SEMANTICS AS SOURCE: GENERATIVE FLOWS FROM SEMANTIC EMBEDDINGS INSTEAD OF NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained semantic representations have become increasingly important in modern image generators, serving as conditioning signals, as feature-alignment targets, and even as the latent space itself. However, in every case the generator still begins from a content-free input, typically Gaussian noise. We propose semantics-as-source (SAS), which directly uses semantic embeddings as the source of a generative flow. The flow starts from the image's pooled semantic embedding, perturbed with noise, and the model learns the spatial residual that recovers the image's full latent token grid, which a frozen decoder maps to pixels. At inference, a small prior over embeddings supplies the starting point. Compared with conditioning and alignment on the same backbone and training budget, SAS gives better images and faster training across model scales and conditional and unconditional settings. On ImageNet-1k with SiT-XL/1, our source formulation achieves an unguided FID of 2.71 after only 40 epochs, compared with 2.96 for conditioning and 4.12 for alignment. SAS also roughly halves the expected squared distance the flow must cover, and the source route passes the Gaussian source's final 40-epoch FID within 10 epochs. Swapping only the source into RAE's tuned DiT-DH-XL recipe passes their released 80-epoch model by epoch 50 and beats it at an identical budget; with only 8 sampling steps our model matches the FID RAE reaches with 50 steps. With a decoupled head and autoguidance, SAS reaches FID 1.86 on ImageNet-1k 256^2 after 160 epochs. These results suggest that semantic information is most effective not merely as guidance for a generative model, but as a structured starting point from which generation proceeds.

1 INTRODUCTION

Pretrained visual representations have increasingly become a central ingredient of modern image generators. REPA (Yu et al., 2024) aligns the generator's features with them, RCG (Li et al., 2024) conditions the generator on them through a learned prior over representations, and RAE (Zheng et al., 2026) makes the encoder's token grid the latent space itself. Yet all of these systems keep the same kind of starting point, where generation begins from an input that carries no content, typically Gaussian noise. As a result, the models need to spend much of their early training discovering this semantic structure (Yu et al., 2024).

We propose *semantics-as-source* (SAS), which uses a frozen semantic embedding as the source distribution of a generative flow. The key intuition behind our approach is that since the pretrained encoder already provides the semantic structure needed for image generation, we can leverage this structure as a starting point. In our construction, each training trajectory starts from the image's own pooled semantic embedding, perturbed with noise, and the flow learns the spatial residual that recovers the image's full RAE latent token grid, which a frozen decoder maps to pixels. At inference, a trained lightweight prior over embeddings supplies the starting point (Fig. 1). Generation is thereby split in two stages, where the prior samples the global semantic structure in the pooled embedding space, and the flow generates the spatial residual that adds the finer details.

SAS outperforms the two established ways of using the same embedding, conditioning as in RCG and feature alignment as in REPA, as well as RAE's Gaussian source, in three ways: better samples, faster training, and fewer sampling steps. All variants share one frozen encoder (DINOv2-B), the

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.

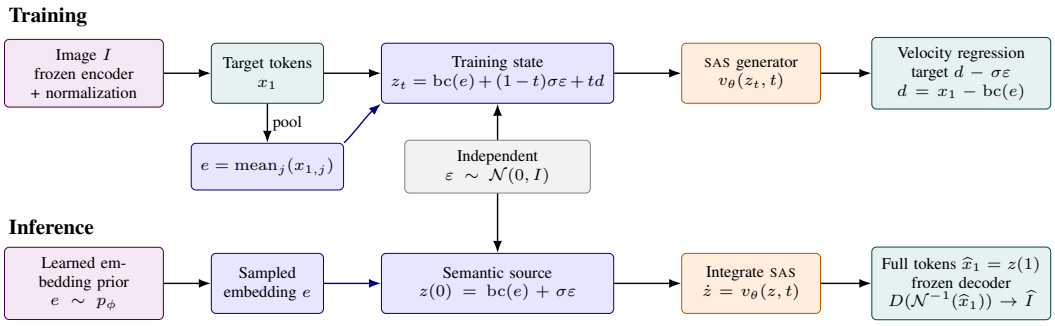


Figure 1: **Semantics as the source of the transport.** *Training:* the frozen encoder maps image I to normalized tokens x_1; their spatial mean e is broadcast to every position and noised, so each trajectory starts at $z_0 = \text{bc}(e) + \sigma\varepsilon$ built from its own image, and the generator regresses the velocity $d - \sigma\varepsilon$. *Inference:* a learned prior p_ϕ samples e, the flow is integrated from $\text{bc}(e) + \sigma\varepsilon$, and the frozen decoder maps the generated tokens to pixels.

RAE token space, the backbone, the data, and the training budget, so they differ only in where the semantics enter.

(i) **Better samples.** SAS gives the best FID in every setting. In the main setting (class-conditional SiT-XL/1, ImageNet-1k, 40 epochs), its unguided FID is **2.71**, against 2.96 for conditioning, 4.12 for alignment, and 6.79 for the Gaussian source. In RAE's training setup, it lowers the 80-epoch FID from 2.18 to 2.03 (Secs. 4.1 and 4.7).

(ii) **Faster training.** In the main setting, SAS matches conditioning's 40-epoch FID after 25 epochs and passes the Gaussian source's with only 25% of the training compute. In RAE's recipe, it passes the released 80-epoch model by epoch 50 (Sec. 4.2).

(iii) **Fewer sampling steps.** SAS roughly halves the squared distance the flow must cover, so it holds up with few sampling steps. With 8 steps, it matches the FID that RAE's released 80-epoch model reaches with 50 (Sec. 4.3).

To summarize, we introduce semantics-as-source (SAS), which uses a noise-perturbed semantic embedding, instead of pure noise, as the starting point of a generative flow. Results show that SAS gives better samples, trains faster, needs fewer sampling steps, and combines with other advances, reaching an FID of 1.86 after 160 epochs with a decoupled head and autoguidance on Imagenet 256^2. (Sec. 4.8).

2 Related Work

Flow matching and interpolants. Flow matching and stochastic interpolants (Lipman et al., 2023; Liu et al., 2022; Albergo & Vanden-Eijnden, 2022; Albergo et al., 2025) learn a velocity field transporting a source distribution to data along prescribed interpolation paths, generalizing score-based diffusion (Ho et al., 2020; Song et al., 2020). Most instantiations, including the transformer backbones we build on (Peebles & Xie, 2023; Ma et al., 2024), keep the source Gaussian. The framework nevertheless allows arbitrary sources and couplings. In particular, Albergo et al. (2023) formalize data-dependent couplings in which source samples are drawn conditionally on the target, while retaining a square-loss velocity regression objective. Our paired semantic source is an instance of this framework, with the coupling determined by a frozen image representation.

Non-Gaussian sources, bridges, and couplings. Bridge models connect prescribed endpoint distributions (Liu et al., 2023; Zhou et al., 2024; De Bortoli et al., 2021), while minibatch-OT methods improve noise–data couplings without changing the source marginal (Tong et al., 2023; Pooladian et al., 2023). Informative sources can use inference-time conditions: class statistics (Issachar et al., 2025), learned shifts in CAR-Flow (Chen et al., 2026), or text representations in CSFM and Cross-Flow (Kim et al., 2026; Liu et al., 2025); adaptive conditional priors also appear in speech models such as PriorGrad (Lee et al., 2021). Closer to our construction, LeDiFlow (Zwick et al., 2025) uses an auxiliary regressor to produce an initial distribution near the target, and MM-FM (Luo

et al., 2026) fits a Gaussian mixture using frozen DINOv2 representations and couples each target to its assigned mode in RAE latent space. Our method broadcasts each training image's continuous, mean-pooled token embedding and adds noise, yielding a spatially constant source mean without a learned map or mode-specific statistics. A learned embedding prior supplies this vector at inference for unconditional or class-conditional generation. This construction lets us compare source injection with conditioning and alignment under a shared encoder, latent space, and training budget.

Semantics for generation: conditioning, alignment, and latent spaces. RCG (Li et al., 2024) conditions a generator on SSL embeddings sampled from a representation diffusion model; its conditioning mechanism motivates our COND arm, and we use an RCG-style prior shared by the embedding-dependent arms. REPA (Yu et al., 2024) aligns intermediate diffusion features with a frozen encoder's tokens, motivating our ALIGN arm; REPA-E extends alignment to end-to-end tokenizer training (Leng et al., 2025). RAE (Zheng et al., 2026) makes the frozen SSL token grid the diffusion space, while LightningDiT/VA-VAE (Yao et al., 2025) aligns the latent space with SSL features during tokenizer training. These Gaussian-source systems motivate our NOISE baseline and the choice to hold the latent space fixed. We instantiate the routes in RAE space and additionally test a source swap in RAE's tuned DiT-DH recipe (Sec. 4.7). Source design also interacts with the representation: CSFM reports larger gains in RAE than in SD-VAE latents (Kim et al., 2026). Other placements include jointly generating image latents and a semantic token, as in REG (Wu et al., 2026). Alignment can also complement representation autoencoders (Singh et al., 2026).

Transport geometry and few-step sampling. Informative sources have been linked to shorter transport paths and fewer sampling steps (Issachar et al., 2025; Chen et al., 2026; Luo et al., 2026). We examine this relationship across the source, conditioning, and alignment routes by reporting chord energy alongside raw and chord-normalized velocity-to-chord deviation. This distinguishes smaller endpoint displacement from lower deviation relative to that displacement, helping interpret improvements in unnormalized trajectory statistics. The measurements describe learned trajectories under a common integration protocol; their relationship to generation quality is evaluated through the accompanying sampling curves. Our few-step comparisons use undistilled base flows with a shared solver and timestep schedule. Methods such as MeanFlow (Geng et al., 2026), which learns an average-velocity field for one-step generation, address sampling efficiency through the training objective, providing a separate axis from the source and coupling choices studied here.

3 SEMANTICS AS SOURCE

3.1 PRELIMINARIES: FLOW MATCHING IN A SEMANTIC TOKEN SPACE

Flow matching learns a velocity field $v_\theta(x_t, t)$ that transports a source distribution p_0 to a data distribution p_1. Given a coupling $(z_0, x_1) \sim \pi$, flow matching sets $x_t = (1-t)\, z_0 + t\, x_1$ and minimizes

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\,(z_0,x_1)\sim\pi} \left\| v_\theta(x_t, t) - (x_1 - z_0) \right\|^2, \tag{1}$$

with $t = 0$ the source end. Eq. 1 is valid for any p_0 and any coupling π. The standard choice is $p_0 = \mathcal{N}(0, I)$ with the independent coupling $\pi = p_0 \otimes p_1$, which is what SAS changes.

We work in the token space of a representation autoencoder (Zheng et al., 2026). The data point is the frozen-encoder token grid $x_1 \in \mathbb{R}^{256\times 768}$ of a 256^2 image (DINOv2-B; Oquab et al., 2023; Darcet et al., 2024), normalized with RAE's released per-channel statistics, and decoded to pixels by RAE's released frozen decoder. Its per-dimension energy after normalization is $\bar{E} := \mathbb{E}\|x_1\|^2/D = 1.30$ on ImageNet-1k, with $D = 256 \cdot 768$.

3.2 THE SOURCE CONSTRUCTION

Let $e = \frac{1}{256}\sum_j x_{1,j} \in \mathbb{R}^{768}$ be the spatial mean of the image's own normalized tokens. SAS starts each trajectory at a noise-augmented broadcast of this vector:

$$z_0 = \mathrm{bc}(e) + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_{256\times 768}), \tag{2}$$

where $\mathrm{bc}(e)$ copies e to all 256 positions. Let $\rho := \mathbb{E}\|\mathrm{bc}(e)\|^2/\mathbb{E}\|x_1\|^2$ be the fraction of the space's energy in the constant subspace. We set $\sigma^2 = (1-\rho)\bar{E}$, which gives $\mathbb{E}\|z_0\|^2 = \mathbb{E}\|x_1\|^2$, so the

source has the data's energy and no renormalization is needed. We find the noise term necessary as embedding-to-image is one-to-many, and a deterministic source would regress to conditional means.

The SAS construction yields the following four identities, which explain its effectiveness relative to a Gaussian start:

(a) The source supplies global semantics. SAS splits each token grid into its spatial mean and the residual around it. The mean e carries the image's global semantics (a linear classifier trained on frozen e reaches 94.0% top-1 accuracy on ImageNet-100), and the source already contains it. The target is therefore $v^* = d - \sigma\varepsilon$, where the model does not need to learn to generate the global content and can focus on the spatial residual d.

(b) Half the transport energy. The target in Eq. 1 is the displacement $x_1 - z_0$, so its expected energy sets both the scale of what the network regresses and the distance the sampler needs to integrate. For SAS it is $\mathbb{E}\|x_1 - z_0\|^2/D = (1-\rho)\bar{E} + \sigma^2 = 2(1-\rho)\bar{E} = 1.31$, against $2\bar{E} = 2.60$ for a Gaussian source of the same energy with independent coupling (and $\bar{E} + 1 = 2.30$ for a unit Gaussian). The ratio to the matched Gaussian is exactly $1 - \rho \approx 0.5$, as pairing removes the share of the data's energy that lives in the spatial mean, because the source already sits there. This especially benefits few-step sampling, as Euler's per-step error grows with how much the velocity changes along the step, which scales with the displacement, so of two fields with the same relative curvature, the one with the shorter chord incurs less absolute error at every step count.

(c) Conservation. The source's semantics do not fade after $t{=}0$. Instead they persist along the whole path. The spatial mean of the interpolant is $\frac{1}{256}\sum_j x_{t,j} = e + (1-t)\,\sigma\bar{\varepsilon}$, where the averaged noise $\bar{\varepsilon}$ has per-dimension std $1/16$. Mean-pooling the input therefore recovers e at every t, with per-dimension SNR $16\sqrt{\rho/(1-\rho)} \approx 16$. The target's global semantics are thus directly readable from the input at every timestep, so the loss only has to explain the residual $p(d \mid e)$. Under a Gaussian source the same readout is $t\,e + (1-t)\bar{\varepsilon}$, which carries no information at $t{=}0$ and stays weaker than SAS's for roughly the first half of the path where the field must average over all images.

(d) No shortcut. Handing the model part of the answer at $t{=}0$ raises a natural concern: it might learn to copy the source rather than to generate. SAS rules this out by giving away only the spatial mean. The source fixes the image's global semantics e, which is identical at every token position, but none of its spatial residual d: everything that varies across positions (layout, pose, texture) starts as fresh noise that is unrelated to the image. Along the path, this noise and the true detail d are mixed in the same token positions, and the model cannot separate them without having learned what plausible detail looks like for a given e. Outside the mean, SAS is therefore ordinary noise-to-data flow matching, conditioned on e. The training signal is not spent re-deriving global content the input already contains, nor can it be satisfied by copying, so it goes to the part of generation that actually needs learning. The training loss reflects this: it starts at the transport energy of (b), $\bar{L} = 1.31$, and in every source-variant run of the route comparisons it stays above $0.1\bar{L}$ instead of collapsing toward zero (Sec. 4.6).

3.3 Sampling: a learned prior over embeddings

At inference the image's own e is unavailable, so the pipeline samples it: $e \sim p_\phi(e)$, or $p_\phi(e \mid y)$ in class-conditional settings; then $z_0 = \mathrm{bc}(e) + \sigma\varepsilon$, a 50-step Euler ODE, and the frozen decoder. The prior is an RCG-style representation diffusion model (Li et al., 2024), which is a residual MLP trained by flow matching on the 1.28M cached embeddings. Its training cost is $<5\%$ of a backbone run.

3.4 What SAS changes relative to a Gaussian start

The deployed sampler also begins with Gaussian noise: $\varepsilon_e \in \mathbb{R}^{768}$ for the prior and $\varepsilon \in \mathbb{R}^{256\times768}$ for the detail term. Composed, the pipeline is a map from Gaussian noise to images, as is the noise variant. What differs is where the noise enters and what the backbone flow has to do with it.

Coupling. The noise variant pairs each training image with an independent draw, so every image is a full-length target from an unrelated starting point and the field must average over all such pairs. SAS pairs each image with its own embedding. By identity (b) this removes about half of the expected transport energy, and by (c) the endpoint's content is fixed at $t = 0$ rather than resolved along the

Table 1: **Semantics as source is the best route in every setting** (FID-10K at the final step). Variants differ only in where the semantics enter.

Setting	Steps	SAS (ours)	COND	ALIGN	NOISE
SiT-B/1, IN-1k	400K	**22.75**	23.54	33.45	54.83
SiT-B/1-cc, IN-1k	200K	**18.21**	20.52	23.37	35.38
SiT-XL/1, IN-1k	200K	**7.19**	7.45	10.52	17.81
SiT-XL/1-cc, IN-1k	200K	**5.75**	5.99	7.10	7.79

Table 2: **Official evaluation in the main setting** (SiT-XL/1-cc, 40 epochs, no guidance, embeddings from the learned prior). Adding alignment on top of the source route gives the best result.

Variant	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
SAS (ours)	2.71	5.81	226.2	0.830	0.561
SAS+ALIGN (ours)	**2.58**	**5.60**	**229.5**	0.832	**0.564**
COND	2.96	6.31	225.4	0.833	0.552
ALIGN	4.12	7.22	201.3	0.838	0.511
NOISE	6.79	8.18	169.6	0.825	0.537

way. The shuffled-pair control (Sec. 4.4) isolates this effect, which shows that with the pairing destroyed and the source marginal $p_0 = \int \mathcal{N}(\mathrm{bc}(e), \sigma^2 I)\, dp(e)$ unchanged, the variant falls below the Gaussian-source variant. The difference from a Gaussian start is the coupling, not the shape of p_0.

Factorization. The prior models the global structure $p(e)$ in 768 dimensions. The backbone models only $p(x_1 \mid e)$, i.e., the spatial residual d that fills in finer details around it. The noise variant folds both into one D-dimensional transport, and spends its early compute recovering the global structure that the prior supplies directly (Sec. 4.2).

4 EXPERIMENTS

This section first compares the source route with conditioning, alignment, and a Gaussian source on SiT backbones. We compare sample quality in Sec. 4.1 and training speed in Sec. 4.2. Sec. 4.3 measures transport energy and few-step sampling, and Sec. 4.4 tests whether the gain requires pairing each image with its own embedding. Sec. 4.5 examines the learned prior used at inference, and Sec. 4.6 covers the noise scale, memorization, and training diagnostics. Sec. 4.7 swaps the source into RAE's tuned DiT-DH-XL recipe, and Sec. 4.8 combines the source route with a decoupled head and autoguidance.

Setup. COND receives the same embedding from the same prior through adaLN (Li et al., 2024), ALIGN adds REPA's loss toward the same encoder (Yu et al., 2024), and NOISE is RAE's Gaussian source (Zheng et al., 2026); everything else is shared and untuned. We train SiT-B/1 and SiT-XL/1 (Ma et al., 2024) on ImageNet-1k, with and without class labels (*-cc*). The main setting is SiT-XL/1-cc for 40 epochs. Training curves use FID-10K (10K samples) and headline rows use official FID (50K samples, ADM suite; Dhariwal & Nichol, 2021).

4.1 SEMANTICS AS SOURCE GIVES THE BEST SAMPLE QUALITY

The source route produces the best samples in every setting (Table 1), and it leads at every checkpoint from 50K steps on. Under the official protocol in the main setting (Table 2), it reaches an unguided FID of 2.71 after 40 epochs, while conditioning reaches 2.96, alignment 4.12, and the Gaussian source 6.79. It also has the best sFID, IS, and recall.

The comparison with conditioning is the most telling. Both use the same prior, embedding, and backbone, and differ only in whether the embedding starts the flow or is a side input. Making it the starting point lowers FID by 3 to 11 percent across settings and adds no learned parameters. Against the Gaussian source, which is RAE's configuration, the gap is much larger: its FID is 35 percent higher than ours in the main setting and 1.9 to 2.7 times higher elsewhere. A semantic token space with a Gaussian source leaves a large margin, and the source route collects it.

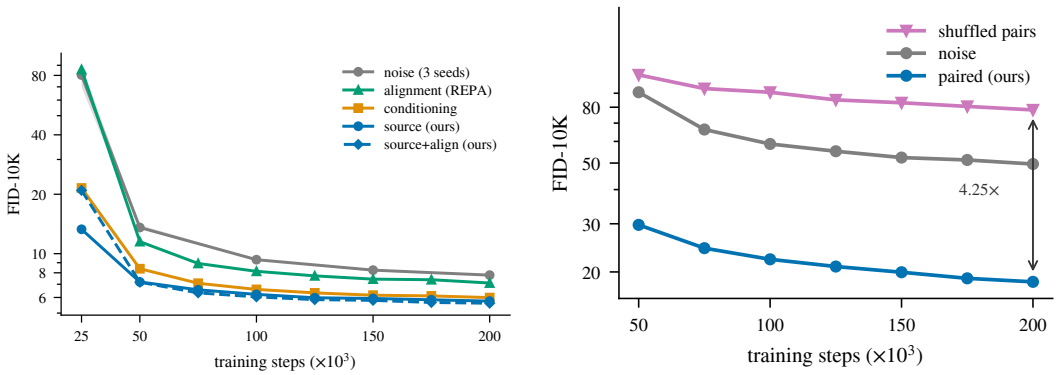


Figure 2: **Left: the source route leads throughout training** (main setting, FID-10K over training steps). Adding alignment to the source route (dashed) starts slower but is ahead from 50K steps on. **Right: the pairing matters** (SiT-B/1, ImageNet-100). Shuffling embeddings across images keeps the source distribution but breaks the pairing; the shuffled variant falls behind even the noise variant after 75K steps.

Alignment is the weakest semantic variant, landing between conditioning and the Gaussian source everywhere; in this space its target is the encoder that already defines the data. It still helps on top of the source route. Following RAEv2's observation that alignment complements a representation-autoencoder space (Singh et al., 2026), we added REPA's unchanged loss to SAS in the main setting. The combined model starts slower (FID-10K of 20.92 at 25K steps, where plain SAS is already at 13.29), catches up by 50K, and finishes ahead at 5.61 compared with 5.75 (Fig. 2, left; Table 5). FID improves from 2.71 to 2.58, making it our best unguided result on the plain SiT-XL/1 backbone. SAS is thus complementary to recent alignment-based improvements such as RAEv2, and the two can be combined for further gains.

4.2 Semantics as source trains faster

The source route also reaches a given quality much sooner. After only 25K steps (5 epochs) in the main setting, its FID-10K is 13.29, while conditioning is at 21.51 and the three noise seeds are between 69.8 and 90.2. It matches conditioning's final 200K-step FID at 125K steps, saving 37.5% of the training, and still saves about 13% in unconditional SiT-B/1. It passes the Gaussian source's final FID within 25% of the training compute in the main setting, 12.5% in unconditional SiT-XL/1, and 50% in class-conditional SiT-B/1 (Appendix B). Against RAE's tuned recipe, it passes the released 80-epoch model by epoch 50 (Fig. 4).

4.3 Semantics as source needs fewer sampling steps

Identity (b) of Sec. 3.2 predicts that the paired source roughly halves the transport energy, the expected squared distance from source to target, so SAS should tolerate coarse integration better than a Gaussian source. The clearest test is RAE's DiT-DH-XL recipe (Sec. 4.7), where SAS differs from the released 80-epoch model only in its source and the two reach nearly the same FID-10K with 50 steps (5.14 and 5.28; Fig. 3, top right). With only 8 steps SAS matches the released model's 50-step FID (5.29 against 5.28), and with 4 steps its FID rises by under 10 percent, while the released model's more than doubles. The degradation persists after ten times more training: the released 800-epoch model is better with 50 steps (4.63), yet its FID roughly triples with 4 steps, and with 8 steps it is worse than our 80-epoch model (6.78 against 5.29). The SiT-B/1 settings show the same pattern (Fig. 3, top left): with 16 steps SAS is within 2 percent of its own 50-step FID on ImageNet-100 and ImageNet-1k, while the noise variant is still 26 to 40 percent worse.

We also measure the chord energy $\|\hat{x}_1 - z_0\|^2$ and the velocity–chord deviation $\|v_\theta - (\hat{x}_1 - z_0)\|^2$ of each generated trajectory to test this prediction directly (Fig. 3, bottom). The measurements agree with the prediction: SAS's chord energy is 0.74 to 0.96, while every Gaussian-source variant has 1.44 to 1.88. Its raw deviation is also the lowest at every scale, while its deviation relative to chord energy is on par with the noise variant. The measurements imply that SAS makes few-step

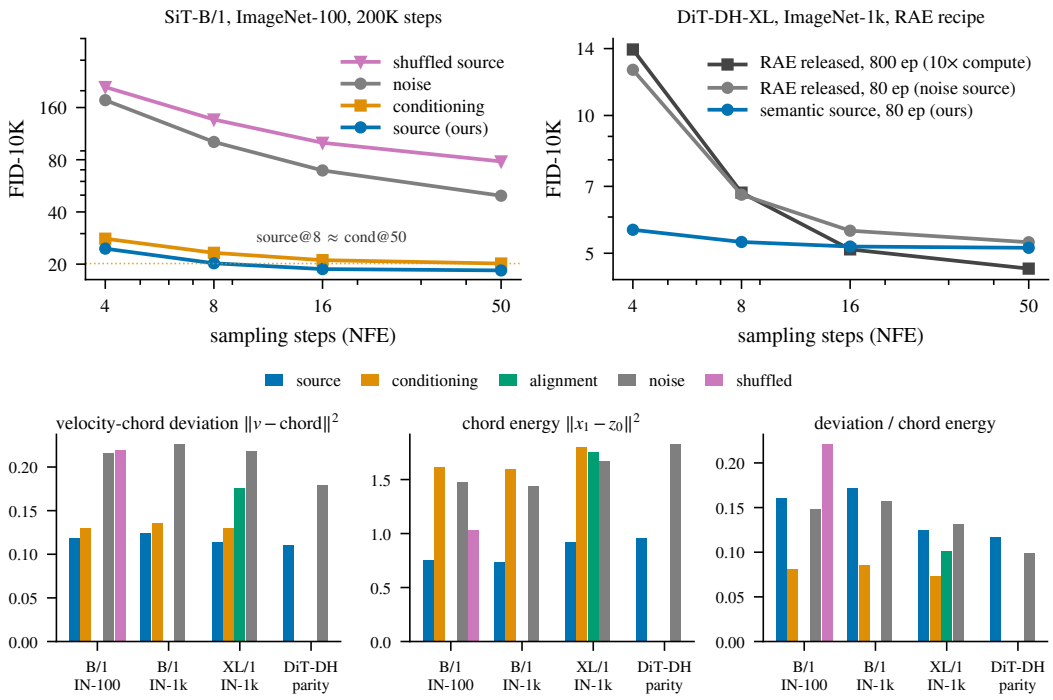


Figure 3: **Top: few-step sampling** (FID-10K by number of Euler steps). Left: ImageNet-100 variants at 200K steps; at 4 steps the noise variant's FID more than triples, while SAS's rises by about a third. Right: on RAE's DiT-DH-XL recipe (Sec. 4.7), SAS and the released 80-epoch model are nearly tied at 50 steps, but the released noise-source models degrade sharply at low step counts after 80 *and* 800 epochs; SAS does not. **Bottom: the source route shrinks the displacement, not the deviation per unit of it.** Velocity–chord deviation (left) is lowest for SAS, then conditioning, then noise. SAS's chord energy (middle) is about half that of the Gaussian-source variants, while normalized by chord energy (right) conditioning deviates least. The shuffled variant has the largest deviation on both measures.

sampling easier than a Gaussian source does by shortening the path. It keeps the noise variant's relative deviation over roughly half the displacement, and so incurs less absolute error at every step count, as identity (b) predicts.

4.4 The gain comes from the coupling, not the shape of the source

Relative to the noise variant, SAS changes two things at once. It changes the geometry of the source distribution, replacing the isotropic Gaussian with the semantically structured marginal $p_0 = \int \mathcal{N}(\mathrm{bc}(e), \sigma^2 I)\, dp(e)$, and it changes the coupling π, pairing each image with its own embedding instead of an independent draw. To better understand the effects of the two changes, we run a control experiment on ImageNet-100, which keeps the source geometry and removes only the coupling: embeddings are shuffled within each batch, so the source marginal and the architecture match SAS exactly, but no image starts from its own embedding. The shuffled variant therefore differs from the noise variant only in geometry, and from SAS only in coupling. Geometry alone does not help (Fig. 2, right): from 75K steps on, the shuffled variant is worse than the Gaussian source. The coupling accounts for the whole gain: the shuffled variant ends at an FID of 78.2 and SAS at 18.4, and the gap widens from three and a half times at 50K steps to more than four times at 200K. The benefit of a semantic source thus comes from the coupling, not from the shape of p_0: the flow benefits from starting at the semantics of the image it produces.

4.5 The learned prior is not a bottleneck

At inference SAS draws its starting embedding from a lightweight learned prior rather than from a real image. A natural concern is whether this prior captures the space of pooled embeddings

Table 3: **Replacing only the source in RAE's DiT-DH-XL** (official FID-50K, no guidance). The 2.03 row is the seed-0 run at epoch 80. A second training seed is below RAE's reported 2.16 from epoch 50 onward and reaches 2.09 at epoch 80.

Model	Ep.	FID↓	sFID↓
RAE released	80	2.18	5.18
+ sem. source	80	**2.03**	5.08

Table 4: **Comparison with prior work** (official FID-50K). Our models are trained without guidance; AG applies autoguidance at inference.

Model	Guid.	Ep.	FID↓
DiT-XL/2 (Peebles & Xie, 2023)	CFG	1400	2.27
SiT-XL/2 (Ma et al., 2024)	CFG	1400	2.06
LightningDiT (Yao et al., 2025)	none	800	2.17
RAE (Zheng et al., 2026)	none	80	2.16
Ours, RAE recipe (parity)	none	80	2.03
Ours + DDT head	none	80	2.32
Ours, frozen-T+DDT	AG	160	**1.86**

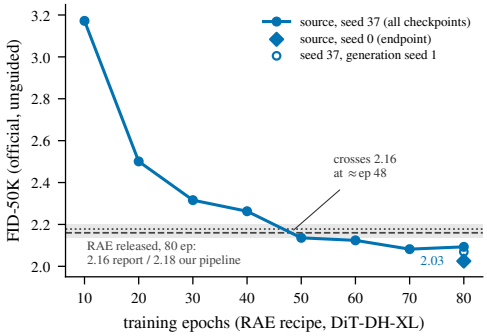


Figure 4: **SAS stays below RAE's released model** (official FID-50K, no guidance). Every checkpoint of one training seed at RAE's release cadence, plus the seed-0 run at epoch 80. The band shows RAE's released 80-epoch model with the 0.023 generation-seed spread on either side.

accurately, or whether it becomes the bottleneck of generation. In the main setting it does not limit SAS (Appendix C): the class-conditional prior gives lower FID than real embeddings at all eight checkpoints (at 200K, 5.75 compared with 6.17 for SAS, and 5.99 compared with 6.45 for conditioning). Real embeddings led only in unconditional SiT-B/1 with our first prior (FID 18.51, where the prior gave 22.75); replacing just the prior, without retraining the flow, brought it to 18.41. The prior is a small, swappable module whose quality can be checked in embedding space before the flow is trained.

4.6 Robustness

SAS's one hyperparameter, the noise scale σ, has a broad optimum. At half, one, and two times the energy-matched σ^* of Sec. 3.2, σ^* gives the best prior-embedding FID at every checkpoint, and all three converge after 150K steps (3.21, 3.27, and 3.29 with 50K samples), so σ mainly affects early training speed (Appendix B). The models do not copy training data: in the space of I-JEPA (Assran et al., 2023), samples from every unconditional SiT-B/1 variant are less similar to their nearest training image than real held-out images are. From 50K steps on, every source-variant run in the route comparisons also keeps the signatures of Sec. 3.2: loss above a tenth of its opening value, mean tokens aligned with the source embedding (median cosine at least 0.9), and mean-channel velocity energy near $\sigma^2/256$.

4.7 A drop-in replacement for RAE's Gaussian source

We further test whether SAS carries over to a more recent architecture, RAE's DiT-DH-XL. We do not use it for the main study for two reasons. Its additional components, a wide DDT head and a training recipe tuned for the Gaussian source, could interact differently with each route and blur what the comparison attributes to where the semantics enter. The wide head also raises the per-step cost, and training every route in every setting with it would exceed our compute budget. We ported RAE's stage-2 code, checked it on their released checkpoint (2.18 under our pipeline, where RAE reports 2.16), and retrained it from scratch with one change: each trajectory starts at $\mathrm{bc}(e) + \sigma\varepsilon$ instead of a Gaussian sample. Within RAE's 80-epoch budget this reaches an official FID of 2.03 (Table 3, Fig. 4), 0.15 lower than the released model under the same pipeline and more than six times the generation-seed spread.

The smaller transport energy carries over too: chord energy falls from 1.83 to 0.96, and with few sampling steps our model degrades far less than the released models after 80 and 800 epochs (Sec. 4.3).

4.8 BUILDING ON THE SOURCE ROUTE

Because SAS changes only where the flow starts, the rest of the pipeline remains free to improve. We stack three improvements on the class-conditional SiT-XL/1 setting: a decoupled head, a learned source, and a stronger form of guidance. First, a 2048-wide DDT head (Wang et al., 2026) brings unguided SAS to an official FID of 2.32 at 80 epochs. Second, we replace the broadcast with $T(e) = \mathrm{bc}(e) + R(e)$, where a regressor R, trained separately to predict the spatial residual from e, is kept frozen. This learned source is slightly worse than the plain broadcast without guidance but the best choice under guidance. Third, autoguidance with an early checkpoint of the same run (Karras et al., 2024) beats every classifier-free guidance setting we searched. Together, they reach an official FID of 1.86 at 160 epochs (1.88 for a replicate; Table 4), lower than guided DiT-XL/2 and SiT-XL/2 with roughly one ninth of their training updates.

5 DISCUSSION AND LIMITATIONS

What the source route buys. Starting from semantics changes the problem the flow solves. SAS directly uses the global semantics available from the pretrained encoder, and a prior samples them in 768 dimensions, where that choice is cheap. The flow then only needs to learn the spatial residual for finer details, from a start that already removes about half the transport energy and keeps the global semantics as a conserved quantity. Conditioning supplies the same embedding as a side input, but its transport still starts from noise; alignment shapes features toward semantics the model must still discover, and in a space that already is the encoder's it has little left to add, though it still helps on top of the source route. Our shuffled control experiment shows that the gains depend on the pairing. Our results suggest that in a semantic token space, start the flow from the semantics is beneficial. This adds no parameters to the flow, holds up with few sampling steps, and combines with alignment and with RAE's recipe.

Limitations and future work. All of our experiments use ImageNet at 256^2 resolution, so text-conditioned and higher-resolution generation remain untested. Our compute budget also limited the controlled comparisons to SiT backbones, so we have not compared SAS with conditioning and alignment on more recent model architectures. Finally, the training speedups vary across settings and become smaller against well-tuned baselines: SAS needs 62.5 to 87 percent of the steps that conditioning needs, and it overtakes RAE's released model at epoch 50 of 80. For future work, a natural extension is text-to-image generation, where the prior maps a caption to an embedding and the flow stays unchanged. The source could also carry more than the pooled mean, for example a coarse spatial grid of token means, which would move layout from the flow to the prior and leave the flow an even smaller residual to learn.

6 CONCLUSION

We introduced semantics-as-source (SAS), which starts a generative flow from a noise-perturbed semantic embedding instead of pure noise. With the same encoder, backbone, and training budget, SAS outperforms conditioning and alignment in every setting we train and trains faster. It roughly halves the transport energy and holds up with few sampling steps where a Gaussian source does not, and a shuffled-pair control shows that the gain comes from pairing each image with its own embedding. Because SAS changes only where the flow starts, it combines with alignment and with RAE's tuned DiT-DH-XL recipe, where it lowers the 80-epoch FID from 2.18 to 2.03, and with a decoupled head and autoguidance it reaches 1.86 after 160 epochs. These results suggest that pretrained semantic representations are most useful not as guidance for a flow that starts from noise, but as the point where the flow starts.

AI USE STATEMENT

In this work, we used generative AI tools to aid in writing the paper and in developing the code for our experiments. We have reviewed all AI-assisted writing, and also verified the correctness of all AI-assisted code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Michael S Albergo, Mark Goldstein, Nicholas M Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv preprint arXiv:2310.03725*, 2023.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Chen Chen, Pengsheng Guo, Liangchen Song, Jiasen Lu, Rui Qian, Tsu-Jui Fu, Xinze Wang, Wei Liu, Yinfei Yang, and Alex Schwing. Car-flow: Condition-aware reparameterization aligns source and target for better flow matching. *Advances in Neural Information Processing Systems*, 38: 94919–94945, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a conditional prior distribution for flow-based generative models. *arXiv preprint arXiv:2502.09611*, 2025.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching. *arXiv preprint arXiv:2602.05951*, 2026.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. *arXiv preprint arXiv:2106.06406*, 2021.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37: 125441–125468, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image Schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2755–2765. IEEE, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Gaoxiang Luo, Frank Cole, Sihang Zhang, Yuxiang Wan, Yulong Lu, and Ju Sun. Flow matching for multimodal distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23260–23271, June 2026.

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*, pp. 23–40. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 40633–40642, 2026.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

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. *arXiv preprint arXiv:2410.06940*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, volume 2024, pp. 8160–8171, 2024.

Pascal Zwick, Nils Friederich, Maximilian Beichter, Lennart Hilbert, Ralf Mikut, and Oliver Bringmann. Lediflow: Learned distribution-guided flow matching to accelerate image generation. *arXiv preprint arXiv:2505.20723*, 2025.