unseen category pairs (Appendix C).

4.1 Class-conditioned generation

Setup Following Self-Flow, we train SiT-B/2 (Ma et al., 2024) (138M parameters) for 800k steps and SiT-XL/2 (687M) for 1.3M steps on ImageNet (Deng et al., 2009) at 256×256. Both methods share the conditioning, data order, noise samples, and optimizer, with one seed per method. Appendix A.4 provides the remaining details.

Evaluation We evaluate along three axes. **Generation accuracy** checks with OWLv2 (Minderer et al., 2023) and the position rule of our annotations whether a generated image shows the requested relation. Since an image may show the relation only because it is the natural layout, e.g., *lamp above table*, we sample each of 100 test prompts from the same noise with the original and the flipped relation, where both images can be correct only if the model follows the relation condition. Half of the prompts have *seen* category pairs, which co-occur in the training annotations, e.g., cup and laptop, and half have *unseen* ones. **Sensitivity** is the relative change of the predicted velocity on the same real image when only the relation is flipped, which is zero for a model that ignores the relation, on 307 annotated validation images with seen pairs (Appendix B.1). **Image quality** follows Self-Flow, i.e., FID and sFID of 50k unguided class-conditional samples (Appendix A.7).

Results In Table 1, we show that with SiT-B/2, SSF generates the requested relation considerably more often than Self-Flow, i.e., for 68% of the test prompts compared to 43% with the original relation, and for 48% vs. 17% with the flipped one. Importantly, for 37% of the test prompts, both of its images show the requested layout, compared to 3% for Self-Flow (a gain of 34 points). With SiT-XL/2, the gain is similar, i.e., both images are correct for 32% of the prompts, compared to 2%. For both models, the gain holds for unseen category pairs as well (34% vs. 2% and 30% vs. 4%, Appendix C). This indicates that Self-Flow generates nearly the same layout regardless of the relation condition, whereas SSF follows it.

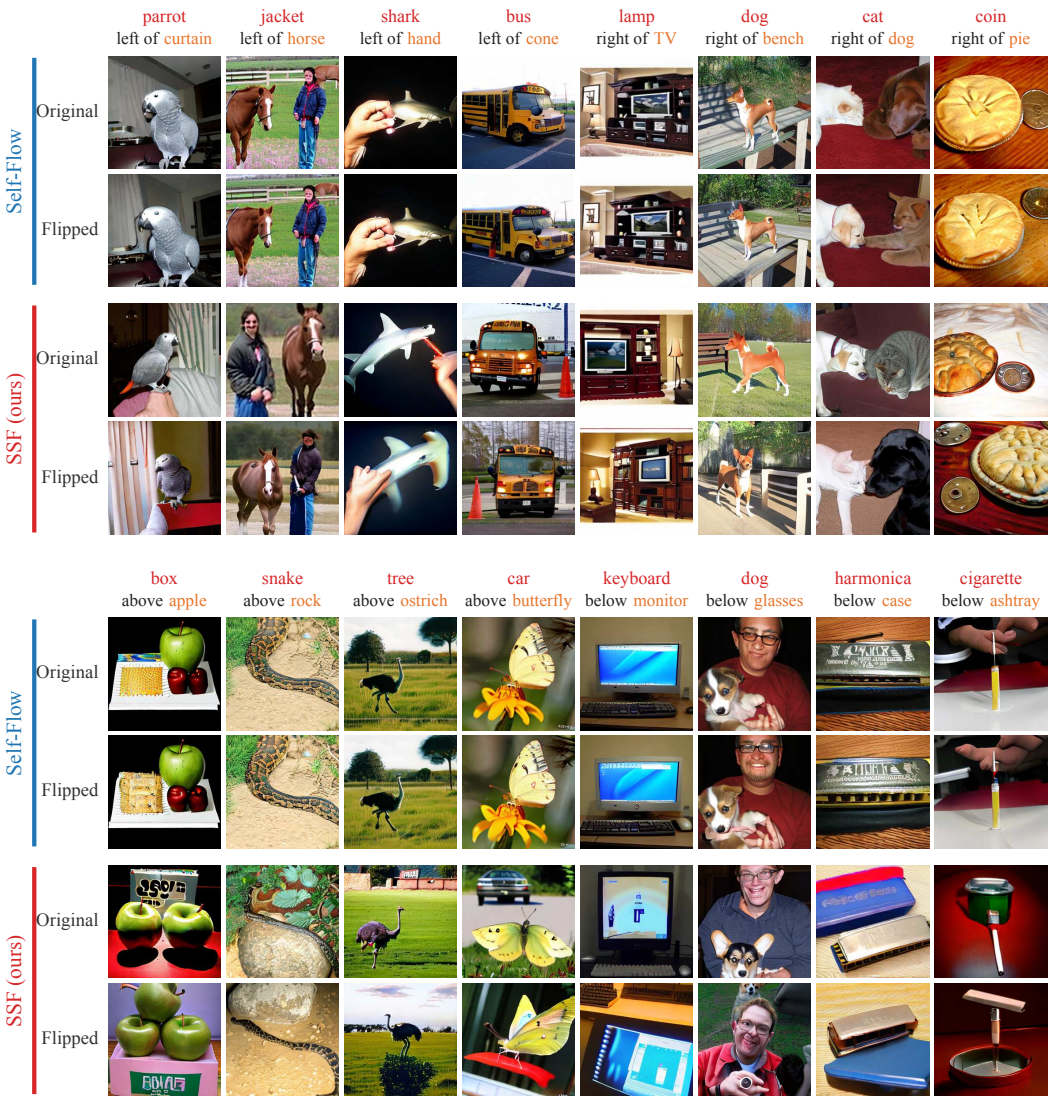


Figure 4: ImageNet samples of SiT-XL/2 for the original prompt and the flipped one, which has the opposite relation, e.g., *right of* instead of *left of*, from the same noise with the guidance of Table 1. The subject is in red and the reference object in orange.

The sensitivity confirms this observation, i.e., flipping the relation condition changes the prediction of SSF 15 times more than that of Self-Flow with SiT-B/2, and nine times more with SiT-XL/2. Note that Self-Flow does use its condition, since swapping the category of the subject changes its prediction 18 and 10 times more than flipping the relation, respectively (Appendix B.1). With SiT-B/2, we also observe that in distinguishing the true from the flipped relation in real images, Self-Flow remains near the chance level of 50% for 800k steps, whereas SSF reaches 75% on seen and 73% on unseen category pairs (Appendix B.2).

The class-conditional FID and sFID are nearly unchanged, i.e., with SiT-B/2, both the FID and the sFID improve slightly, and with SiT-XL/2, the FID increases by 0.13, the sFID improves, and precision and recall are unchanged (0.70 and 0.66 for both).

Human evaluation We also ask university students to compare the SiT-XL/2 images of Self-Flow and SSF for the same prompt and noise, i.e., in which image the relation is correct, and which image looks better, with 411 ratings in total (Appendix D).

In Figure 3, we show that raters find the relation correct in the image of SSF about twice as often as in that of Self-Flow (69% vs. 37% of the ratings, and 66% vs. 28% for the flipped prompt). Raters prefer the look of Self-Flow in 46% of the ratings and that of SSF in 26%, and find them about the same in 28%. Note also that a head-to-head comparison can show a large gap in preference even for small differences in quality. On inspection, this preference often arises because SSF draws the unusual layout that the prompt requests, e.g., *keyboard above monitor*, whereas Self-Flow ignores the relation and draws the same natural scene for both prompts, as the qualitative results below show.

Qualitative results In Figure 4, we show four prompts for each relation, and observe that Self-Flow often retains the same layout for the original and the flipped prompt, whereas SSF rearranges the objects to follow the relation while preserving the overall scene, e.g., for *jacket left of horse*, it places the person on the left of the horse for the original prompt and on the right for the flipped one. For *above* and *below*, the flipped prompt often asks for an unusual layout, e.g., *keyboard above monitor* or *dog above glasses*, which SSF draws while Self-Flow keeps the natural scene, and hence the images of SSF can look less natural. Appendix E shows uncurated samples.

4.2 TEXT-TO-IMAGE

We next repeat the comparison with caption conditioning, where the relation is a word in the caption, e.g., *a cup to the left of a laptop*. We report the results after 200k steps in Table 2, and earlier checkpoints in Appendix C. Appendix A.5 provides the training details.

Evaluation We use the tests of Section 4.1, where the relation is now flipped in the caption, e.g., from *left of* to *right of*, on 1,186 held-out captions of COCO val2017 images (Lin et al., 2014), which were not used for training. For the **generation accuracy**, Qwen3-VL (Bai et al., 2025) judges the generated images. Since the judge may accept the same image for both relations, e.g., when a scene contains several instances of an object, an image counts as correct only if the judge accepts it for its own caption and rejects it for the opposite one, and **both** requires the two images from the same noise to be correct. The **sensitivity** is the relative change of the prediction on real images when only the relation in the caption is flipped, and the **FID-5k** compares 5,000 unguided samples with COCO val2017 images (Appendix A.7).

Table 2: Comparison to Self-Flow in text-to-image training at 200k steps. Generation accuracy is the share of captions whose image the judge accepts for its caption and rejects for the opposite one. Sensitivity is as in Table 1, and FID-5k uses unguided samples.

	Generation acc. (%)			Sensitivity (%)	Image quality
Method	Original ↑	Flipped ↑	Both ↑	Relation ↑	FID-5k ↓
Self-Flow	32.0	26.8	9.3	7.2	26.50
SSF (ours)	**38.9**	**36.8**	**15.4**	**9.2**	**24.93**

Results In Table 2, we show that at 200k steps, SSF generates the requested relation more often than Self-Flow, for both the original and the flipped caption (38.9% vs. 32.0% and 36.8% vs. 26.8%). Importantly, both images are correct for 15.4% of the captions, compared to 9.3% (a gain of 6.1 points), mostly for *left of* and *right of*. Moreover, without the judge, the flipped image is nearly identical to the original one for 30.9% of the captions with Self-Flow, compared to 20.2% with SSF, i.e., Self-Flow ignores the relation word more often, as illustrated in Figure 5 (see Appendix E for uncurated samples). SSF also responds more strongly to the relation (9.2% vs. 7.2%). The FID-5k also improves (24.93 vs. 26.50).

5 DISCUSSION

SSF introduces an inductive bias, i.e., it uses annotations to decide where the stronger noise goes. Random masking does not explicitly encourage the reconstruction to depend on the relation condition, since the context at the lower noise level often reveals the property, e.g., most of the subject

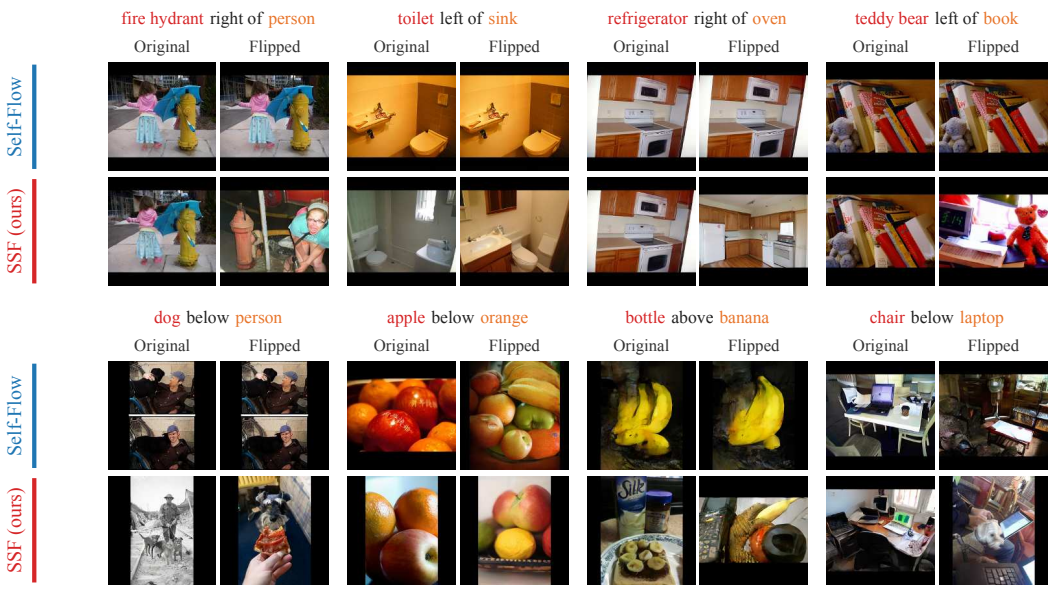


Figure 5: Text-to-image samples for the original caption and the flipped one, which has the opposite relation, from the same noise. The subject is in red and the reference object in orange.

remains at the lower noise level and reveals its location, and hence the relation. Moreover, a random mask noises the reference as often as any other object, although the relation is defined with respect to it. SSF targets this shortcut by corrupting the subject together with plausible alternatives while keeping the reference clean, which encourages the model to locate the subject relative to the reference through the relation condition.

This may explain why generative models often fail on semantic properties such as spatial relations and counting (Huang et al., 2023; Zafar et al., 2026).

6 Limitations

Our compute is modest, and thus we could not train the models to convergence, e.g., we train SiT-XL/2 for 1.3M steps, whereas Self-Flow trains it for 4M steps (Appendix A.1), and hence our results may not fully capture the behavior of SSF at convergence. Unfortunately, for the same reason, we could run only a few ablations, and a full study was possible only for spatial relations, although our counting results indicate that SSF may also help with other semantic properties (Appendix G).

In addition, both the annotations that SSF uses for training and our evaluation are automatic, and hence depend on the quality of the detector and of the pseudo-labels of the vision–language model.

7 Conclusion

We introduced Semantic Self-Flow (SSF), which places the stronger noise of Self-Flow where recovery depends on a chosen semantic property. We find that the random mask of Self-Flow makes semantic properties such as spatial relations hard to learn, since it may noise the reference and thus break the relation, and it often leaves the subject visible and thus poses too little challenge. Our relation-aware mask improves relation learning in all our settings, with a nearly unchanged class-conditional FID. We hope that SSF encourages future work on other semantic properties, such as counting.

AI USE STATEMENT

We used AI assistants to run experiments, to help revise parts of the paper, to ease the reporting of exact results, and to illustrate figures. The appendix relies on AI assistance much more than the main paper, since we asked the assistants to report every part of our experiments exactly. We believe that this form of reporting suits an appendix, since other AI agents could also use it to understand all implementation details. The ideas in the paper and the experimental setups are our own, and the authors take full responsibility for the content of the paper.

ETHICS STATEMENT

Our work improves how image generators follow semantic properties of the condition, such as spatial relations. Like other generative models, such models could be misused to create misleading images, and SSF does not introduce risks beyond those of existing generators. We train only on publicly available data, i.e., ImageNet, COCO, and spatial pairs derived from COCO, together with synthetic shapes. Our human evaluation asked volunteers to judge generated images, and it collected no personal information beyond an optional name. Our relation annotations are produced automatically by pretrained models and may inherit their biases and errors, and we therefore release them to allow their inspection.

REPRODUCIBILITY STATEMENT

We will share all code and data used in our experiments, including the relation annotations. Appendix A details the training and evaluation settings, and Table 3 lists the hardware and the training time of every experiment.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, and Yezhou Yang. Getting it Right: Improving Spatial Consistency in Text-to-Image Models. In *European Conference on Computer Vision (ECCV)*, 2024.

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 Transactions on Graphics (SIGGRAPH)*, 2023.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.06507.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Colin Conwell and Tomer Ullman. Testing Relational Understanding in Text-Guided Image Generation. *arXiv preprint arXiv:2208.00005*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked Diffusion Transformer is a Strong Image Synthesizer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking Spatial Relationships in Text-to-Image Generation. *arXiv preprint arXiv:2212.10015*, 2022.

Yasaman Haghighi, Bastien van Delft, Mariam Hassan, and Alexandre Alahi. LayerSync: Self-aligning Intermediate Layers. In *International Conference on Learning Representations (ICLR)*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Yujin Jeong, Arnas Uselis, Seong Joon Oh, and Anna Rohrbach. Diffusion Classifiers Understand Compositionality, but Conditions Apply. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. Representation Alignment for Diffusion Transformers without External Components. In *International Conference on Learning Representations (ICLR)*, 2026.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for End-to-End Tuning of Latent Diffusion Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yueming Pan, Ruoyu Feng, Qi Dai, Yuqi Wang, Wenfeng Lin, Mingyu Guo, Chong Luo, and Nanning Zheng. Semantics Lead the Way: Harmonizing Semantic and Texture Modeling with Asynchronous Latent Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Idan Schwartz, Vésteinn Snæbjarnarson, Hila Chefer, Serge Belongie, Lior Wolf, and Sagie Benaim. Discriminative class tokens for text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22668–22678, 2023.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for Representation Alignment: Global Information or Spatial Structure? In *International Conference on Learning Representations (ICLR)*, 2026.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research (JMLR)*, 11:3371–3408, 2010.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, Kai Wang, and Yang You. REPA Works Until It Doesn't: Early-Stopped, Holistic Alignment Supercharges Diffusion Training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation Rectification in Diffusion Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Rui Xu, Jiepeng Wang, Hao Pan, Yang Liu, Xin Tong, Shiqing Xin, Changhe Tu, Taku Komura, and Wenping Wang. ComboStoc: Combinatorial Stochasticity for Diffusion Generative Models. *ACM Transactions on Graphics (SIGGRAPH)*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Sapir Esther Yiflach, Yuval Atzmon, and Gal Chechik. Data-driven loss functions for inference-time optimization in text-to-image. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Oz Zafar, Yuval Cohen, Lior Wolf, and Idan Schwartz. Detection-driven object count optimization for text-to-image diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1885–1894, 2026.

Gaoyang Zhang, Bingtao Fu, Qingnan Fan, Qi Zhang, Runxing Liu, Hong Gu, Huaqi Zhang, and Xinguo Liu. CoMPaSS: Enhancing Spatial Understanding in Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast Training of Diffusion Models with Masked Transformers. *Transactions on Machine Learning Research (TMLR)*, 2024.

APPENDIX

A Experimental setup and compute 15
We describe the training and evaluation details of every experiment, the relation annotations, the hardware, and the training time.

B Does the model use the relation condition? 19
We measure whether the model uses the relation condition, and whether it uses it in the right direction on real images.

C Additional results 21
We report generation results for different guidance scales, the relation accuracy when both objects appear, and the text-to-image results during training.

D Human evaluation 24
We describe the blind human study of relation following and image quality.

E Additional qualitative results 25
We show uncurated samples for randomly drawn ImageNet prompts and text-to-image captions.

F Synthetic shapes 30
We describe the synthetic data, the masks, and the results for every share of images with the relation-aware mask.

G Counting on synthetic shapes 32
We apply SSF to counting on synthetic shapes, and describe the masks of each method and the results during training.

A EXPERIMENTAL SETUP AND COMPUTE

Within each experiment, Self-Flow and SSF use the same data, seeds, architecture, and optimizer, and differ only in the corruption of the annotated images.

A.1 COMMON SETTINGS AND COMPUTE

All models are SiT networks (Ma et al., 2024) trained with the Self-Flow objective and a masking budget of 25% of the tokens. Unless stated otherwise, we use AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, no weight decay, a constant learning rate, gradient clipping at a norm of 1, and bf16 mixed precision. We evaluate the EMA weights, which also serve as the teacher, with a decay of 0.9999 on ImageNet and in text-to-image training. The counting runs, described in Appendix G, take 15 minutes for Self-Flow and 25 minutes for SSF per 20k steps on a single NVIDIA L40S GPU.

Table 3: Training runs. The time is the training time of a single run, i.e., of one method, excluding the evaluation of checkpoints.

Experiment	Model	Steps	GPUs	Time
Synthetic shapes	SiT-B/1	20k	1 RTX PRO 6000	2 h
ImageNet	SiT-B/2	800k	2 B200	58 h
ImageNet	SiT-XL/2	1.3M	4 RTX PRO 6000, then 4 B200	171 h
Text-to-image	SiT-XL/2	200k	4 B200	23 h

A.2 RELATION ANNOTATIONS

We annotate the ImageNet images in three stages. First, Qwen3-VL-8B-Instruct (Bai et al., 2025) lists the objects in each image together with their boxes. Second, OWLv2 (Minderer et al., 2023) localizes each listed object, and we keep an object only if its detection score is at least 0.15 and its box overlaps the box proposed by Qwen3-VL with an IoU of at least 0.5. SAM (Kirillov et al., 2023) with a ViT-H backbone then segments the object, and we keep masks with a predicted IoU of at least 0.8 that cover at least 0.5% of the image. Third, Qwen3-VL verifies each remaining object in the image, and only accepted objects form pairs.

For each pair of objects from different categories, we assign the relation from the centers of their visible boxes. We require a separation of at least 10% of the image along the direction of the relation, an offset along the other direction of at most 0.75 of this separation, and a box IoU of at most 0.2, and discard the remaining pairs as ambiguous. This yields 104,956 annotated training images (8.2%). We annotate the 50k validation images with the same pipeline, and the annotation of all 1.28M training images takes about 33 hours on up to four B200 GPUs.

A.3 DECOYS

On ImageNet and in text-to-image training, we build a bank of the token regions of all annotated training subjects, indexed by their relation and by the cell of their reference object, i.e., the center of the reference on an 8×8 grid and its size tercile, with a fallback to the nearest occupied cell. For each alternative relation, we draw 16 regions from the cell of the current reference object and use the first one that does not overlap the reference, or otherwise the one with the least overlap. Each region keeps its original position in its source image, and hence the masked regions follow approximately the same distribution regardless of the true relation.

To verify how closely the implemented masks follow this construction, we replay the training data stream and the masking code on 20,758 annotated training examples of SiT-B/2, and SiT-XL/2 gives the same statistics within one point. The higher noise level covers on average 74% of the subject tokens, or 80% excluding the tokens shared with the reference, which are protected, and all maskable subject tokens in 43% of the examples. The subject and the decoys exceed the budget in 60% of the examples, mostly for large subjects, e.g., subjects of at most 16 tokens are masked completely in 78% of the examples. Of the tokens at the higher noise level, 43% belong to the subject, 45% to the decoys, 11% to the border around them, and 1% are random, and the reference never receives the

higher noise level, except in one of 32,599 examples whose reference covers 190 of the 256 tokens. Since the budget is sampled uniformly from the subject and the decoys, the decoys keep the same fraction of their tokens as the subject even when the subject alone exceeds the budget, i.e., in 23% of the examples. Note that a decoy is on average smaller than the subject, i.e., the three decoys cover 46 tokens in total, compared to 43 tokens for the subject, which a model could use to identify the subject, although the leakage classifier does not benefit from the mask (Table 4). The examples in Figure 1 have small subjects, which are masked completely. In text-to-image training, the subject coverage is 78%.

A.4 IMAGENET

Both model sizes use ImageNet at 256×256 with SD-VAE latents (Rombach et al., 2022), a batch size of 256, a learning rate of 10^{-4}, and $\lambda = 0.8$. In addition to the class, the models are conditioned on the names of the subject and the reference object, from a vocabulary of 3,020 object names, and on the relation, each through a learned embedding. As in Equation (6), the sum of the embeddings modulates every block through adaLN, and the embeddings are also appended to the image tokens with learned position embeddings, where these condition tokens are modulated as clean tokens. Images without a relation annotation, i.e., 92% of the training images, receive null tokens for these three conditions. To enable classifier-free guidance, we drop all conditions together with a probability of 0.1. For SiT-XL/2, we follow the architecture of Self-Flow, where the student features are taken from layer 8 through a two-layer projector and the teacher features from layer 20. From step 1.11M onward, both SiT-XL/2 runs continue to train on all ImageNet images, but keep relation annotations only for pairs in which both objects are detected with a confidence of at least 0.5, i.e., for 1.3% of the images, and apply horizontal flips that swap *left* and *right*.

Note that both SiT-B/2 models sample the noise levels of all images by Equation (10), rather than by Equation (3) for the Self-Flow baseline and for the images of SSF without annotations, and hence the two methods differ only in which tokens are masked. With SiT-XL/2, where the baseline is the original Self-Flow, the results follow a similar trend, e.g., both images of Self-Flow are correct for 2% of the test prompts, compared to 3% with SiT-B/2 (Table 1).

We train SiT-B/2 on two NVIDIA B200 GPUs per method, which takes 58 hours for 800k steps (117 GPU-hours). SiT-XL/2, with 687M parameters including the additional embeddings, trains on four GPUs per method. Since GPUs were not constantly available to us, we had to switch GPU types during training, i.e., the first 1M steps ran on NVIDIA RTX PRO 6000 GPUs at 1.8 steps per second, which took 153 hours, and the remaining 300k steps ran on B200 GPUs at 4.6 steps per second, which took 18 hours. Hence, each SiT-XL/2 run took 171 hours, i.e., about 610 GPU-hours on RTX PRO 6000 and 70 GPU-hours on B200.

A.5 TEXT-TO-IMAGE

We train SiT-XL/2 from scratch at 256×256 with SD-VAE latents, conditioned on captions encoded by a frozen T5-base encoder (Raffel et al., 2020). The training data consists of COCO 2017 training images (Lin et al., 2014), which we resize to fit 256×256 and pad with black instead of cropping them, such that both objects of a pair remain in the image, and hence the generated images often contain black borders. At each step, 75% of the examples are human captions, and 25% are spatial pairs from SCOP (Zhang et al., 2025) with template captions and COCO instance masks. SSF applies the relation mask of Section 3.3 to the SCOP pairs, while both methods use the Self-Flow mask for the human captions. Each SCOP pair specifies the COCO instances of the subject and the reference, from which its template caption is formed, and hence no matching between caption words and objects is needed. We use 236,574 human captions, i.e., two per image, and 28,028 SCOP pairs in both directions, where the 2,826 pairs with a compound relation, e.g., *to the right of and below*, also appear in both word orders, i.e., 61,708 template captions. For a compound relation, SSF builds the mask for one of its component relations, chosen anew at each epoch. The 1,186 test captions are template captions of SCOP pairs in COCO val2017 images, each with a single relation, i.e., 389 *left of*, 360 *right of*, 234 *above*, and 203 *below*.

We use a batch size of 256, a learning rate of 10^{-4}, $\lambda = 0.8$, and a caption dropout of 0.1. Each model, with 829M parameters, trains on four NVIDIA B200 GPUs at 2.4 steps per second and reaches 200k steps after 23 hours (about 91 GPU-hours).

A.6 LEAKAGE CLASSIFIER

Table 4: Relation leakage study on COCO and on synthetic shapes (Appendix F). Numbers are accuracies from spatial cues.

Classifier input	COCO	Synthetic
Majority class	0.317	0.249
Reference box only	0.626	0.388
+ subject mask	0.932	0.847
+ subject-and-decoys mask	0.602	0.376

In Table 4, we report the accuracy of a classifier that predicts the relation from the reference box and the mask. First, the reference box alone predicts the relation with an accuracy of 62.6% on COCO images (Lin et al., 2014), compared to 38.8% on synthetic shapes (Appendix F). Second, masking the subject without decoys largely reveals the relation (93.2% on COCO and 84.7% on synthetic shapes), whereas the SSF mask does not improve over the reference box alone (60.2% and 37.6%). Thus, the decoys prevent the mask from making the relation trivial to infer.

To measure how much a mask reveals the relation, we train a classifier to predict the relation from the reference box and the mask, without the relation condition. Its input is the flattened 16×16 binary mask together with the center and the corners of the reference box in token units, standardized with the training statistics, and the reference-only row uses the box alone. We train a logistic regression and an MLP with two hidden layers of 128 units with Adam for 40 epochs, select the classifier and the epoch by their accuracy on a validation split, and evaluate the selected classifier once on the test set. On COCO, we use 8,000 subject and reference pairs from 3,687 training images, of which 1,600 pairs from 20% of the images are held out for validation, and test on 1,830 pairs from 672 held-out images. On synthetic shapes (Appendix F), we use 8,000 training images, i.e., 4,000 counterfactual pairs, of which 20% of the pairs are held out for validation, and test on the 2,000 test images, i.e., 1,000 counterfactual pairs, whose combinations of shapes, colors, and relation axis do not appear in training. Compared to the reference box alone, the subject mask increases the accuracy by 30.6 points on COCO and by 45.9 points on synthetic shapes (95% confidence intervals [28.3, 33.1] and [43.6, 48.2], paired bootstrap over test images), whereas the SSF mask changes it by -2.4 [-4.1, -0.8] and -1.2 [-3.3, 0.9] points. Averaged over five random validation splits, the accuracy of the reference box alone, the subject mask, and the SSF mask is 62.2%, 93.1%, and 60.7% on COCO, and 38.1%, 85.0%, and 37.2% on synthetic shapes.

A.7 EVALUATION

On ImageNet, we evaluate sensitivity and recognition only on validation images in which both objects are detected with a confidence of at least 0.4. Recognition averages the denoising error over the noise levels 0.2, 0.4, 0.6, and 0.8, with two noise samples per level that are shared by all candidate relations, and sensitivity uses the noise level 0.8. Text-to-image uses the same protocol, where sensitivity is the relative change of the velocity over the whole image. For ImageNet generation with SiT-XL/2, we sample the 100 test prompts with classifier-free guidance over the full condition c, the class alone c_{cls}, i.e., with null subject, reference, and relation, and no condition,

$$v = v(c) + (w_1 - 1)\big(v(c) - v(\varnothing)\big) + (w_2 - 1)\big(v(c) - v(c_{\text{cls}})\big),$$

with $w_1 = w_2 = 4$. The *joint* setting in Table 6 instead uses standard guidance, $v = v(\varnothing) + w\big(v(c) - v(\varnothing)\big)$ with $w = 4$. We then detect the objects with OWLv2 at a threshold of 0.1. For each image, the detector keeps the highest-scoring box of each of the two objects, which does not depend on the relation, and we decide the relation by the offset between the centers of the two boxes along the axis of the relation. Hence, an unchanged image passes at most one of two opposite relations, and both images are correct only if the layout changes with the relation condition. For text-to-image, we sample with a guidance scale of 4 on the caption, and Qwen3-VL judges the relations while OWLv2 detects the objects. The judge, Qwen3-VL-8B with greedy decoding, receives one image at a time with the following prompt, where {S}, {R}, and {O} are the subject, the relation, and the reference, and only the answer *yes* counts as correct.

```
Inspect this image independently. Question: does the
image show {S} {R} {O}? First decide whether each named
object is clearly visible, then whether the stated
spatial relation between them is visible. Return JSON
only: {"subject_present":true,"object_present":true,
"relation":"yes"} with relation one of yes, no, unclear.
```

Since the prompt does not specify which instance is meant, one image can pass both captions, e.g., when a scene contains several people, which our criterion excludes, since it also requires the judge to reject the image for the opposite caption. We compute 95% confidence intervals with a paired bootstrap over the test prompts.

FID and sFID use 50k samples from the stochastic sampler of Self-Flow with 250 steps and no guidance, and the evaluation code of ADM (Dhariwal & Nichol, 2021). For SiT-B/2, the sampler uses the diffusion coefficient of SiT, i.e., t instead of $2t$, and the MSE-finetuned VAE decoder, and hence its FID is comparable between the two methods but not with that of SiT-XL/2. For ImageNet, the 50k class labels are drawn uniformly at random and shared by both methods, with null subject, reference, and relation, i.e., FID measures class-conditional generation. Sampling 50k images takes about 8 GPU-hours on NVIDIA A100 GPUs. For text-to-image, the FID-5k uses 5,000 unguided samples against 5,000 COCO val2017 images, padded in the same way.

B Does the model use the relation condition?

The generation accuracy in the main paper evaluates whether the generated images follow the relation, but it depends on the sampler and on the detector. In this section, we complement it with two tests that examine the model directly. First, sensitivity measures whether the model uses the relation condition at all (Appendix B.1). Second, recognition measures whether the model uses it in the right direction, i.e., whether the true relation explains a real image better than the flipped one (Appendix B.2).

B.1 Sensitivity to the relation condition

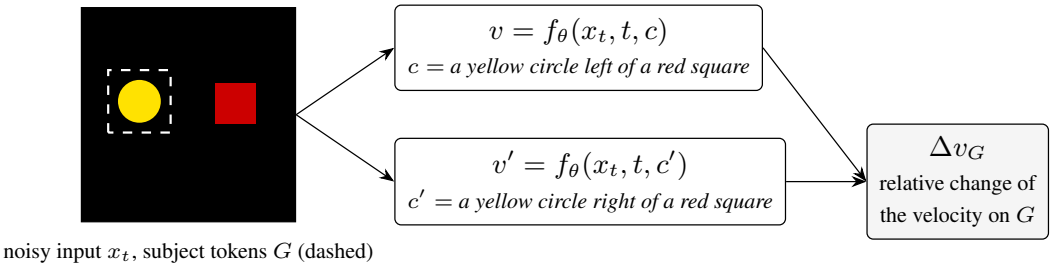


Figure 6: Measuring the sensitivity to the relation condition. We keep the noisy input x_t, the noise level t, and the checkpoint fixed, flip only the relation in the condition, and measure the relative change of the predicted velocity on the subject's tokens G. A model that ignores the relation obtains $\Delta v_G = 0$.

Intuitively, if a model uses the relation condition, changing only the relation, e.g., from *left of* to *right of*, should change its prediction for the subject, whereas a model that ignores the relation condition predicts the same velocity for both prompts. Unlike generation accuracy, this test requires neither sampling nor a detector. We keep the latent, the noise, the noise level, and the checkpoint fixed, flip only the relation condition from c to c' (e.g., from *left* to *right*, or from *above* to *below*), and measure the relative change of the predicted velocity on a token group G (Figure 6),

$$\Delta v_G = \frac{1}{|G|} \sum_{i \in G} \frac{\|f_\theta(x_t, t, c)_i - f_\theta(x_t, t, c')_i\|_2}{\frac{1}{2}\big(\|f_\theta(x_t, t, c)_i\|_2 + \|f_\theta(x_t, t, c')_i\|_2\big)}. \tag{11}$$

To verify that the response is specific to the relation, rather than to any change of the prompt, we compare it to a control that changes the prompt without changing the relation, i.e., the category of the subject.

On ImageNet, we measure the relative change of the velocity over the whole image, i.e., $\|f_\theta(x_t, t, c) - f_\theta(x_t, t, c')\|_2 / \|f_\theta(x_t, t, c)\|_2$ over all tokens, at the noise level 0.8 and on the 307 validation images with seen pairs. Flipping the relation changes the prediction of SSF by 7.7%, compared to 0.5% for Self-Flow, with SiT-B/2 at 800k steps, and by 10.0%, compared to 1.1%, with SiT-XL/2 at 1.3M steps. On the 241 images with unseen pairs, the change is 8.9% vs. 0.5% and 11.9% vs. 1.3%, respectively. In contrast, swapping the category of the subject changes the prediction of both models considerably, i.e., by 9.6% for SSF and 9.1% for Self-Flow with SiT-B/2, and by 14.9% and 11.2% with SiT-XL/2. Hence, Self-Flow uses its condition but ignores the relation condition in particular, whereas SSF uses both.

B.2 Recognition of the relation in real images

Sensitivity measures how much the prediction changes, but not whether it changes in the right direction. Recognition tests the direction on real images, using the model itself as a zero-shot classifier (Li et al., 2023a). Intuitively, a model that understands the relation should find a real image easier to denoise under the true relation than under the flipped one. Specifically, we add noise to the image and denoise it twice, once conditioned on the true relation and once on the flipped one, and a prediction is correct if the true relation yields the lower denoising error, averaged over noise levels (chance

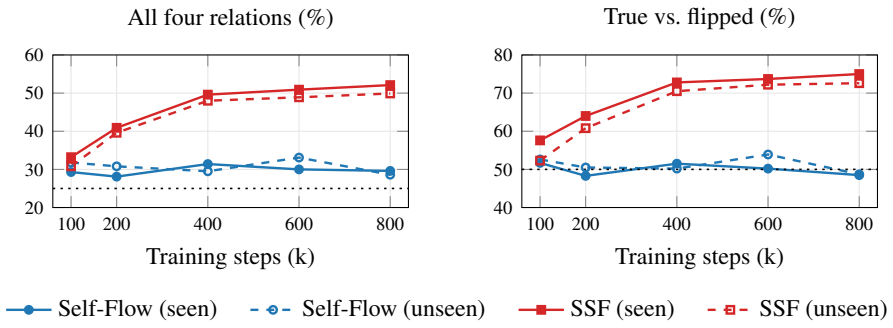


Figure 7: Recognition during training with SiT-B/2 on ImageNet, one seed. The left panel shows the accuracy of predicting the true relation among all four relations (chance 25%), and the right panel shows the accuracy of distinguishing the true relation from the flipped one (chance 50%). Solid and dashed lines denote seen and unseen category pairs, respectively.

Table 5: Recognition on ImageNet at the end of training, i.e., SiT-B/2 at 800k steps and SiT-XL/2 at 1.3M steps. True vs. flipped has a chance level of 50%, and all four relations a chance level of 25%.

		True vs. flipped (%)		All four relations (%)	
Model	Method	Seen ↑	Unseen ↑	Seen ↑	Unseen ↑
SiT-B/2	Self-Flow	48.5	48.7	29.6	28.6
	SSF (ours)	**75.0**	**72.6**	**52.1**	**49.9**
SiT-XL/2	Self-Flow	53.4	47.7	32.6	31.5
	SSF (ours)	**76.6**	**66.0**	**53.1**	**43.6**

50%). Unlike generation accuracy, this test uses real images and requires neither sampling nor a detector, and it is cheap enough to track during training.

We also report the accuracy of predicting the true relation among all four relations, as in Jeong et al. (2025) (chance 25%), which additionally requires distinguishing horizontal from vertical relations. For SiT-XL/2, both tests use 548 annotated ImageNet validation images, i.e., 307 with seen and 241 with unseen category pairs, in which both objects are detected with a confidence of at least 0.4. For SiT-B/2, we use up to 1,024 validation images per split. The unseen pairs test whether the model generalizes the relation to category pairs that never appear together in the training annotations.

In Figure 7, we show that with SiT-B/2, the accuracy of Self-Flow in distinguishing the true from the flipped relation remains near the chance level for both seen and unseen pairs, and its four-way accuracy shows little improvement during training. In contrast, SSF pulls ahead by 200k steps and reaches 75% on seen pairs and 73% on unseen pairs at 800k steps. In Table 5, we summarize both models at the end of training, and show that SiT-XL/2 at 1.3M steps follows the same pattern in both tests.

C ADDITIONAL RESULTS

In this section, we report results that complement the main paper.

Table 6: Generation accuracy (%) of SiT-XL/2 at 1.3M steps for different guidance scales.

Guidance scale		Self-Flow			SSF (ours)		
w_1	w_2	Original	Flipped	Both	Original	Flipped	Both
2.5	3	43	20	1	55	41	**26**
4	3	40	22	1	57	42	**30**
4	4	45	19	2	63	46	**32**
6	3	48	20	6	63	44	**34**
4 (joint)		36	17	0	53	40	**28**

Guidance scale In Table 6, we report the generation accuracy of SiT-XL/2 at 1.3M steps for different guidance scales w_1 and w_2 (Appendix A.7), where *joint* uses standard guidance with a single scale. We observe that SSF generates both the original and the flipped layout far more often than Self-Flow for every guidance scale (26% to 34% vs. 0% to 6%). Hence, the gain of SSF does not depend on a specific choice of guidance. Moreover, even without guidance, SSF at 1M steps generates both layouts for 10.3% of the samples, compared to 0.3% for Self-Flow, on a larger set of 300 test prompts with two noise samples each.

Relation accuracy when both objects appear Following the relation requires generating both objects, and SSF generates both objects less often under the flipped prompt (59% vs. 68%). We therefore also compute the accuracy only over the images in which both objects are detected. With the guidance of Table 1, SSF follows the relation in 91% of these images with the original prompt, compared to 66% for Self-Flow, and in 78% vs. 28% with the flipped prompt. This indicates that the gain of SSF is not due to differences in object presence.

Generation with SiT-B/2 In Table 7, we report the generation accuracy of SiT-B/2 at 800k steps with the protocol of Table 1, where the two methods differ only in the placement of the mask (Appendix A.4). SSF generates both layouts far more often than Self-Flow, for seen and unseen category pairs alike (40% vs. 4% and 34% vs. 2%). The gain in both holds at 900k steps, the last checkpoint of Self-Flow (37% vs. 0%), with standard guidance of scale 4 (23% vs. 2%), and with two other noise seeds for sampling (gains of 36 and 40 points). Moreover, both objects appear more often in the images of SSF (74% vs. 59% for the original prompt, and 60% vs. 55% for the flipped one), and among the images with both objects, SSF shows the requested relation for 92% vs. 73% of the original and 80% vs. 31% of the flipped prompts. Self-Flow barely changes its image when the relation is flipped, i.e., the mean absolute pixel change is 9.2 of 255, compared to 57.0 for SSF.

Table 7: Generation accuracy (%) of SiT-B/2 at 800k steps, where the two methods differ only in the placement of the mask, with the paired bootstrap 95% confidence interval of the difference over prompts.

Pairs	Method	Original	Flipped	Both
All	Self-Flow	43	17	3
	SSF (ours)	**68**	**48**	**37**
	Difference	+25 [14, 37]	+31 [20, 42]	+34 [24, 44]
Seen	Self-Flow	52	22	4
	SSF (ours)	**70**	**52**	**40**
	Difference	+18 [4, 32]	+30 [14, 44]	+36 [22, 50]
Unseen	Self-Flow	34	12	2
	SSF (ours)	**66**	**44**	**34**
	Difference	+32 [14, 50]	+32 [16, 48]	+32 [18, 46]

Seen and unseen category pairs In Table 8, we split the generation accuracy of SiT-XL/2 in Table 1 into the 50 test prompts with seen and the 50 with unseen category pairs. SSF improves both, e.g., both layouts are correct for 30% vs. 4% of the prompts with unseen pairs, which indicates that the relation generalizes to category pairs that never appear together in the training annotations. Note that the intervals reflect the sampling of test prompts, not of training runs.

Table 8: Generation accuracy (%) of SiT-XL/2 at 1.3M steps for seen and unseen category pairs (50 prompts each), with the paired bootstrap 95% confidence interval of the difference over prompts.

Pairs	Method	Original	Flipped	Both
Seen	Self-Flow	50	20	0
	SSF (ours)	**70**	**50**	**34**
	Difference	+20 [4, 36]	+30 [14, 46]	+34 [22, 48]
Unseen	Self-Flow	40	18	4
	SSF (ours)	**56**	**42**	**30**
	Difference	+16 [0, 32]	+24 [12, 38]	+26 [14, 38]

Relation axis In Table 9, we split the generation accuracy by the axis of the relation. For *left of* and *right of*, the flipped layout is a mirror image of the original one, and SSF follows the flipped prompt as often as the original one, e.g., for 65% vs. 61% of the prompts with SiT-B/2. For *above* and *below*, the flipped layout is often unnatural, e.g., *stool above piano* instead of *stool below piano*, and the accuracy of SSF drops from 74% to 33%, while both objects appear less often (48% vs. 76%). In contrast, Self-Flow reaches 52% for the original vertical prompts, which usually describe the typical layout, but only 9% for the flipped ones, since it generates nearly the same image for both. Hence, the lower accuracy under the flipped prompt mainly reflects unnatural vertical layouts, and the gain of SSF in both holds for both axes.

Table 9: Generation accuracy (%) on ImageNet by the axis of the relation, with the number of test prompts in parentheses, and the share of images in which both objects are detected for the original and the flipped prompt.

Model	Axis	Method	Original	Flipped	Both	Both objects
SiT-B/2	Left, right (46)	Self-Flow	32.6	26.1	6.5	57 / 48
		SSF (ours)	**60.9**	**65.2**	**45.7**	72 / 74
	Above, below (54)	Self-Flow	51.9	9.3	0.0	61 / 61
		SSF (ours)	**74.1**	**33.3**	**29.6**	76 / 48
SiT-XL/2	Left, right (46)	Self-Flow	26.1	28.3	4.3	63 / 63
		SSF (ours)	**63.0**	**67.4**	**50.0**	72 / 74
	Above, below (54)	Self-Flow	61.1	11.1	0.0	72 / 72
		SSF (ours)	**63.0**	**27.8**	**16.7**	67 / 46

Text-to-image during training In Table 10, we report the sensitivity and the FID-5k of the text-to-image models of Section 4.2 at 50k, 100k, and 200k steps. SSF responds more strongly to the relation than Self-Flow at every checkpoint, and its FID-5k is lower at 100k and 200k steps.

Table 10: Text-to-image sensitivity and image quality during training. Dashes indicate no evaluation.

		Sensitivity (%)	Image quality
Method	Steps	Relation ↑	FID-5k ↓
Self-Flow	50k	1.41	–
SSF (ours)	50k	**2.66**	–
Self-Flow	100k	3.64	27.06
SSF (ours)	100k	**4.96**	**24.76**
Self-Flow	200k	7.16	26.50
SSF (ours)	200k	**9.20**	**24.93**

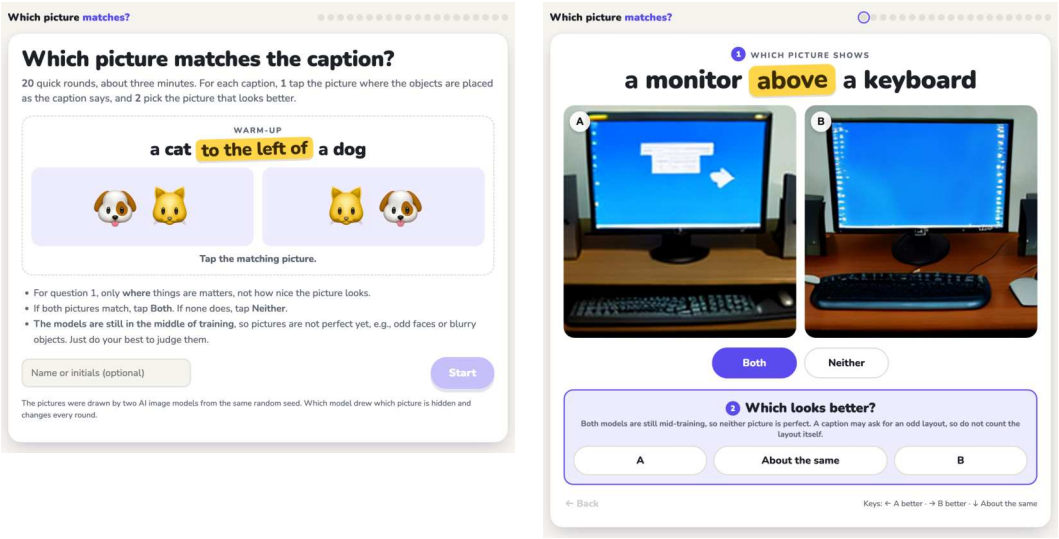
Unchanged-image control Since the two images are judged separately, both could count as correct even if the model ignores the relation, e.g., when an unchanged image satisfies both captions. We therefore also give the judge each image with the opposite caption. At 200k steps, the judge accepts the original image for both captions for 31.6% of the captions with Self-Flow and 32.6% with SSF. If we only checked each image with its own caption, both images would count as correct for 51.7% vs. 40.9% of the captions, but for 70% and 77% of these, the judge also accepts at least one of the two images for the opposite caption. Hence, we count an image as correct only if the judge also rejects it for the opposite caption, which gives 15.4% vs. 9.3% for both (a gain of 6.1 points). This gain comes almost entirely from *left* and *right* (22.6% vs. 14.2% for 749 captions), since for *above* and *below*, the judge rarely distinguishes the two relations (3.0% vs. 0.9% for 437 captions). Without the judge, the flipped image is nearly identical to the original one for 30.9% of the captions with Self-Flow and 20.2% with SSF (a difference of 10.7 points, [7.5, 13.9]).

D HUMAN EVALUATION

We complement the automatic evaluation with a blind human study, whose interface is shown in Figure 8. In each round, a rater sees a caption and the two images generated by Self-Flow and by SSF from the same noise, and answers two questions, i.e., in which image the relation is correct (A, B, both, or neither), and which image looks better (A, B, or about the same). The rater is not told which model generated which image, and each model appears on the left in half of the rounds, in a random order. A warm-up round explains the task, and raters are told that the models are still in training. University students who are not authors gave 411 ratings in total, which Figure 3 pools. The study ran in two rounds.

In the first round, ten raters, nine of whom completed all 20 rounds, rated 20 of the 100 ImageNet test prompts of Table 1, with the same images and guidance, i.e., 182 ratings, where each rater sees every prompt once, half with the original and half with the flipped relation, and the first question asked which picture matches the caption. To keep the task easy to judge, we kept the prompts where OWLv2 detects both objects in all four images, i.e., for both models and both relations, and then checked the four images of each prompt blindly for recognizable and non-overlapping objects. Raters judge the image of SSF to show the caption for 76% vs. 48% of the images with the original relation and 68% vs. 32% with the flipped one, and a majority vote per image agrees with the automatic check for 84% of the images. Raters prefer the image of Self-Flow in 40% vs. 21% of the rounds, with the remaining 40% about the same.

In the second round, 12 raters rated 6 to 30 image pairs each, i.e., 229 ratings, of SiT-XL/2 at 1.5M steps for 13 hand-picked test prompts with up to five noise seeds, where we kept the prompt and seed pairs in which OWLv2 detects both objects in all four images, and the first question asked in which image the relation holds. The relation holds in 67% of the images of SSF, compared to 34% for Self-Flow, and the gap is positive for every rater. Raters prefer the look of the image of Self-Flow in 51% of the ratings and that of SSF in 30%. The agreement between pairs of raters on the relation is 70%.



(a) Instructions and warm-up

(b) A round with both questions

Figure 8: Interface of the human study. (a) The instructions and a warm-up round. (b) A round, where question 2 becomes active after question 1 is answered.

E ADDITIONAL QUALITATIVE RESULTS

To complement the selected samples of Figure 4, we show uncurated samples, i.e., randomly drawn prompts, each generated with the original and the flipped relation from the same noise. For ImageNet, Figures 9 to 12 show 15 prompts for each of the four original relations, and Figure 13 shows six text-to-image captions at 200k steps. In addition to Figure 5, Figure 14 shows more selected text-to-image samples, among the captions for which both images of SSF are correct (Section 4.2).

On ImageNet, Self-Flow generates nearly the same image for the original and the flipped relation for almost all prompts, whereas SSF changes the layout for many of them, although it sometimes distorts or omits an object. In text-to-image training, both models keep nearly the same image for most of the six captions at 200k steps, and the effect of SSF is less visible than on ImageNet. For these six captions, both images are correct only for the fork, for both methods, and the judge accepts several unchanged images for both relations.

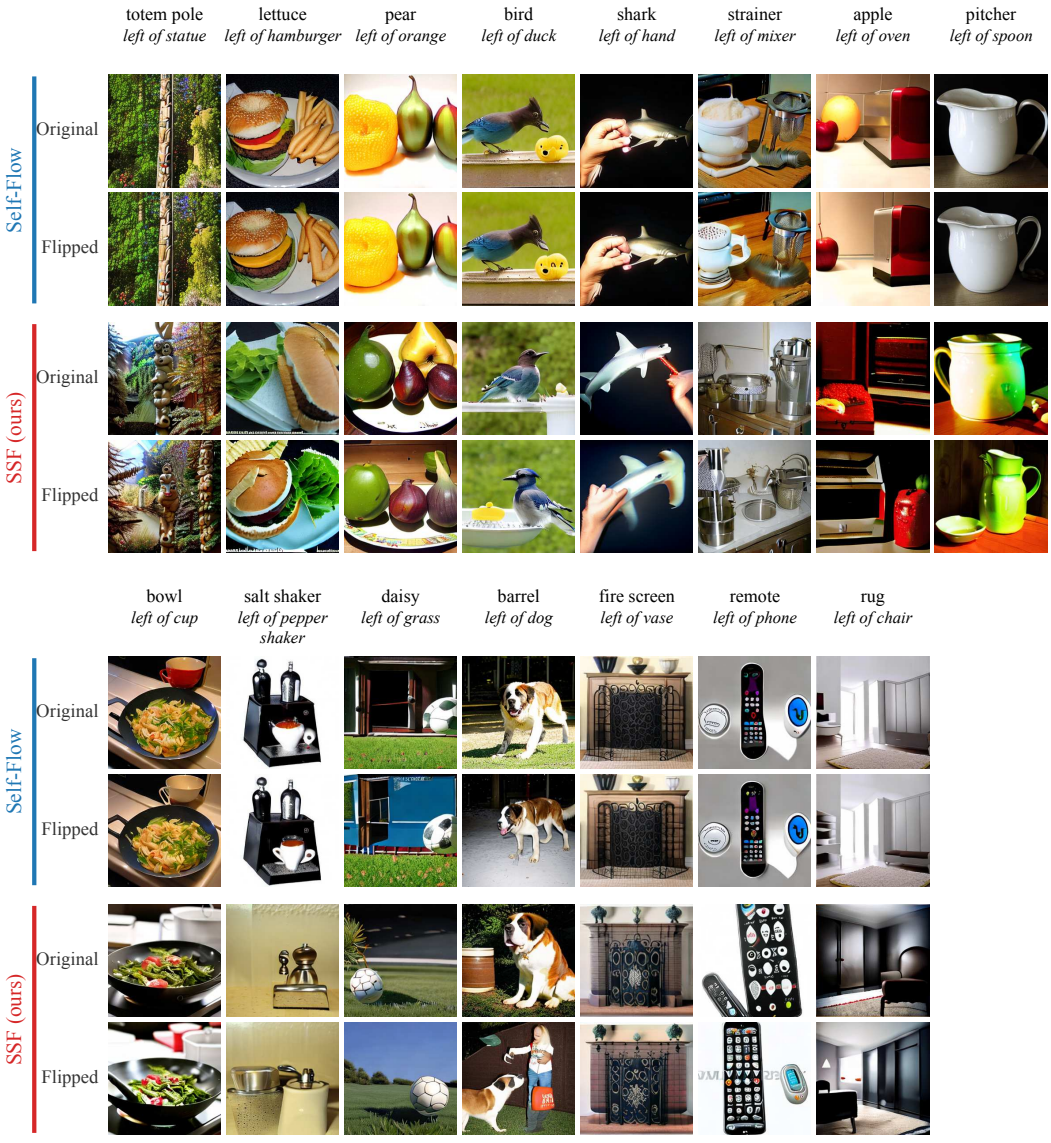


Figure 9: Uncurated same-noise samples on ImageNet for 15 randomly drawn test prompts whose original relation is *left of*, generated with the guidance of Table 1. Each column shows one prompt with the original and the flipped relation.

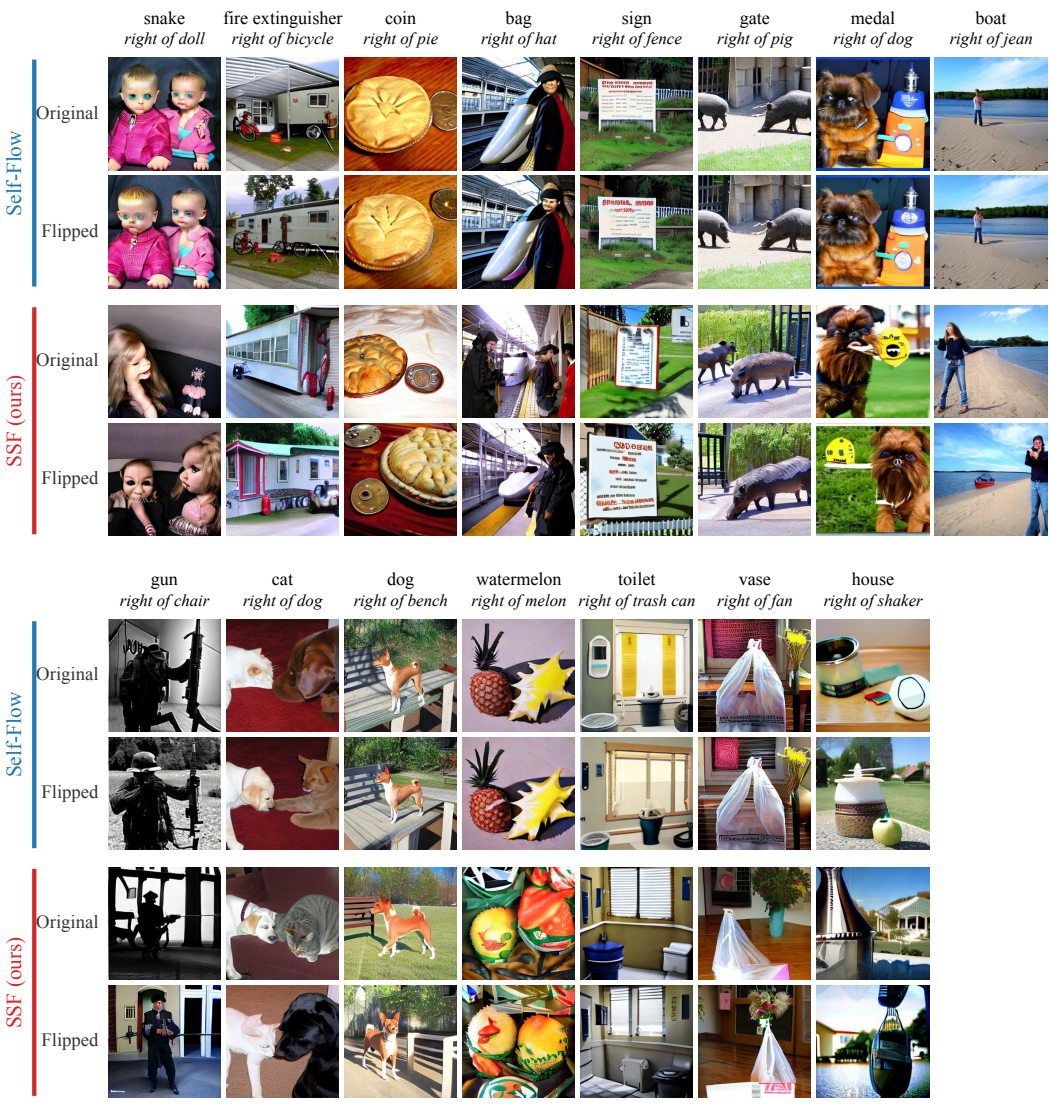


Figure 10: Uncurated same-noise samples on ImageNet for 15 randomly drawn test prompts whose original relation is *right of*, generated with the guidance of Table 1. Each column shows one prompt with the original and the flipped relation.

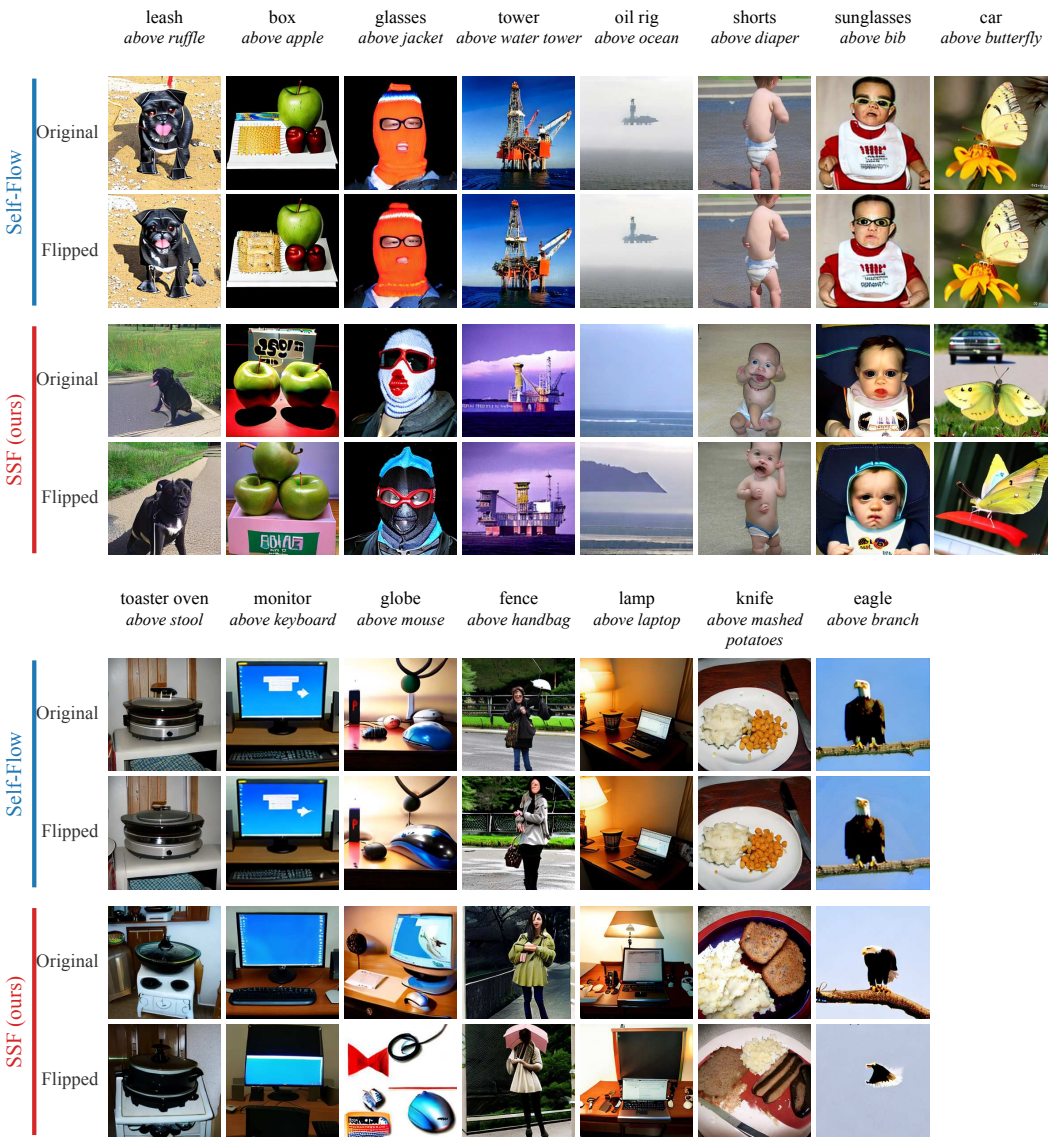


Figure 11: Uncurated same-noise samples on ImageNet for 15 randomly drawn test prompts whose original relation is *above*, generated with the guidance of Table 1. Each column shows one prompt with the original and the flipped relation.

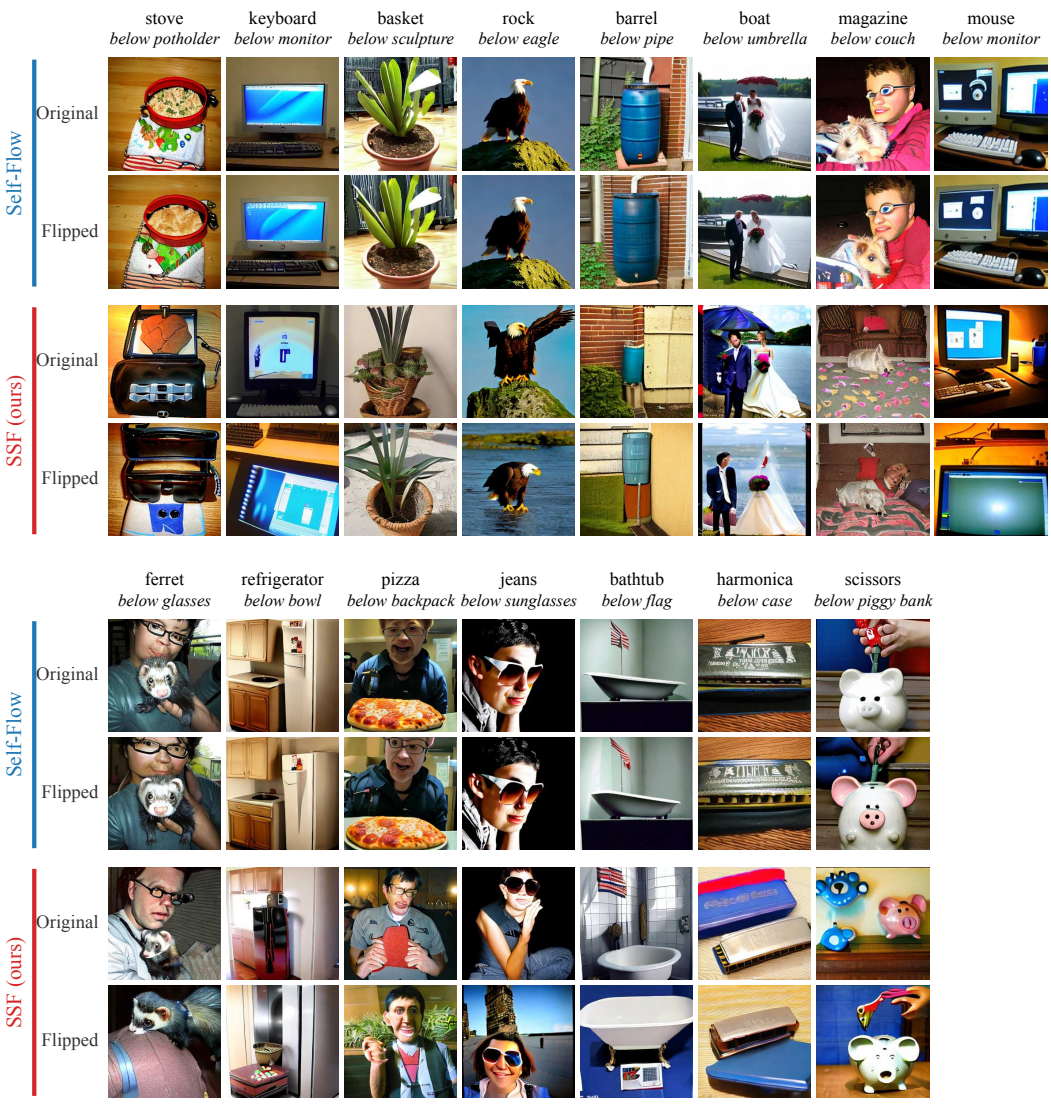


Figure 12: Uncurated same-noise samples on ImageNet for 15 randomly drawn test prompts whose original relation is *below*, generated with the guidance of Table 1. Each column shows one prompt with the original and the flipped relation.



Figure 13: Uncurated same-noise text-to-image samples at 200k steps for 6 of the 1,186 test captions, drawn at random and generated with a guidance scale of 4. Each column shows one caption with the original and the flipped relation.

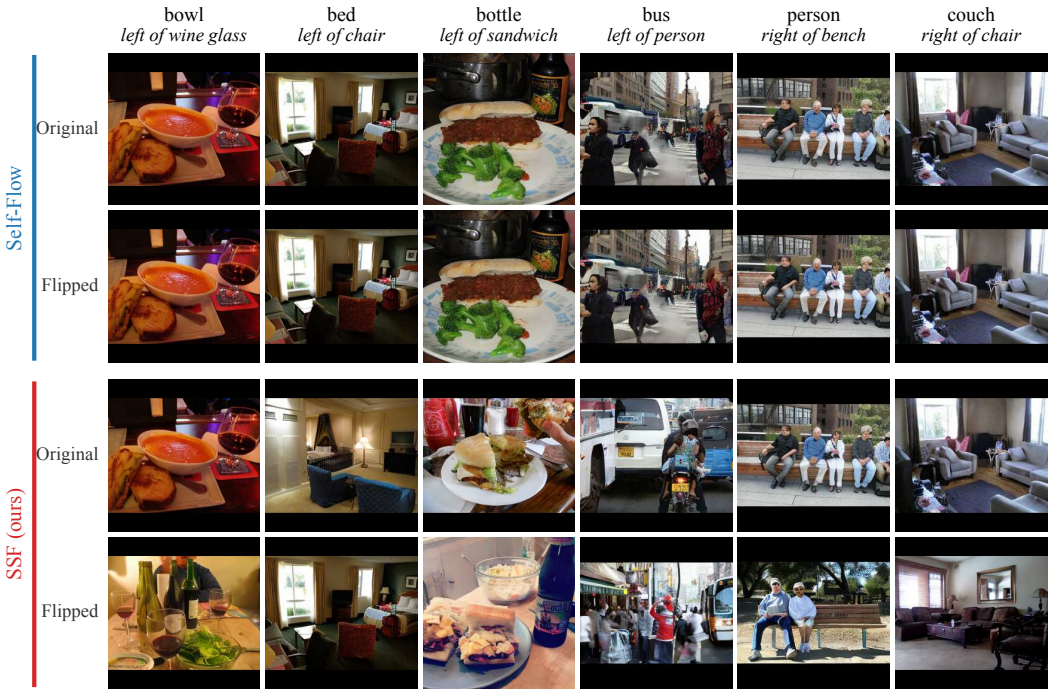


Figure 14: Additional text-to-image samples for the original and the flipped caption, generated from the same noise.

F SYNTHETIC SHAPES

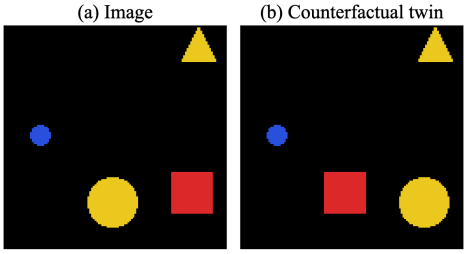


Figure 15: Synthetic shapes. (a) *A yellow circle left of a red square*, with two distractors. (b) Its counterfactual twin.

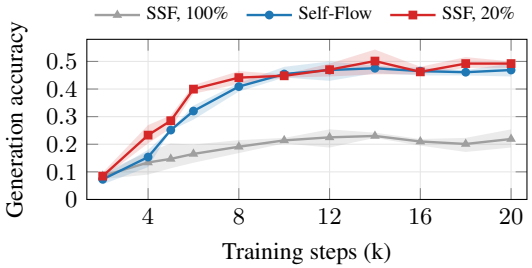


Figure 16: Generation accuracy on synthetic shapes during training with the relation-aware mask on 20% or on all training images. Lines show the mean over three seeds, and the shading shows their range.

Data and training Each 128×128 image contains a subject, a reference object, and up to two distractors, drawn from colored circles, squares, and triangles on a black background, such that only the positions of the two objects predict the relation, and each image is paired with a counterfactual twin that changes only the relation (Figure 15). We train on 10,000 counterfactual pairs a SiT-B/1 with 12 layers, a width of 768, and 12 heads, i.e., 132.7M parameters, on SD3.5-VAE latents with 256 tokens, with QK-normalization, where the student features are taken from layer 4 and the teacher features from layer 8. The model is conditioned on learned embeddings of the shape and the color of each object and of the relation, which are added to the timestep embedding and appended to the image tokens, and all conditions are dropped together with a probability of 0.1. We train for 20k steps with a batch size of 128, a learning rate of $3 \cdot 10^{-4}$ after 500 warmup steps, $\lambda = 0.5$, and an EMA decay of 0.999, with three seeds. All runs mask 64 of the 256 tokens, assign the masked tokens the higher noise level of Equation (10), and show the teacher the image at the lower level, and hence differ only in which tokens are masked. Every 2k steps, we measure the generation accuracy on 256 held-out prompts with a guidance scale of 3, i.e., the fraction of generated images in which a shape detector finds both objects in the requested relation (chance 0.25), and we report its average over training, i.e., the area under the accuracy curve from 2k to 20k steps divided by 18k.

Masks The relation-aware mask covers the subject and, as decoys, its seven rotations and reflections around the center of the reference object, with a one-token border, never covers the reference, and fills the rest of the budget with random tokens. It covers 97% of the subject tokens and hides the subject completely in 84% of the images, where half of its budget goes to the decoys. For mixed sampling, each training image receives this mask with probability p and the random mask of Self-Flow otherwise.

Results In Figure 16 and Table 11, we show that applying the relation-aware mask to all images lowers the accuracy far below that of Self-Flow throughout training, possibly since the seven decoys leave two candidate locations for each relation, and since the model then never sees images with the random mask, from which it learns the appearance and the natural layout of the objects. In contrast, applying it to 20% of the images, which works best among the shares we tried, learns the relation faster than Self-Flow in all three seeds, whereas 10% is not enough, and 30% and 50% give smaller gains. Hence, mixed sampling is necessary. Note that the mixed runs were trained on four NVIDIA B200 GPUs, the runs on all images on one NVIDIA L40S GPU, where Self-Flow reaches 0.373 with one seed, and Self-Flow on one NVIDIA RTX PRO 6000 GPU.

Table 11: Generation accuracy on synthetic shapes, averaged over training (mean and standard deviation over three seeds) and at 20k steps.

Mask	Images with the mask	Average	At 20k steps
Random (Self-Flow)	–	0.387 ± 0.004	0.469
Relation-aware	10%	0.378 ± 0.014	0.471
Relation-aware	20%	0.413 ± 0.008	0.492
Relation-aware	30%	0.394 ± 0.002	0.497
Relation-aware	50%	0.397 ± 0.014	0.454
Relation-aware	100%	0.191 ± 0.010	0.219

G COUNTING ON SYNTHETIC SHAPES

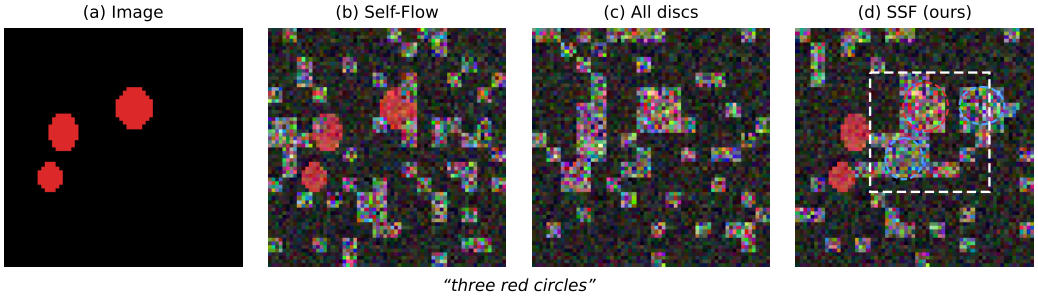


Figure 17: Masks for counting on the same training image, conditioned on *three red circles*. (b) Self-Flow noises random tokens, so the visible discs reveal the count. (c) Noising all discs reveals the count through the number of noised shapes. (d) SSF, here with $K = 3$, noises the disc inside a random window (dashed square, red) together with decoy discs (blue) until three shapes are noised.

Table 12: Counting accuracy (%) for both targets K, over three seeds. We use $K = 5$ by default.

	Training steps	
Method	20k	40k
Self-Flow	64.5	91.8
SSF ($K = 3$)	66.8	85.1
SSF ($K = 5$)	**74.5**	**92.3**

Table 13: Earlier runs with $K = 3$ (%), mean and standard error over two seeds (one for all discs).

	Training steps		
Method	10k	15k	20k
Self-Flow	0±0	17±1	62±1
All discs	2	14	18
SSF ($K = 3$)	**12**±2	**55**±13	**74**±5

Spatial relations are only one of the semantic properties that generative models often fail on (Schwartz et al., 2023; Zafar et al., 2026), and in principle, SSF applies to any property that recovery can be made to depend on. We therefore explore counting on synthetic shapes. Similar to spatial relations, the count can often be read from the image, e.g., a random mask leaves most discs visible, and thus the model can copy the count from the image rather than from the prompt. Hence, SSF noises the discs together with decoy discs, such that the noised shapes reveal little about the count (Figure 17). SSF reaches a higher counting accuracy than Self-Flow at the same number of training steps, i.e., 74.5% at 20k steps compared to 64.5%, and it matches Self-Flow at 40k steps (Table 12). This indicates that SSF could also help with other semantic properties, and we leave a complete study to future work.

Data and model We generate 64×64 images that contain one to five red discs of 7 to 11 pixels at random, non-overlapping positions, conditioned on the count, e.g., *three red circles*. Each scene has counterfactual twins that add or remove a disc at a fixed latent position. We train a small pixel-space SiT with four layers, a width of 192, and a patch size of 4, using the Self-Flow objective for 20k steps (40k steps for the three-seed runs in Table 12) with a batch size of 128 and a learning rate of $3 \cdot 10^{-4}$. All methods noise exactly 64 of the 256 tokens at the higher noise level.

Masks As shown in Figure 17, SSF noises the discs whose center lies inside a random 8×8-token window and adds decoy discs until K shapes are noised, while the discs outside the window remain at the lower noise level and provide the context, similar to the reference object in spatial relations. We use $K = 5$ by default and also report $K = 3$. If the window contains more than K discs, all of them are noised without decoys, and a decoy that would exceed the masking budget of 64 tokens is dropped, while the real discs are never truncated, i.e., K is the target number of noised shapes. Thus, the noised shapes look similar regardless of the count, which encourages the model to use the count condition to tell how many of them are real.

Evaluation and results We generate 40 images per count prompt with 50 Euler steps and a guidance scale of 3, and count the discs in each image. Table 12 reports both targets we tried, $K = 3$ and $K = 5$, with three seeds, where we evaluate the 40k checkpoints on 1,000 images. With $K = 5$, SSF learns faster than Self-Flow and matches it at 40k steps (92.3% vs. 91.8%), whereas with $K = 3$, SSF is only slightly ahead at 20k steps and falls behind at 40k steps (85.1%). In earlier runs with $K = 3$ (Table 13), SSF also learns faster than Self-Flow during the first 20k steps, whereas noising all discs provides no gain.