PERCEPTUAL FLOW MATCHING FOR FEW-STEP GENERATIVE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose **Perceptual Flow Matching** (**PFM**), a simple yet effective framework for few-step generation in flow-matching models. Rather than performing velocity regression in the conventional VAE latent space, PFM supervises flow matching in a *perceptual feature space* using pretrained perceptual models. This simple change substantially improves the few-step generation capability of flow-matching models, reducing the number of sampling steps from 35–50 to 4–8 while preserving generation quality. Unlike existing acceleration and distillation approaches, PFM requires neither teacher models nor auxiliary score networks and can be integrated into standard flow-matching training pipelines with minimal modifications. Extensive experiments on image generation, video generation, and image editing tasks demonstrate that PFM consistently produces high-quality results while producing fewer artifacts than existing distillation-based methods. We further show that perceptual supervision shifts the regression minimizer from mean-seeking to mode-seeking, biasing predictions toward on-manifold modes that remain accurate under coarse few-step integration. Our results reveal that standard flow-matching training can naturally yield high-quality few-step generators when supervised in an appropriate representation space. We hope this insight inspires future research into representation-aware objectives for efficient generative modeling.

1 INTRODUCTION

Flow matching has emerged as a powerful framework for generative modeling and has achieved remarkable success across image generation, image editing, and video generation (Lipman et al., 2023; Labs et al., 2025; Esser et al., 2024; Liu et al., 2023; Wan et al., 2025; Wu et al., 2025; Bai et al., 2026). However, despite its impressive generation quality, practical deployment remains challenging due to the high computational cost of inference. Similar to diffusion models, flow matching models typically require dozens of sampling steps (e.g., 35–50 steps) to produce high-quality outputs, resulting in substantial latency and computational overhead.

To address this challenge, prior works have explored few-step generation through distillation (Salimans & Ho, 2022; Yin et al., 2024b;a; Dao et al., 2025), consistency training (Song et al., 2023), and continuous-time flow field modeling (Lu & Song, 2025; Geng et al., 2025; Sabour et al., 2025). While these approaches have demonstrated promising results, they often require auxiliary models, or substantial modifications to the original training framework, increasing both computational cost and implementation complexity. Moreover, distillation-based approaches can be prone to overfitting and mode collapse, which leads to undesirable synthetic artifacts (Geng et al., 2025).

We propose Perceptual Flow Matching (PFM), a simple yet effective framework for few-step generation. Rather than supervising flow matching through conventional velocity regression in the VAE latent space, PFM formulates the training objective as a perceptual regression loss on decoded samples, using feature representations extracted from pretrained perceptual models (Zhang et al., 2018; Oquab et al., 2023). We further show that classifier-free guidance can be naturally incorporated during training (Ho & Salimans, 2022), eliminating the need for an additional unconditional forward pass at inference time. Notably, PFM requires no teacher models, auxiliary score networks, or distillation procedures. It retains the original flow matching framework (Lipman et al., 2023) and only replaces the supervision space, making it easy to implement and reducing training cost. As PFM is trained directly on real data and closely follows the standard flow matching training paradigm,

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.

requiring only a change in the supervision space, it is scalable to large-scale training and produces fewer synthetic artifacts compared to distillation-based methods.

The key insight we find is that the supervision space determines the geometry of the one-step prediction. Under the standard Euclidean objective, flow matching is effectively mean-seeking: when the posterior $q(x_0 \mid x_t)$ is multimodal at high noise levels, the optimal prediction lies between plausible data modes, producing an off-manifold and blurry estimate that is difficult to correct with only a few sampling steps. In contrast, pretrained perceptual models encode the structure of natural images in perceptual space, assigning larger distances to off-manifold degradations such as blur while keeping valid images in coherent manifold. As a result, perceptual supervision makes blurry averages less favorable and biases the prediction toward perceptually valid samples, substantially improving few-step generation.

We conduct extensive experiments on image generation, image editing, and video generation tasks to validate the effectiveness of PFM. On the text-to-image COCO 2014 val benchmark, PFM achieves state-of-the-art performance among 8-step generation methods, attaining an FID of 33.93, a CLIP score of 31.70 and a HPSv3 score of 11.42. On image editing and video generation benchmarks, PFM substantially outperforms its few-step baseline counterparts, while achieving results comparable to, and in some cases surpassing, the original multi-step models. Our results suggest that standard flow-matching training can naturally yield strong few-step generators when optimized in an appropriate representation space. We hope this insight motivates further investigation into the role of supervision space in generative modeling.

2 Related Work

2.1 Few-Step Generation and Distillation

Reducing the sampling cost of diffusion and flow-based generative models has been a central research topic in recent years. Existing approaches can largely be categorized into four paradigms. Progressive Distillation (Salimans & Ho, 2022) accelerates generation by repeatedly training a student model to emulate multiple sampling steps of a pretrained teacher, gradually reducing the number of required inference steps. Consistency Models and Consistency Distillation (Song et al., 2023) learn a trajectory-consistent mapping that enables one-step or few-step generation through teacher-generated trajectories and consistency constraints. Distribution Matching Distillation (DMD) (Yin et al., 2024b) and its variants (Yin et al., 2024a; 2025) align the distribution of student-generated samples with those of a stronger teacher through additional score estimators, critics, or auxiliary objectives. More recently, methods such as MeanFlow (Geng et al., 2025) have explored alternative training targets for flow matching, replacing the conventional instantaneous velocity objective with an average-velocity formulation and enforcing consistency between the two through Jacobian-vector product (JVP).

2.2 Perceptual Supervision

Perceptual losses (Johnson et al., 2016; Zhang et al., 2018) have been widely adopted in image synthesis, image restoration, and representation learning to improve perceptual quality. Existing works (Yu et al., 2024; Leng et al., 2025; Wu et al., 2026) primarily employ perceptual supervision to improve reconstruction fidelity or visual realism. In contrast, we investigate perceptual supervision as a mechanism for enabling few-step generation. Our results suggest that the invariances encoded by perceptual representations fundamentally improve the few-step capability of flow matching models, eliminating the need for explicit distillation.

3 Method

3.1 Flow Matching Preliminaries

Flow Matching (FM) (Lipman et al., 2023) learns a continuous transport map between a simple noise distribution and the data distribution. Given a clean latent representation x_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, an interpolated sample is constructed as:

$$x_t = (1 - \sigma_t)x_0 + \sigma_t\epsilon, \tag{1}$$

where $\sigma_t \in [0, 1]$ denotes the noise level associated with timestep t. The corresponding target velocity field is:

$$v(x_t, t) = \epsilon - x_0. \tag{2}$$

A flow matching model $v_\theta(x_t, t, c)$ is trained to predict this velocity field by minimizing a mean squared error objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|v_\theta(x_t, t, c) - v(x_t, t)\|_2^2\right], \tag{3}$$

where c denotes optional conditioning information such as text prompts or reference images. Given a predicted velocity field, a clean sample estimate can be recovered by inverting from Eq. 1 as:

$$\hat{x}_0 = x_t - \sigma_t v_\theta(x_t, t, c). \tag{4}$$

Conventional flow matching applies supervision directly to the velocity prediction in Eq. 3. In contrast, our method revisits the supervision space used during training and instead supervises the recovered clean prediction $\hat{x}_0$ in a perceptual feature space.

3.2 Perceptual Flow Matching

Conventional flow matching supervises the predicted velocity field through a regression objective in VAE latent space. In this work, we revisit the supervision space used during training and propose **Perceptual Flow Matching (PFM)**, which decodes the recovered latent prediction into pixel space and applies supervision in a perceptual feature space.

Given a noisy latent x_t, the model predicts a velocity field $v_\theta(x_t, t, c)$ and recovers a clean sample estimate $\hat{x}_0$ via Eq. 4. Both the prediction $\hat{x}_0$ and the ground-truth sample x_0 are then decoded into pixel space using a pretrained VAE decoder $D(\cdot)$:

$$\hat{y}_0 = D(\hat{x}_0), \qquad y_0 = D(x_0). \tag{5}$$

Instead of comparing $\hat{x}_0$ and x_0 in latent space, PFM measures the discrepancy of their decoded images in a perceptual feature space induced by $\phi(\cdot)$:

$$\mathcal{L}_{\text{PFM}} = \mathbb{E}_{x_0,\epsilon,t}\left[d\big(\phi(\hat{y}_0), \phi(y_0)\big)\right], \tag{6}$$

where $d(\cdot, \cdot)$ denotes a perceptual distance metric.

At inference time, PFM models can be sampled using standard consistency sampling (Song et al., 2023; Yin et al., 2024b): at each step, the model predicts $\hat{x}_0$ from x_t via Eq. 4, which is then re-noised to obtain $x_{t'}$ for the next timestep. The number of sampling steps can be freely adjusted without retraining, allowing users to trade off quality and speed as needed.

3.3 Classifier-free Guidance

Classifier-free guidance (CFG) (Ho & Salimans, 2022) is a widely adopted inference-time technique for improving generation quality and condition alignment. Given conditional and unconditional predictions, CFG constructs a guidance-enhanced prediction as:

$$v_{\text{cfg}} = v_\theta(x_t, t, \varnothing) + w\Big(v_\theta(x_t, t, c) - v_\theta(x_t, t, \varnothing)\Big), \tag{7}$$

where w denotes the guidance scale. While effective, CFG requires an additional unconditional network evaluation at every sampling step, which increases inference cost. We explore two strategies to bake classifier-free guidance in training:

Prediction-side. Inspired by GFT (Chen et al., 2025), we incorporate guidance directly into the model's prediction before computing the perceptual loss:

$$\tilde{v} = \beta\, v_\theta(x_t, t, c) + (1 - \beta)\, \text{sg}\big[v_\theta(x_t, t, \varnothing)\big], \quad \beta = 1/w, \tag{8}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator and $v_\theta(x_t, t, \varnothing)$ denotes the unconditional prediction. The perceptual loss is then applied to $\hat{x}_0$ recovered from $\tilde{v}$ via Eq. 4, so that gradients flow through the conditional branch while the model learns to internalize the effect of guidance.

Algorithm 1 Perceptual Flow Matching (PFM) with optional CFG baking

```
# ========================================
# PFM
# ========================================
v_pred = model(x_t, t, c)
x0_pred = velocity_to_x0(v_pred, x_t, t)

loss = perceptual_loss(
  decode(x0_pred),
  decode(x0)
)

# ========================================
# PFM with prediction side CFG baking
# ========================================
v_pred = model(x_t, t, c)
v_uncond = model(x_t, t, null)
v_pred = beta * v_pred + (1 - beta) * stopgrad(v_uncond)
x0_pred = velocity_to_x0(v_pred, x_t, t)

loss = perceptual_loss(
  decode(x0_pred),
  decode(x0)
)
```

Target-side. Alternatively, guidance can be baked into the supervision target rather than the model's prediction. We construct a CFG-enhanced target and directly optimize the model toward it:

$$x_{\text{cfg}} = \alpha x_0 + (1-\alpha)\Big[\hat{x}_{0,\varnothing} + w\big(\hat{x}_{0,c} - \hat{x}_{0,\varnothing}\big)\Big], \tag{9}$$

where $\hat{x}_{0,c}$ and $\hat{x}_{0,\varnothing}$ denote the clean-sample estimates reconstructed from the conditional and unconditional predictions, respectively. $\alpha \in [0,1]$ controls the interpolation between the ground-truth sample and the CFG-enhanced target for training stability. The perceptual loss (Eq. 6) is then computed using x_{cfg} as the supervision target, enabling guidance-enhanced generation in a single forward pass at inference.

Both strategies can be incorporated into the standard flow matching framework with minimal modification, requiring only an additional unconditional forward pass during training. Empirically, however, we find that PFM already substantially reduces the need for classifier-free guidance at inference time (Section 4.3). Models trained solely with the perceptual objective can generate satisfactory results without classifier-free guidance. We attribute this to the learned perceptual supervision geometry, which biases the one-step prediction toward semantically valid samples. CFG baking can be optionally used to further strengthen the guidance effects.

3.4 Why Perceptual Space Works?

During training, standard flow matching minimizes the squared Euclidean distance between the model's one-step prediction of $\hat{x}_0$ and the true x_0[1]; the minimizer of this prediction is the posterior mean $\hat{x}_0^\star = \mathbb{E}[x_0 \mid x_t]$. At high noise levels the posterior $q(x_0 \mid x_t)$ is multimodal, so this mean may lie between data modes, producing an off-manifold and visually blurred prediction. Many-step sampling can gradually correct such estimates through iterative evaluations, but in the few-step regime there are too few updates to fully recover from this initial off-manifold prediction.

Perceptual supervision changes the geometry of the minimizer. Under the perceptual space, the minimizer becomes a feature-space barycenter rather than the Euclidean posterior mean. Because ϕ is pretrained on large-scale natural-image data, its representation encodes the structure of valid images: natural samples tend to lie in coherent feature manifold, whereas off-manifold artifacts such as blurry interpolations are assigned disproportionately large distances. As a result, the inter-mode average is no longer a cheap solution, and the learned prediction is biased away from blurry barycenters and toward perceptually valid samples that can be realized with few sampling steps. In this sense,

[1]Standard flow matching predicts velocity v, an affine transform of x_0 ($\hat{x}_0 = x_t - tv$ on the linear path), so the two objectives are equivalent and share the minimizer $\hat{x}_0^\star = \mathbb{E}[x_0 \mid x_t]$.

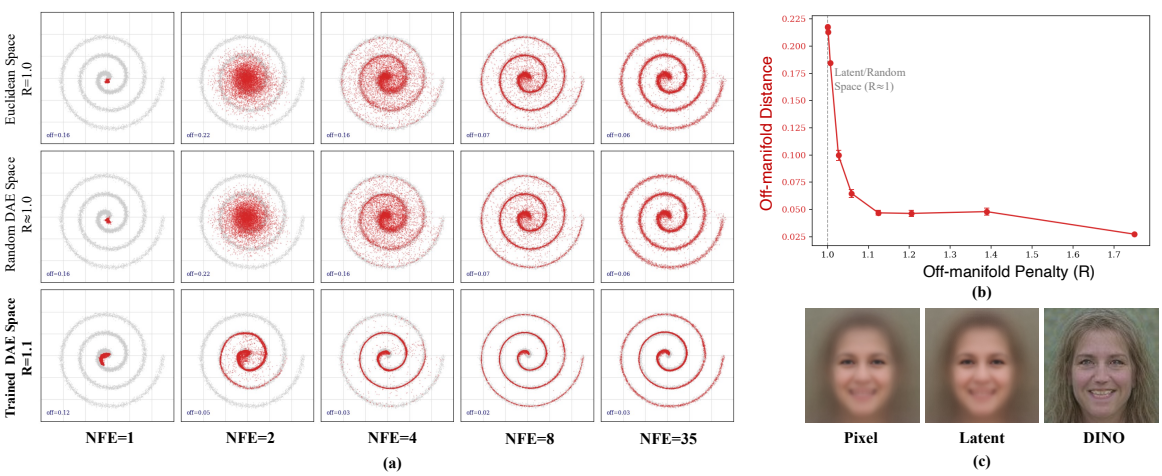


Figure 1: **(a)** Flow matching (FM) on 2D spiral data, supervised in three spaces: (i) velocity regression in Euclidean space (standard FM), (ii) x_0 regression in a random denoising-autoencoder (DAE) perceptual space, and (iii) x_0 regression in a pretrained DAE perceptual space. In Euclidean space, the prediction collapses toward the data mean at small NFE, whereas supervision in the pretrained DAE space places samples on the real data manifold even at small NFE. **(b)** Off-manifold distance (at NFE=2) as the off-manifold penalty R_ϕ increases: larger R_ϕ yields lower off-manifold distance, i.e. better few-step generation. **(c)** Average of 1,000 similar face images from FFHQ in pixel, VAE latent, and DINO feature space. The pixel and VAE-latent averages are blurry (off-manifold), whereas averaging in DINO feature space preserves sharp facial structure.

Table 1: Off-manifold penalty R_ϕ across supervision spaces. *Diff* is R_ϕ relative to the pixel space.

Supervision space	R_ϕ	Diff
VAE Latent (standard FM)	1.594	0.000
VGG	1.946	+0.352
DINO	1.940	+0.346
SigLIP	1.900	+0.306
ConvNeXt	**2.026**	**+0.432**
CLIP	1.953	+0.359

Table 2: Effect of the perceptual-space dimension D in the toy experiment. With the trained DAE feature space, increasing feature capacity raises R_ϕ and improves few-step performance.

D	R_ϕ	off-dist@NFE2 ↓	off-dist@NFE4 ↓
2	1.095	0.116	0.075
8	1.165	0.068	0.036
16	1.187	0.054	0.029
64	1.205	**0.046**	**0.027**

perceptual supervision shifts the effective behavior of regression from Euclidean mean-seeking toward perceptual mode-seeking.

To validate this mechanism, we measure how a supervision space treats off-manifold averages. Let a and b be two on-manifold samples, and let $m = \frac{1}{2}(a+b)$ denote their Euclidean midpoint. Although a and b are valid samples, their midpoint can lie off the data manifold, making m a simple proxy for the blurry, averaged image produced by an ℓ_2 regression objective. Writing the feature-space distance as $d_\phi(u,v) = \|\phi(u) - \phi(v)\|^2$, we define:

$$R_\phi(a,b) = \frac{2d_\phi(m,a)}{d_\phi(b,a)}. \tag{10}$$

Since m is the geometric midpoint of the chord $a-b$, any isometric or locally linear feature space gives $R_\phi \approx 1$. In this case, the off-manifold midpoint is treated as a cheap interpolation. In contrast, $R_\phi > 1$ indicates that ϕ expands the distance from the midpoint to the data manifold, penalizing blurry averages and pushing the barycenter toward a valid mode.

[2]Euclidean distance is the special case $\phi = \text{identity}$.

Following the manifold assumption, we conduct a toy experiment. The underlying data lie on a smooth one-dimensional manifold (an Archimedean spiral) in a $\mathbb{R}^2$ latent space. We train a 5-layer ReLU MLP with 256-dim hidden units as the generator following Li & He (2025) on the 2D latent space. The perceptual model is a self-supervised denoiser autoencoder (DAE) trained on perceptual space $\mathbb{R}^D$, where the 2D latent is lifted by a fixed column-orthogonal projection $P \in \mathbb{R}^{D\times d}$ ($P^\top P = I_d$). We compare three supervision spaces: (1) standard velocity regression in the latent space, (2) regression through the feature space of an untrained DAE, and (3) regression through the feature space of a trained DAE. As shown in Figure 1, both latent-space supervision and the untrained-DAE feature space exhibit mean-seeking behavior: with few sampling steps, their samples concentrate near the off-manifold average of the distribution, the toy analog of pixel-space blur. These samples gradually move back onto the manifold only when many integration steps are used. In contrast, supervision through the trained DAE feature space, which yields ($R_\phi > 1$), produces samples that lie close to the manifold with far fewer steps. The untrained DAE provides a nonlinear, non-Euclidean feature space, yet its $R_\phi \approx 1$ and its failure indicate that the key factor is the learned off-manifold penalty rather than feature nonlinearity alone; this is consistent with our RandViT ablation on SD3-Medium (Section 4.3). As shown in Figure 1, few-step quality improves as the off-manifold penalty R_ϕ increases: larger R_ϕ consistently reduces the off-manifold distance of generated samples and improves few-step generation performance. Table 2 further shows that increasing the feature dimension D of the perceptual DAE space raises R_ϕ and improves few-step generation performance.

Finally, we verify that pretrained perceptual spaces exhibit a stronger off-manifold penalty on natural images. For each anchor image a from FFHQ (Karras et al., 2019), we retrieve its k nearest neighbors in pixel space (k=8) and select one neighbor as another on-manifold sample b. Since real images do not provide an exact inter-mode midpoint, we construct a practical off-manifold proxy m by averaging the 100 images most similar to a. This averaged face is visibly blurry and off-manifold, resembling the blur commonly produced by coarse few-step prediction. We then measure R_ϕ in the SD3 VAE latent space and in several pretrained perceptual spaces (Table 1). The VAE latent space gives the weakest penalty ($R_\phi \approx 1.59$), treating the blurry average as only moderately farther from the anchor than another real image. Thus, under coarse one-step prediction, latent-space supervision provides limited pressure against blur, helping explain why few-step sampling struggles even though many-step integration can still recover high-quality samples. In contrast, pretrained perceptual features such as VGG, DINOv2, and ConvNeXt assign substantially larger penalties ($R_\phi \approx 1.9$–2.0). Surprisingly, the ranking of R_ϕ across supervision spaces closely matches their relative performance on the text-to-image benchmark (Table 7). This indicates R_ϕ plays a vital role in the few-step generation performance.

Furthermore, we examine why averaging in perceptual space is more compatible with few-step generation. We take $1{,}000$ similar face images from FFHQ and average them in three spaces: (1) directly in pixel space, (2) in SD3 VAE latent space, decoded with the SD3 VAE decoder, and (3) in DINOv2 feature space, decoded with a representation autoencoder (RAE) (Zheng et al., 2025). As shown in Figure 1(c), the DINOv2 average preserves sharp facial structure, whereas the pixel and latent averages are blurry. This suggests that averaging in a semantic feature space is far more benign than averaging in pixel or latent space. Because pretrained perceptual features organize images along semantically meaningful dimensions, their average is more likely to decode to a coherent on-manifold image rather than an off-manifold blur. As a result, even when the high-noise posterior is multimodal, perceptual supervision provides a more coherent one-step target, improving robustness under coarse few-step integration.

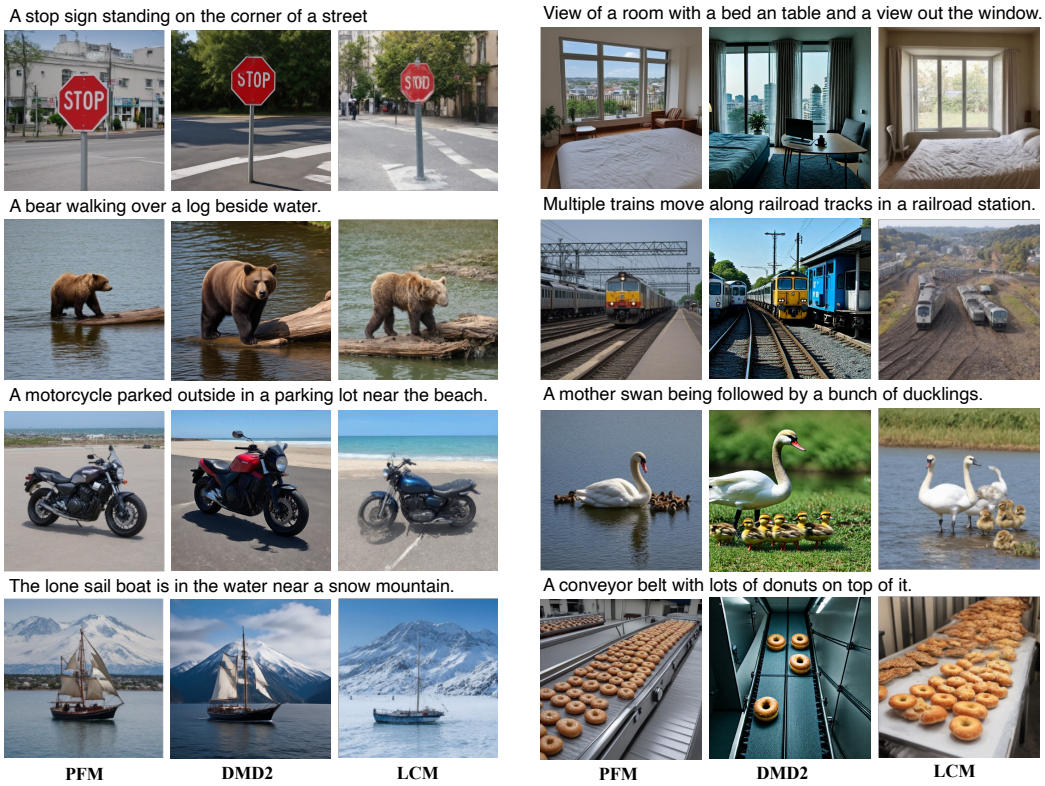
4 Experiments

4.1 Experimental Setting

We evaluate PFM across various tasks, including text-to-image generation, image editing, and text-to-video generation. For text-to-image generation, we apply PFM to SD3-Medium (Esser et al., 2024) and train on our in-house dataset. We evaluate on COCO 2014 val (Lin et al., 2014) using FID (Heusel et al., 2017), CLIP Score (Radford et al., 2021), and HPSv3 (Ma et al., 2025), which measure image fidelity, text-image alignment, and aesthetic quality, respectively. For image editing, we apply PFM to Qwen-Image-Edit (Wu et al., 2025), train on our in-house dataset, and evaluate on

Table 3: **Text-to-image generation results on COCO 2014 val set.** We manually implemented LCM and DMD2 on SD3-Medium based on their original codebase and trained on the same dataset.

Method	NFE	FID ↓	CLIP ↑	HPSv3 ↑
SD3-Medium	50×2	28.26	33.5384	10.90
LCM	4	**30.59**	30.62	6.68
DMD2	4	36.20	31.24	8.28
PFM (Ours)	4	31.16	**31.89**	**8.50**
LCM	8	34.18	30.42	9.19
DMD2	8	36.16	31.54	10.19
PFM (Ours)	8	**33.93**	**31.70**	**11.42**



(a) NFE=4 comparison

(b) NFE=8 comparison

Figure 2: Qualitative comparison on SD3-Medium at different sampling steps.

the MagicBrush (Zhang et al., 2023) benchmark. For text-to-video generation, we apply PFM to the Wan-2.1 1.3B model, and evaluate on the VBench (Huang et al., 2024) benchmark. Detailed training settings and hyper-parameters are provided in the Appendix.

4.2 Main Results

Text-to-image generation. For text-to-image generation, we compare our method with two distillation approaches: LCM (Luo et al., 2023) and DMD2 (Yin et al., 2024a). All methods are trained for the same 1,000 optimization steps with the same data on SD3-Medium. As shown in Table 3, our method achieves state-of-the-art results on CLIP score, HPSv3 score and FID in 8 steps, demonstrating that simple perceptual guidance can effectively improve both semantic alignment and image

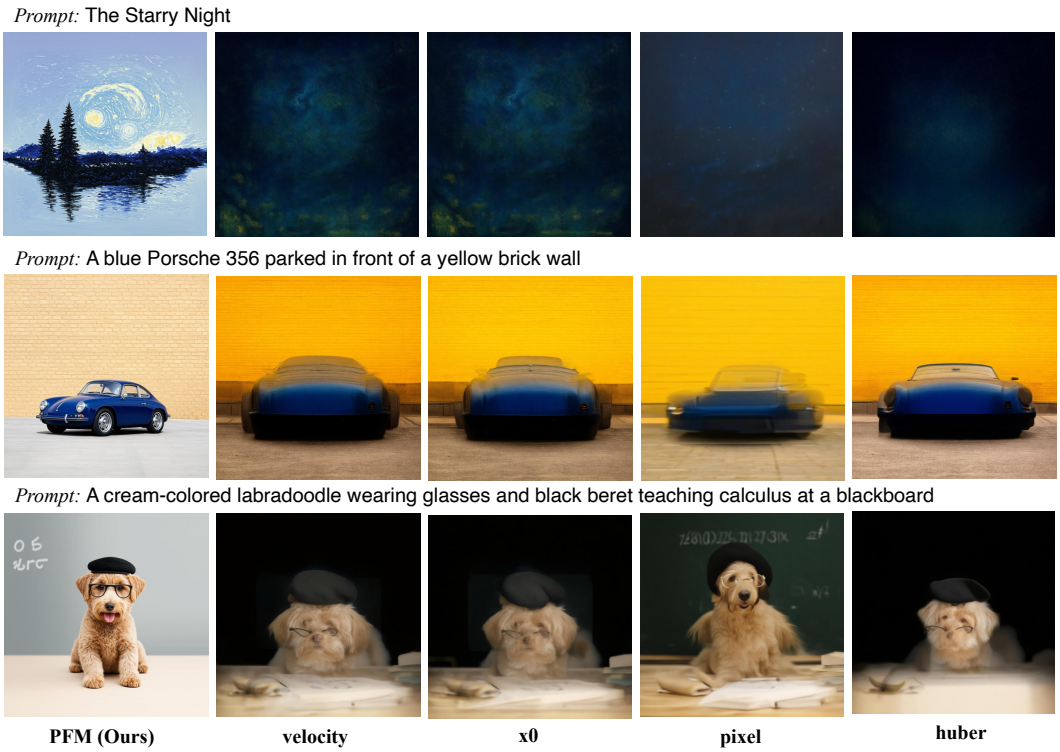


Figure 3: **Ablation of supervision spaces.** All models are fine-tuned from SD3-Medium with the same data and iterations, evaluated at 8 steps. PFM uses no CFG; others use CFG scale 4.0.

quality. We further provide visual comparisons in Figure 2. We can see that LCM produces blurrier images with more artifacts than our method, while DMD2 tends to generate images with overly high saturation and contrast.

Image editing. We apply PFM to Qwen-Image-Edit and evaluate its effectiveness on image editing tasks using the MagicBrush benchmark (Zhang et al., 2023). We use VGG (Simonyan & Zisserman, 2014) and DINO (Caron et al., 2021) as perceptual feature extractors and train the model using only the PFM objective, without CFG baking. We compare PFM against the original Qwen-Image-Edit model with 40-step Euler sampling, as well as its 8-step consistency sampling variant with CFG scale of 4.0.

As shown in Table 4, reducing sampling steps from 40 to 8 leads to noticeable degradation. In contrast, PFM significantly outperforms the 8-step baseline while requiring no CFG at inference. Remarkably, PFM even surpasses the 40-step model on CLIP-I, DINO, L1, and L2, while achieving a comparable CLIP-T score. Qualitative comparisons are provided in Appendix B.

Text-to-video generation. We apply PFM to Wan2.1-1.3B (Wan et al., 2025) and evaluate on VBench (Huang et al., 2024), training with InternVideo2 (Wang et al., 2024) as the perceptual model

Table 4: **Image editing results on MagicBrush.** PFM is applied to Qwen-Image-Edit. "$\times 2$" denotes the additional network evaluation introduced by classifier-free guidance (CFG).

Method	NFE	CFG	CLIP-T ↑	CLIP-I ↑	DINO ↑	L1 ↓	L2 ↓
Qwen-Image-Edit	40×2	4.0	**0.3129**	0.8957	0.8312	0.1060	0.0428
Qwen-Image-Edit	8×2	4.0	0.3073	0.8714	0.7968	0.1394	0.0526
PFM (ours)	8	1.0	0.3026	**0.9402**	**0.9187**	**0.0535**	**0.0156**

Table 5: **VBench evaluation results.** PFM is compared against the Wan2.1-1.3B baselines. Bold indicates the best result.

Method	NFE	Object Class	Multi. Objects	Human Action	Color	Spatial Rel.	Scene	Appear. Style	Temporal Style	Overall Cons.
Wan 2.1	35×2	0.786	**0.592**	0.810	**0.912**	**0.646**	0.283	0.222	**0.219**	0.219
Wan 2.1	8×2	0.763	0.383	0.800	0.679	0.351	0.260	**0.226**	0.186	0.222
PFM	8	**0.820**	0.562	**0.910**	0.789	0.620	**0.356**	0.198	0.214	**0.246**

Method	NFE	Subject Cons.	BG Cons.	Temp. Flicker	Motion Smooth.	Dynamic Degree	Aesth. Quality	Imaging Quality	Quality Score	Sem. Score	Total Score
Wan 2.1	35×2	0.962	0.960	0.979	0.987	0.379	0.535	**0.658**	0.799	0.675	0.774
Wan 2.1	8×2	0.918	0.964	0.994	0.976	**0.514**	0.550	0.465	0.775	0.578	0.735
PFM	8	**0.969**	**0.972**	**0.996**	**0.995**	0.319	**0.651**	0.618	**0.820**	**0.677**	**0.792**

for 4000 iterations without CFG baking. We find that using a video-based foundation model encourages greater motion dynamics compared to image-based models like DINOv2 (Oquab et al., 2023). As shown in Table 5, PFM significantly surpasses the 8-step counterpart and achieves comparable or slightly better results than the 35-step baseline at CFG scale 5.0. Qualitative comparisons are in Appendix B.

4.3 Ablation Studies

Supervision on other spaces. We compare PFM against alternative supervision strategies: velocity regression (standard FM), x_0 regression, pixel-space regression, and Huber loss, all in the VAE latent space. All models are trained from SD3-Medium and evaluated on PartiPrompts with 8-step sampling. As shown in Figure 3, only perceptual supervision produces sharp results; all alternatives yield noticeably blurrier outputs.

Perceptual models. We compare VGG, DINOv2, SigLIP (Zhai et al., 2023), ConvNeXt (Liu et al., 2022), a randomly initialized ViT (RandViT), and VGG+DINOv2. As shown in Figure 7 (Appendix), all pretrained models produce sharp results, whereas RandViT yields blurrier outputs, confirming that learned representations are essential. Higher-resolution models (DINOv2, SigLIP, ConvNeXt) preserve more fine-grained details than VGG (224×224). Quantitative evaluation in Table 7 (Appendix) shows that performance closely tracks the off-manifold penalty R_ϕ measured in Table 1.

Classifier-free guidance. PFM trained with only the perceptual objective already generates high-quality samples without CFG, suggesting that the guidance effect is implicitly learned. The model remains compatible with standard CFG at inference, where small scales suffice (Appendix A).

Inference steps. We evaluate PFM with prediction-side CFG baking on PartiPrompts using 1, 2, 4, and 8 steps. As shown in Figure 9 (Appendix), quality improves consistently; at 4–8 steps the results achieve strong visual quality with most fine details preserved.

Layer depth in the perceptual model. We ablate which DINOv2 layer the perceptual loss is applied to. As shown in Figure 8 (Appendix), shallow layers yield blurrier results, consistent with Section 3.4: shallow features remain nearly pixel-isometric, while deeper layers encode learned semantic structure. However, the deepest layers alone degrade pixel fidelity; we therefore average features across multiple layers.

5 Conclusion

We introduced Perceptual Flow Matching (PFM), which replaces latent-space velocity regression with perceptual supervision on decoded predictions, enabling high-quality 4–8 step generation across image generation, editing, and video without teacher models or distillation. Our ablations show that supervision geometry is the key factor: pretrained perceptual features substantially outperform latent- or pixel-space objectives. Extending PFM to fewer than two steps and identifying the optimal perceptual space remain open for future work.

Reproducibility Statement

We provide detailed training settings, hyperparameters, and evaluation protocols in the Appendix to support reproducibility. All experiments are built upon publicly available models (SD3-Medium, Wan2.1-1.3B, Qwen-Image-Edit) and standard evaluation benchmarks (COCO 2014, MagicBrush, VBench). The core PFM method requires only a change in the supervision space of standard flow matching training, with pseudocode provided in Algorithm 1.

LLM Usage Disclosure

Large language models were used to assist with improving the writing quality of this paper, including grammar correction and sentence rephrasing.

References

Lichen Bai, Tianhao Zhang, Shitong Shao, Dingwei Tan, Qiyu Zhong, Zhengpeng Xie, Haopeng Li, Qinghao Huang, Dandan Shen, Tengjiao Ji, et al. Mainecoon: Pursuing a real-time audio-visual social world model. *arXiv preprint arXiv:2606.17800*, 2026.

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*, pp. 9650–9660, 2021.

Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. *arXiv preprint arXiv:2501.15420*, 2025.

Quan Dao, Hao Phung, Trung Dao, Dimitris N. Metaxas, and Anh Tran. Self-corrected flow distillation for consistent one-step and few-step image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

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

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

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*, pp. 740–755. Springer, 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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations (ICLR)*, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. *arXiv preprint arXiv:2506.14603*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision–ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXV*, pp. 396–416. Springer, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A Classifier-Free Guidance Ablation

Table 6: Effect of classifier-free guidance (CFG) baking and train/val CFG scales. We train with VGG+DINOv2 as perceptual models for 500 steps. 1.0 indicates CFG is not used during inference.

CFG Baking	Train Scale	Val Scale	CLIP	HPSv3
–	–	1.0	30.38	7.41
–	–	1.5	31.85	9.27
–	–	2.0	**32.16**	<u>9.70</u>
Pred side	1.5	1.0	31.17	9.17
Pred side	2.0	1.0	31.57	**9.78**
Pred side	2.5	1.0	<u>31.96</u>	9.40
Target side	1.5	1.0	30.13	7.34
Target side	2.0	1.0	30.44	8.09
Target side	2.5	1.0	30.70	8.62

As shown in Figure 4, PFM trained with only the perceptual objective already generates high-quality samples without classifier-free guidance. As shown in Table 6, the model remains compatible with

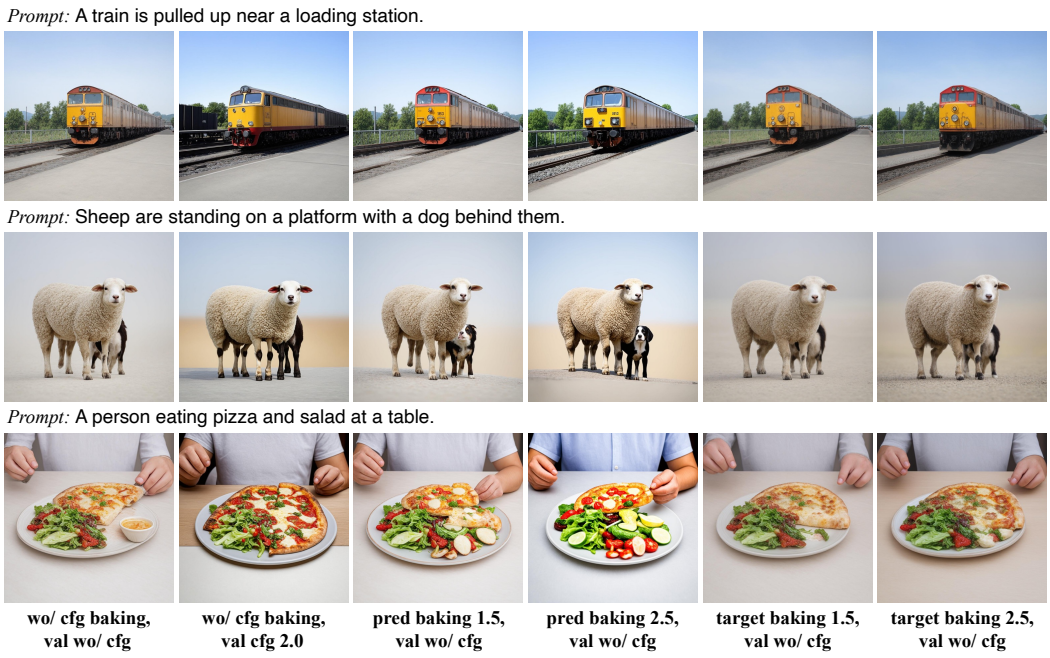


Figure 4: **Ablation of classifier-free guidance.** PFM without CFG baking generates images with good fidelity, and prediction-side CFG baking yields results similar to inference-time CFG.

standard CFG at inference time, where relatively small guidance scales suffice to achieve strong results. When applying the proposed CFG baking strategies (Section 3.3), the model trained with CFG baking yields results closely matching inference-time CFG while requiring only a single forward pass during sampling.

B Additional Results

Table 7: Comparison of vision backbones as feature extractors.

Backbone	CLIP	HPSv3
VGG	29.5011	5.0642
DINO	29.7590	7.4574
ConvNeXt	**31.5326**	**10.1701**
SigLIP	29.2188	5.2237
RandViT	20.9400	-9.4900



Figure 5: **Qualitative results of image editing.** All images are generated with the same prompts. The Qwen-Image-Edit baselines use a CFG scale of 4.0, whereas PFM does not use CFG at inference. PFM closely matches the 40-step baseline with only 8 sampling steps.

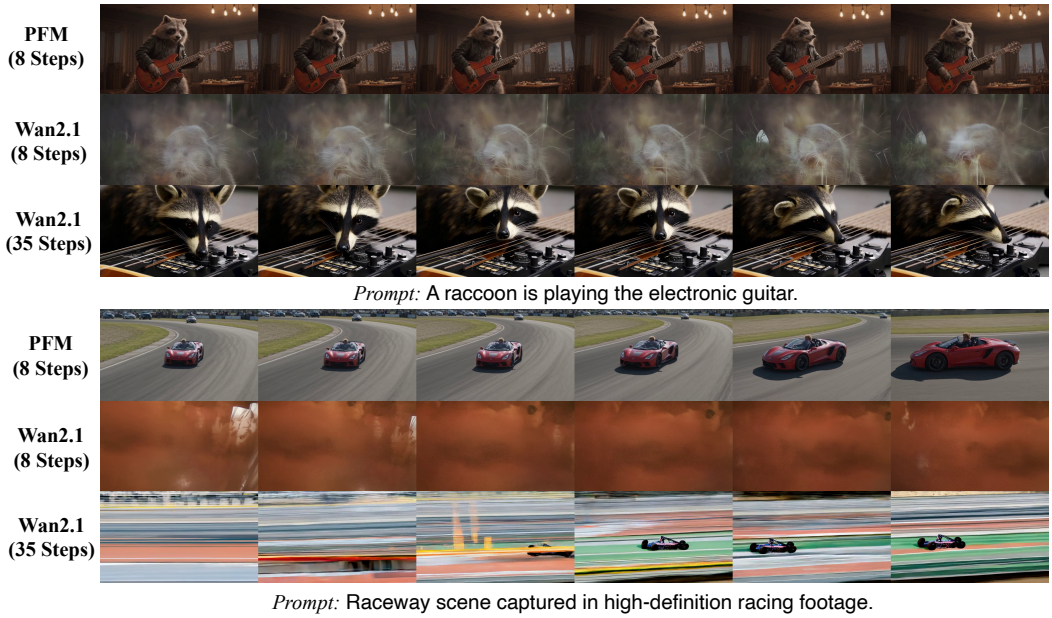


Figure 6: **Qualitative results of text-to-video generation.** The 8-step baseline exhibits significant blurriness, whereas PFM produces high-fidelity results matching the 35-step baseline.

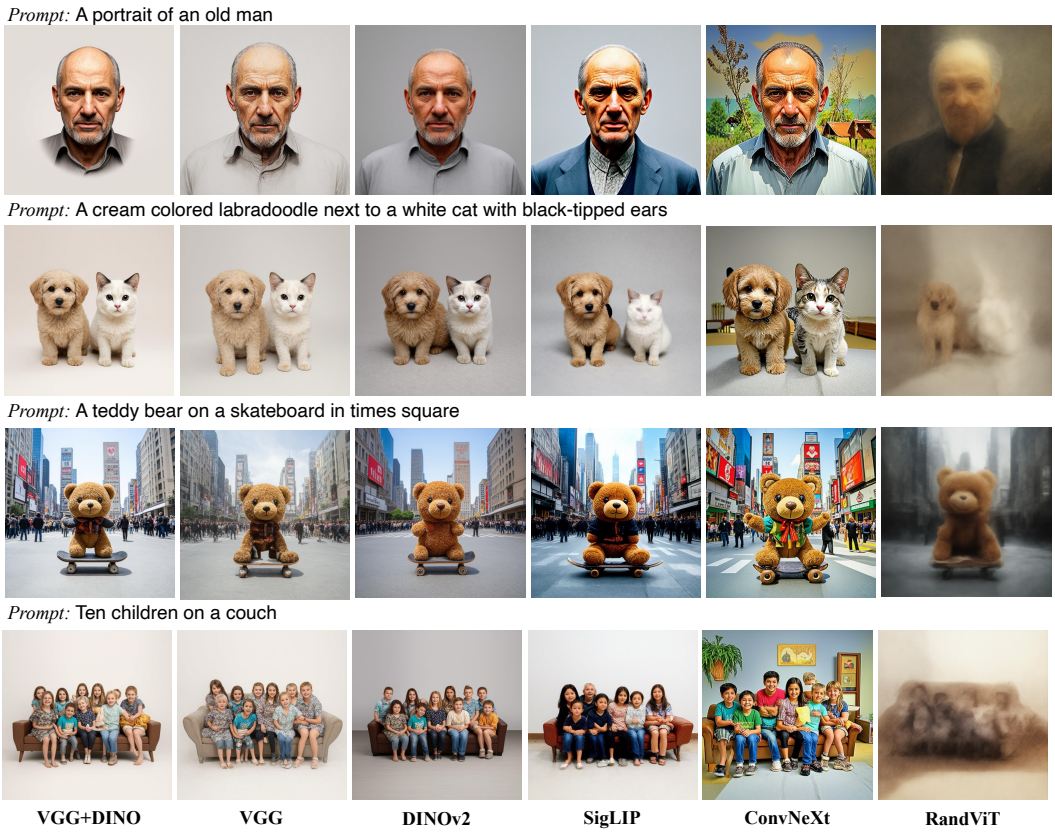


Figure 7: **Ablation of different perceptual models.** All perceptual models yield clear results in 8 steps, while a randomly initialized ViT (RandViT) produces blurry results, similar to standard flow matching.

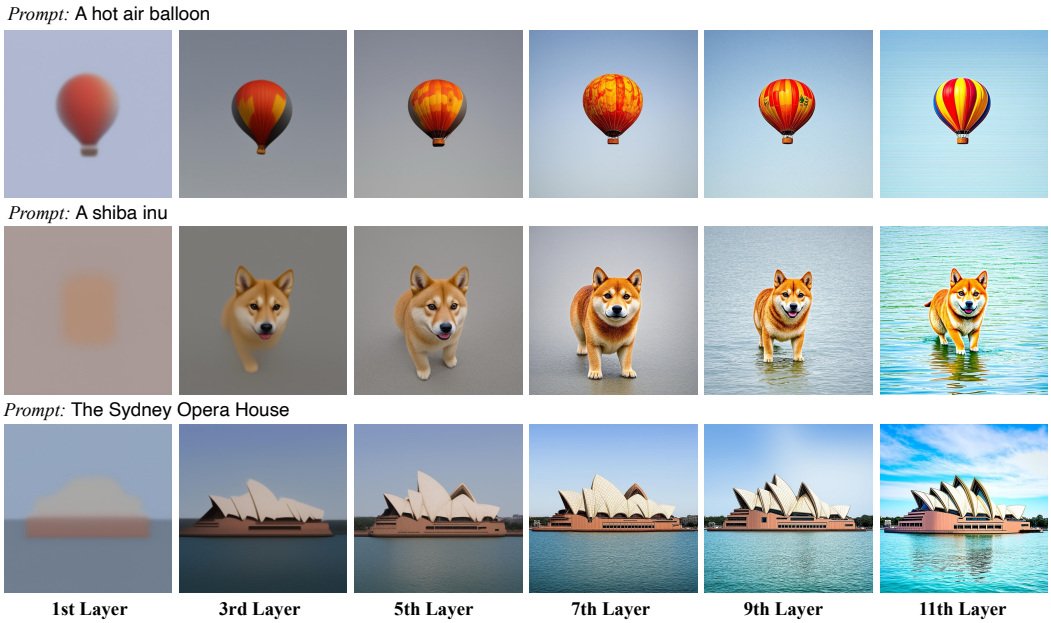


Figure 8: **Ablation on layer depth in the perceptual model.** We compute the perceptual loss using features from different layers of DINOv2. Shallow layers, which remain closer to pixel-level representations, produce blurrier results, while the deepest layers tend to collapse to fixed modes.

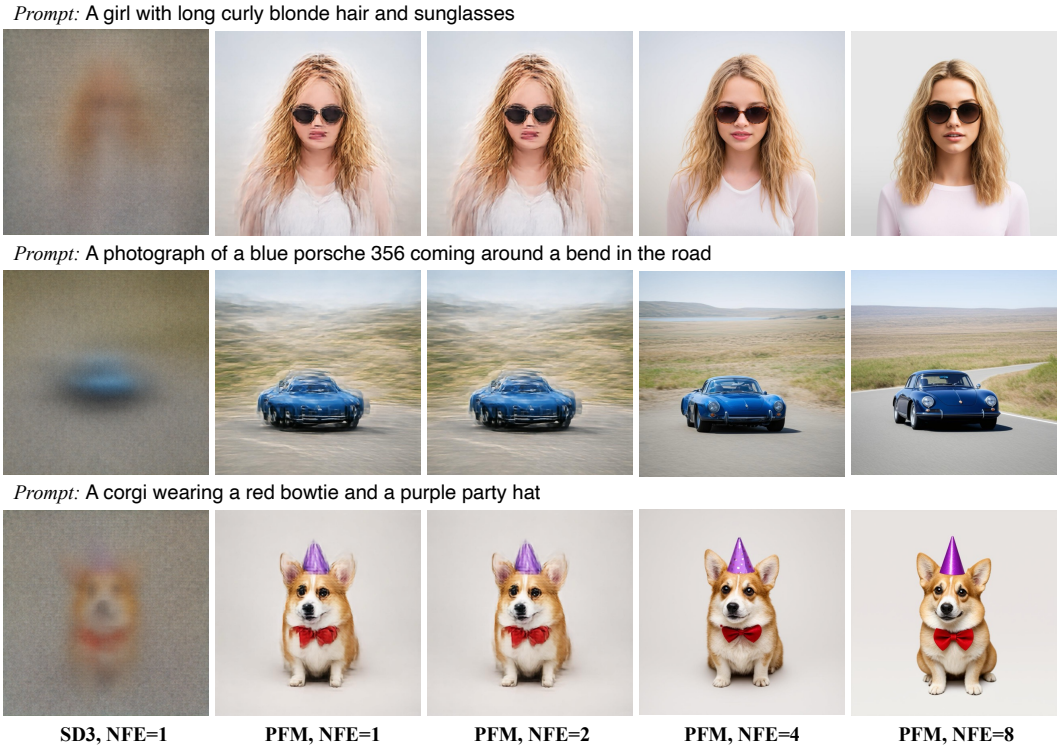


Figure 9: **Ablation of inference steps.** PFM produces high-fidelity results at 4 and 8 NFE and remains sharp even at 1 NFE, whereas the SD3 baseline is blurry at 1 NFE.