INTRA-ENCODER MULTI-VIEW FRÉCHET SUPERVISION FOR VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fréchet Distance (FD) has recently been adopted as a training objective for visual generative models, providing distribution-level supervision by matching the feature statistics of real and generated images in frozen visual representation spaces. Optimizing FD in a single representation space can overfit to the statistics captured by that representation, producing a low in-space FD while substantial discrepancies under other representations and poor visual quality remain. FD-loss addresses this issue with FD-SIM, which jointly aligns real and generated distributions across multiple pretrained visual encoders. This multi-encoder supervision broadens representation coverage but introduces substantial additional encoder computation during training.

To obtain multi-view Fréchet supervision without relying on multiple independent encoders, we propose Intra-Encoder Multi-View Fréchet Supervision (IEMV-FD), which constructs multiple distributional supervision views from the internal representations of a single pretrained visual encoder. Specifically, we extract representations from multiple layers and form both CLS-token and average-pooled patch-token views at each selected layer. Existing Fréchet-based evaluation metrics may use encoders that also participate in training supervision as evaluation judges, creating supervision-evaluation overlap and making it difficult to isolate generalization to unseen representation spaces. We therefore introduce held-out FDr^6, which averages normalized Fréchet distance ratios across 6 evaluation encoders that are not involved in the training of either IEMV-FD or FD-SIM. On ImageNet 256×256 class-conditional generation, IEMV-FD achieves lower held-out FDr^6 than FD-SIM across multiple models, including reductions from 10.76 to 7.37 on JiT-B and from 8.01 to 6.65 on pMF-B, while using a single supervision encoder with only about 11.5% of the supervision-encoder parameters used by FD-SIM. Our code and model weights will be made publicly available.

1 INTRODUCTION

Fréchet Distance (FD) has long been one of the most widely used distribution-level metrics for evaluating visual generative models (Heusel et al., 2017). Directly using FD for generative training, however, is challenging because reliable estimation of feature means and covariances typically requires a population substantially larger than a standard training batch. Recently, Representation Fréchet Loss (FD-loss) (Yang et al., 2026) addressed this practical bottleneck by decoupling the population size used for FD estimation from the batch size used for gradient computation, making FD directly and stably optimizable in frozen visual representation spaces. This development establishes FD as a practical training objective that directly constrains the feature distributions of real and generated images in frozen representation spaces.

The representation space used by FD-loss has a strong effect on the resulting optimization. Results from FD-loss show that optimizing FD in one representation can greatly reduce the distance measured in that space, while substantial discrepancies may remain under other representations. This representation dependence is also reflected in visual quality. In the representation ablation of FD-loss, the Inception-optimized model achieves the lowest FID, while models optimized in modern representation spaces produce better samples despite higher Inception FID (Yang et al., 2026). A single representation therefore provides only a partial view of the discrepancy between real and gen-

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.

erated distributions. Optimizing that view alone can emphasize the statistics captured by one feature space while leaving discrepancies visible to other representation spaces.

FD-loss addresses this limitation with FD-SIM, which jointly optimizes Fréchet objectives from SigLIP, Inception, and MAE (Yang et al., 2026). Combining multiple representations broadens the supervision signal, but each representation is produced by an independent visual encoder. The three FD-SIM supervision encoders contain approximately 755M parameters in total and require separate encoder forward passes during training. These costs motivate a supervision design that preserves representation diversity without requiring multiple independent visual encoders.

To this end, we propose Intra-Encoder Multi-View Fréchet Supervision (IEMV-FD), which obtains multiple Fréchet supervision views from the internal representations of a single frozen visual encoder. Our key observation is that a pretrained encoder already contains multiple diverse internal representations. Features extracted at different network depths capture different information, while different token aggregation strategies provide another source of variation. We combine these two dimensions to construct multiple layer–pooling representation heads. In our main configuration, CLS tokens and average-pooled patch features are extracted from multiple selected layers of a single frozen visual encoder. Each head defines a distinct representation space with its own Fréchet loss, and these objectives are optimized jointly. All supervision views share a single encoder forward pass. Our main configuration therefore provides multi-view supervision using one visual encoder with only about 11.5% of the supervision-encoder parameters used by FD-SIM.

Reliable evaluation of Fréchet-based generative training should determine whether improvements extend beyond the representation spaces used for supervision. The FDr^6 metric introduced by FD-loss averages normalized Fréchet distance ratios across six visual encoders (Yang et al., 2026). However, three of these encoders, namely Inception, MAE, and SigLIP, are also used for training supervision in FD-SIM. Consequently, FDr^6 combines performance in both supervised representation spaces and those not used for supervision. Its aggregate score alone cannot isolate improvements in representation spaces that are not directly optimized during training.

Therefore, we introduce held-out FDr^6, which uses six pretrained visual encoders for evaluation. These include DINOv3-L, PE-L, SwinV2-L, and EVA02-L (Siméoni et al., 2025; Bolya et al., 2025; Liu et al., 2022; Fang et al., 2023), together with ConvNeXtV2-B (Woo et al., 2023) and CLIP ViT-L/14 (Radford et al., 2021) from the original FDr^6 evaluation set. None of these encoders is used for training supervision in either IEMV-FD or FD-SIM. Held-out FDr^6 averages their normalized Fréchet distance ratios to evaluate both methods in representation spaces that are not directly optimized during training.

Our contributions are summarized as follows:

- We propose IEMV-FD, which achieves multi-view Fréchet supervision using only a single frozen visual encoder. By exploiting diverse internal representations, our method mitigates representation-specific FD overfitting while avoiding the cumbersome multi-encoder design of FD-SIM, thereby simplifying the supervision framework for FD-based image generation.
- We introduce held-out FDr^6, which evaluates generative models across six representation spaces whose encoders are not involved in any stage of training for either IEMV-FD or FD-SIM. This provides a common evaluation for examining performance transfer beyond directly supervised representations.
- We demonstrate that IEMV-FD achieves lower held-out FDr^6 than FD-SIM across multiple generator configurations, indicating that its improvements extend beyond the supervised representation spaces. These results are achieved using a single supervision encoder with approximately 11.5% of the supervision-encoder parameters used by FD-SIM.

2 Related Work

Generative Models. Diffusion models generate images through iterative denoising (Ho et al., 2020), while flow matching learns continuous transformations between noise and data distributions (Lipman et al., 2023). Recent advances in efficient image generation include pixel-space denoising architectures such as JiT (Li & He, 2026) and one-step generation methods such as

pixel MeanFlow (Lu et al., 2026). IEMV-FD applies multi-view Fréchet post-training to a multi-step JiT generator, enabling one-step generation while improving generation quality.

Distribution Matching for Generative Training. Distribution matching trains generative models by minimizing discrepancies between real and generated feature distributions. MMD-GAN (Li et al., 2017) combines maximum mean discrepancy with adversarially learned representations, while dos Santos et al. (2019) match the means and covariances of pretrained perceptual features. Subsequent studies investigate the direct optimization of Fréchet Distance. Doan et al. (2020) minimize Fréchet Distance in discriminator feature spaces, and Mathiasen & Hvilshøj (2020) study backpropagation through FID. More recently, Representation Fréchet Loss (Yang et al., 2026) makes Fréchet-based post-training practical by separating the population used for estimating distribution statistics from the minibatch used for gradient computation. Building on this framework, IEMV-FD simplifies multi-view Fréchet supervision using a single frozen encoder and introduces a held-out evaluation metric to assess performance beyond directly supervised representation spaces.

Pretrained Visual Representations for Generative Supervision. Pretrained visual representations provide additional supervision for generative training. Projected GAN (Sauer et al., 2021) uses multi-scale features from a frozen pretrained network for adversarial discrimination, while Vision-aided GAN (Kumari et al., 2022) combines multiple pretrained visual encoders to diversify adversarial feedback. REPA (Yu et al., 2025) instead aligns intermediate generator representations with clean-image features extracted by pretrained visual encoders. These approaches incorporate pretrained visual knowledge through adversarial discrimination or representation alignment. FD-SIM (Yang et al., 2026) further explores multi-representation Fréchet supervision using multiple frozen visual encoders. In contrast, IEMV-FD constructs diverse Fréchet supervision views from internal representations of a single frozen encoder. This design avoids additional independent supervision backbones while improving performance across held-out representation spaces.

Generative Model Evaluation. Fréchet Inception Distance (FID) is widely used to evaluate image generation by comparing the feature distributions of real and generated images in the Inception-v3 representation space (Heusel et al., 2017). However, its dependence on Inception features can affect evaluation outcomes. Kynkäänniemi et al. (2023) demonstrate that FID is sensitive to ImageNet class statistics, while Jayasumana et al. (2024) propose CMMD, which measures distributional discrepancies using CLIP representations without assuming Gaussian feature distributions. To reduce dependence on a single representation, Yang et al. (2026) introduce FDr^6, which averages normalized Fréchet distance ratios across six visual encoders. However, three of these encoders also participate in FD-SIM training supervision, introducing direct overlap between supervision and evaluation. To examine whether improvements extend beyond directly supervised representation spaces, we introduce held-out FDr^6, which evaluates both IEMV-FD and FD-SIM using 6 visual encoders that are not used for training supervision by either method.

3 Preliminaries: Representation-Space Fréchet Loss

Let G_θ denote a generator parameterized by θ, and let h be a frozen image representation function. We denote the real image distribution by P_r and the distribution induced by the generator by P_θ. For representation h, let (μ_h^r, Σ_h^r) and (μ_h^g, Σ_h^g) denote the feature means and covariances of real and generated images, respectively. The Fréchet loss in representation space h is defined as

$$\mathcal{L}_h = \|\mu_h^g - \mu_h^r\|_2^2 + \mathrm{Tr}\left(\Sigma_h^g + \Sigma_h^r - 2\left[(\Sigma_h^r)^{1/2}\Sigma_h^g(\Sigma_h^r)^{1/2}\right]^{1/2}\right). \tag{1}$$

Reliable estimation of feature means and covariances typically requires a population larger than a standard training batch. Following Representation Fréchet Loss (Yang et al., 2026), we decouple the population used for estimating distribution statistics from the batch used for gradient computation. The real-image reference statistics are computed offline, while the generated-image statistics are accumulated across training iterations. For a generated batch $\{x_{g,i}\}_{i=1}^B$, the batch mean and raw second moment in representation space h are

$$\mu_{h,b} = \frac{1}{B}\sum_{i=1}^{B} h(x_{g,i}), \quad M_{h,b} = \frac{1}{B}\sum_{i=1}^{B} h(x_{g,i})h(x_{g,i})^\top. \tag{2}$$

The generated statistics are updated using an exponential moving average (EMA) with decay coefficient β:

$$\mu_h^g \leftarrow \beta \, \mathrm{sg}(\mu_h^g) + (1-\beta)\mu_{h,b}, \tag{3}$$

$$M_h^g \leftarrow \beta \, \mathrm{sg}(M_h^g) + (1-\beta)M_{h,b}, \quad \Sigma_h^g = M_h^g - \mu_h^g(\mu_h^g)^\top. \tag{4}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. Historical EMA statistics contribute to the current distribution estimate without propagating gradients across iterations. Gradients are propagated only through the generated features of the current batch.

4 METHOD

4.1 OVERVIEW

We propose Intra-Encoder Multi-View Fréchet Supervision (IEMV-FD), which constructs multiple Fréchet supervision views from the internal representations of a single frozen visual encoder. Figure 1 illustrates the difference between conventional multi-encoder supervision and our method. While FD-SIM obtains distinct representations from three independent visual encoders, IEMV-FD extracts multiple representations from different internal layers of one encoder. At each selected layer, we construct two image-level representations using the CLS token and average-pooled patch tokens. Each representation defines a separated Fréchet objective, and all objectives are jointly optimized while sharing a single encoder forward pass.

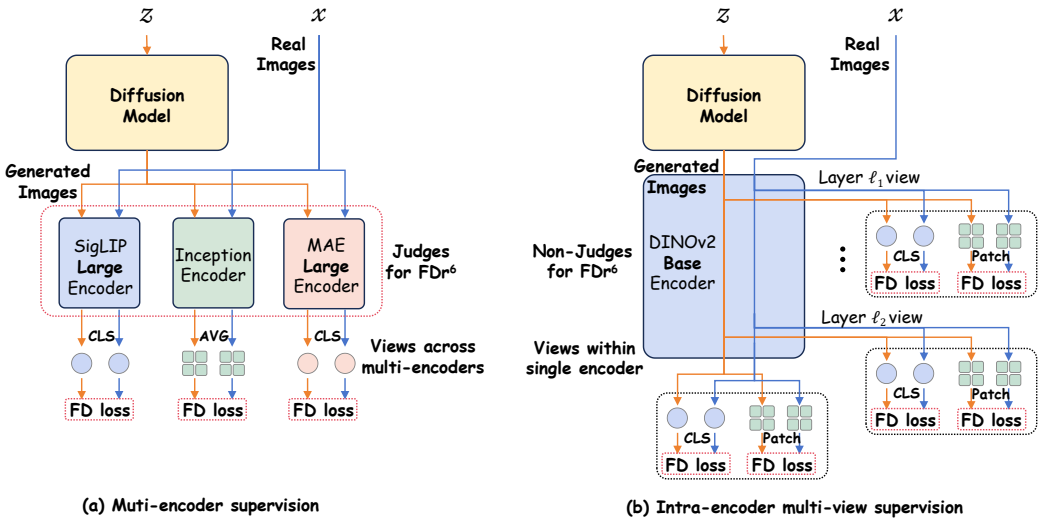


Figure 1: Multi-encoder supervision versus intra-encoder multi-view Fréchet supervision. (a) FD-SIM employs three independent frozen visual encoders, namely SigLIP2-SO400M, Inception-v3, and MAE-Large, to construct separate Fréchet objectives. These encoders contain approximately 755M parameters in total and require independent forward passes, introducing substantial computational overhead during training. (b) IEMV-FD extracts CLS-token and average-pooled patch-token representations from multiple selected layers of a single frozen DINOv2-Base encoder. Each representation view defines a separate Fréchet objective, while all views share a single encoder forward pass. Notably, IEMV-FD uses only approximately 11.5% of the supervision-encoder parameters required by FD-SIM.

4.2 INTRA-ENCODER MULTI-VIEW REPRESENTATIONS

As illustrated in Figure 1(a), multi-encoder Fréchet supervision obtains different representation spaces from separate pretrained visual encoders. In FD-SIM, SigLIP, Inception, and MAE independently extract image features and provide three Fréchet objectives. This design obtains representation diversity across encoders but requires separate encoder forward passes, resulting in inefficient training.

In contrast, Figure 1(b) illustrates how IEMV-FD constructs multiple supervision views within a single frozen visual encoder. Instead of introducing additional visual backbones, we extract internal representations from multiple selected layers and construct different image-level features from the corresponding token sequences. Consider a vision Transformer encoder ϕ containing L Transformer blocks. Given an image x, the output of block ℓ is represented as

$$\Phi_\ell(x) = [c_\ell(x), p_{\ell,1}(x), \ldots, p_{\ell,N}(x)]\,. \tag{5}$$

where $c_\ell(x)$ denotes the CLS token, $p_{\ell,n}(x)$ denotes the n-th patch token, and N is the number of patch tokens. We obtain multiple supervision views by extracting representations from different network depths and applying two token representation methods at each selected layer. Specifically, we define

$$h_{\ell,\mathrm{cls}}(x) = c_\ell(x), \quad h_{\ell,\mathrm{avg}}(x) = \frac{1}{N}\sum_{n=1}^{N} p_{\ell,n}(x). \tag{6}$$

The CLS representation uses the image-level token produced by the encoder, while the average-pooled representation aggregates the patch tokens into an image-level feature. Applying these two operations to multiple internal layers produces a collection of representations that differ in both their extraction depth and aggregation strategy. Let $\mathcal{I} \subseteq \{0, \ldots, L-1\}$ denote the selected set of encoder layers. The resulting supervision-view set is

$$\mathcal{S} = \mathcal{I} \times \{\mathrm{cls}, \mathrm{avg}\}. \tag{7}$$

Each pair $s = (\ell, a) \in \mathcal{S}$ defines a representation function $h_s = h_{\ell,a}$, where s denotes a supervision view.

Importantly, all selected representations are extracted from intermediate activations produced during a single encoder forward pass. Although the method constructs multiple supervision views, it does not require additional independent visual encoders.

4.3 View-Wise Fréchet Supervision

After constructing the intra-encoder representation views, we assign a Fréchet objective to each selected view, as illustrated on the right side of Figure 1. For every $s \in \mathcal{S}$, we maintain view-specific real-image reference statistics $(\mu^r_{h_s}, \Sigma^r_{h_s})$ and generated-image EMA statistics $(\mu^g_{h_s}, \Sigma^g_{h_s})$. The corresponding Fréchet loss $\mathcal{L}_{h_s}$ is computed using Eq. (1). We compute the Fréchet loss independently for each selected view. The final objective of our IEMV-FD is the uniform average over all selected views,

$$\mathcal{L}_{\text{IEMV-FD}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \mathcal{L}_{h_s}. \tag{8}$$

All selected views receive the same scalar weight, and no additional learnable weighting parameters are introduced. This formulation allows different internal representations to contribute independent distribution-matching objectives while sharing the same visual encoder. Unlike conventional multi-encoder supervision, IEMV-FD obtains multiple Fréchet objectives without requiring a separate backbone for each representation space.

5 Experiments

5.1 Experimental Setup and Evaluation Protocol

Experimental Settings. We evaluate IEMV-FD on ImageNet-1k (Deng et al., 2009) class-conditional generation at 256×256 resolution, covering the latent-space iMF-B model (Geng et al., 2026) and four pixel-space models: pMF-B and pMF-L (Lu et al., 2026), and JiT-B and JiT-L (Li & He, 2026). Unless otherwise specified, IEMV-FD uses a single frozen DINOv2-Base encoder (Oquab et al., 2024). For our main experiments, we select zero-indexed layers $\{4, 9, 10, 11\}$ and extract both CLS-token and average-pooled patch-token representations, yielding eight supervision views. Our models are post-trained for 50 epochs, whereas the main FD-SIM checkpoints use three independent supervision encoders and 100 post-training epochs. Additional post-training configuration details are given in Appendix A.

Evaluation Metrics. Following FD-loss (Yang et al., 2026), we evaluate generative performance using Fréchet distance ratios (FDr) computed in pretrained visual representation spaces. However, three of the six encoders used in the original FDr^6 also participate in FD-SIM training supervision. This overlap makes it difficult to isolate improvements in representation spaces not directly optimized during training. To address this issue, we introduce **held-out** FDr^6 as our primary evaluation metric. It follows the normalization procedure of Yang et al. (2026) but employs six evaluation encoders that are not used for training supervision by either IEMV-FD or FD-SIM. We define

$$\mathrm{FDr}^6_{\text{held-out}}(\mathcal{G}) = \frac{1}{6} \sum_{\phi_i \in \mathcal{H}} \mathrm{FDr}_{\phi_i}(\mathcal{G}),$$

where $\mathcal{H}$ contains DINOv3-L, PE-L, SwinV2-L, EVA02-L, ConvNeXtV2-B, and CLIP ViT-L/14. Detailed encoder configurations and the normalization procedure are provided in Appendix B.

Evaluation Protocol. For the Fréchet-based evaluations, we report both the original FDr^6 and held-out FDr^6 using 50,000 class-balanced samples per checkpoint. We use fixed random seeds and matched sampling configurations when comparing methods within the same generator architecture. Detailed sampling configurations are provided in Appendix B.

5.2 Main Results

Quantitative results. Table 1 presents the main quantitative results on ImageNet 256×256. We re-evaluate the released FD-SIM checkpoints under a unified evaluation protocol and compare them with IEMV-FD across five generator configurations. IEMV-FD achieves lower held-out FDr^6 than FD-SIM on all five models. Specifically, held-out FDr^6 decreases from 12.61 to 8.88 on iMF-B, from 8.01 to 6.65 on pMF-B, from 5.28 to 4.00 on pMF-L, from 10.76 to 7.37 on JiT-B, and from 7.83 to 4.88 on JiT-L. These results demonstrate the effectiveness of IEMV-FD across both latent-space and pixel-space generators. Notably, IEMV-FD uses only approximately 11.5% of the supervision-encoder parameters required by FD-SIM.

Cross-Representation Transfer. To further investigate whether the improvements extend beyond the supervised representation spaces, we compare the performance of IEMV-FD and FD-SIM across six held-out visual encoders. As shown in Table 2, IEMV-FD achieves lower FDr than FD-SIM in five of the six held-out representation spaces across all five generator configurations, with particularly notable improvements under DINOv3-L and EVA02-L. Although FD-SIM achieves slightly lower FDr under ConvNeXt, IEMV-FD consistently obtains lower aggregate held-out FDr^6 across all evaluated generators. These results suggest that intra-encoder multi-view supervision provides benefits that generalize to multiple unseen evaluation representations rather than being concentrated in the supervised feature spaces.

Qualitative Comparison. We complement the quantitative results with selected class-conditional image comparisons between IEMV-FD and FD-SIM. Figure 2 presents paired samples generated by JiT-L and pMF-L under identical generation settings within each model, including the same ImageNet class conditions and initial noise. In several examples, IEMV-FD produces more recognizable subjects and clearer class-related structures.

5.3 Training Efficiency

To evaluate the training efficiency of IEMV-FD, we compare it with FD-SIM on JiT-B/16 and pMF-B/16 using a single NVIDIA H200 GPU. Both methods are benchmarked under matched settings within each model, with a batch size of 256 and BF16 precision. We use 10 warm-up steps followed by 30 measured steps, excluding initialization and initial compilation. As shown in Table 3, we report the computational cost of the frozen supervision encoders, peak GPU memory consumption of the entire training pipeline, median training step time, and image throughput. The memory measurements cover the complete training process, including the generator and supervision encoders, rather than the visual encoders alone.

IEMV-FD reduces supervision-encoder computation from 301.87 to 66.65 GFLOPs per image, a reduction of 77.9%. On JiT-B/16, peak GPU memory decreases from 108.04 to 54.57 GiB, while median step time decreases from 0.691 to 0.321 seconds, corresponding to reductions of 49.5%

Table 1: Main results on ImageNet 256×256. FD-SIM (reported) results are from Yang et al. (2026), while all other results are evaluated using our fixed sampling protocol. NFE denotes the number of network evaluations, and the encoder columns describe post-training supervision. Bold metric values indicate the better results between FD-SIM (re-eval.) and Ours. Lower metric values are better. † follows the JiT baseline sampler notation in Li & He (2026).

Model	Method	NFE	# Visual Encoders ↓	Visual Encoder Params	Post-training Epochs	FDr^6 ↓	Held-out FDr^6 ↓
Latent-space models							
iMF-B	Baseline	1	–	–	–	15.42	25.17
	+ FD-SIM (reported)	1	3	755.0M	100	5.56	–
	+ FD-SIM (re-eval.)	1	3	755.0M	100	5.61	12.61
	+ Ours	1	**1**	**86.6M**	**50**	**5.03**	**8.88**
Pixel-space models							
pMF-B	Baseline	1	–	–	–	13.83	15.13
	+ FD-SIM (reported)	1	3	755.0M	100	3.50	–
	+ FD-SIM (re-eval.)	1	3	755.0M	100	3.55	8.01
	+ Ours	1	**1**	**86.6M**	**50**	**3.53**	**6.65**
pMF-L	Baseline	1	–	–	–	9.17	11.06
	+ FD-SIM (reported)	1	3	755.0M	100	2.09	–
	+ FD-SIM (re-eval.)	1	3	755.0M	100	2.12	5.28
	+ Ours	1	**1**	**86.6M**	**50**	**2.11**	**4.00**
JiT-B	Baseline	$50\times2\times2^{\dagger}$	–	–	–	15.81	18.57
	+ FD-SIM (reported)	1	3	755.0M	100	5.53	–
	+ FD-SIM (re-eval.)	1	3	755.0M	100	5.66	10.76
	+ Ours	1	**1**	**86.6M**	**50**	**5.24**	**7.37**
JiT-L	Baseline	$50\times2\times2^{\dagger}$	–	–	–	10.80	13.02
	+ FD-SIM (reported)	1	3	755.0M	100	3.24	–
	+ FD-SIM (re-eval.)	1	3	755.0M	100	3.31	7.83
	+ Ours	1	**1**	**86.6M**	**50**	**3.01**	**4.88**

Table 2: Encoder-wise held-out Fréchet distance ratios on ImageNet 256×256. Parentheses give reductions from the corresponding baseline. Lower is better; bold marks the best result within each model.

Model	Method	Encoder-wise FDr ↓						Held-out FDr^6 ↓
		DINOv3-L	PE-L	SwinV2-L	EVA02-L	ConvNeXt	CLIP-L	
iMF-B	Baseline	70.48	31.80	5.04	12.39	3.27	28.06	25.17
	+ FD-SIM	32.02	21.76	2.32	5.65	**1.14**	12.77	12.61 (↓49.90%)
	+ Ours	**18.87**	**17.52**	**1.90**	**3.52**	1.16	**10.32**	**8.88** (↓64.70%)
pMF-B	Baseline	24.30	27.50	5.36	8.48	1.89	23.26	15.13
	+ FD-SIM	12.57	17.58	1.80	6.64	**0.76**	8.71	8.01 (↓47.10%)
	+ Ours	**9.52**	**16.43**	**1.54**	**3.62**	1.03	**7.76**	**6.65** (↓56.10%)
pMF-L	Baseline	18.36	23.07	3.61	5.30	1.34	14.69	11.06
	+ FD-SIM	5.74	12.96	1.28	5.29	**0.56**	5.85	5.28 (↓52.30%)
	+ Ours	**3.76**	**11.45**	**0.99**	**2.56**	0.71	**4.52**	**4.00** (↓63.80%)
JiT-B	Baseline	26.66	28.80	6.91	10.29	4.48	34.28	18.57
	+ FD-SIM	12.10	19.78	2.80	9.21	**1.32**	19.32	10.76 (↓42.10%)
	+ Ours	**8.93**	**17.41**	**2.02**	**3.68**	1.57	**10.61**	**7.37** (↓60.30%)
JiT-L	Baseline	14.38	22.59	5.36	6.49	3.46	25.82	13.02
	+ FD-SIM	6.96	16.30	2.15	7.42	**0.86**	13.27	7.83 (↓39.90%)
	+ Ours	**4.67**	**13.28**	**1.44**	**2.12**	1.04	**6.71**	**4.88** (↓62.50%)

and 53.6%, respectively. Image throughput increases from 370.37 to 797.65 images/s, achieving a $2.15\times$ speedup. Similar improvements are observed on pMF-B/16, where peak memory decreases by 56.0%, median step time decreases by 40.2%, and throughput increases from 339.88 to 568.71 images/s, yielding a $1.67\times$ speedup. These results demonstrate that the computational savings of intra-encoder multi-view supervision translate into lower end-to-end memory consumption and faster training steps across both evaluated generator architectures.

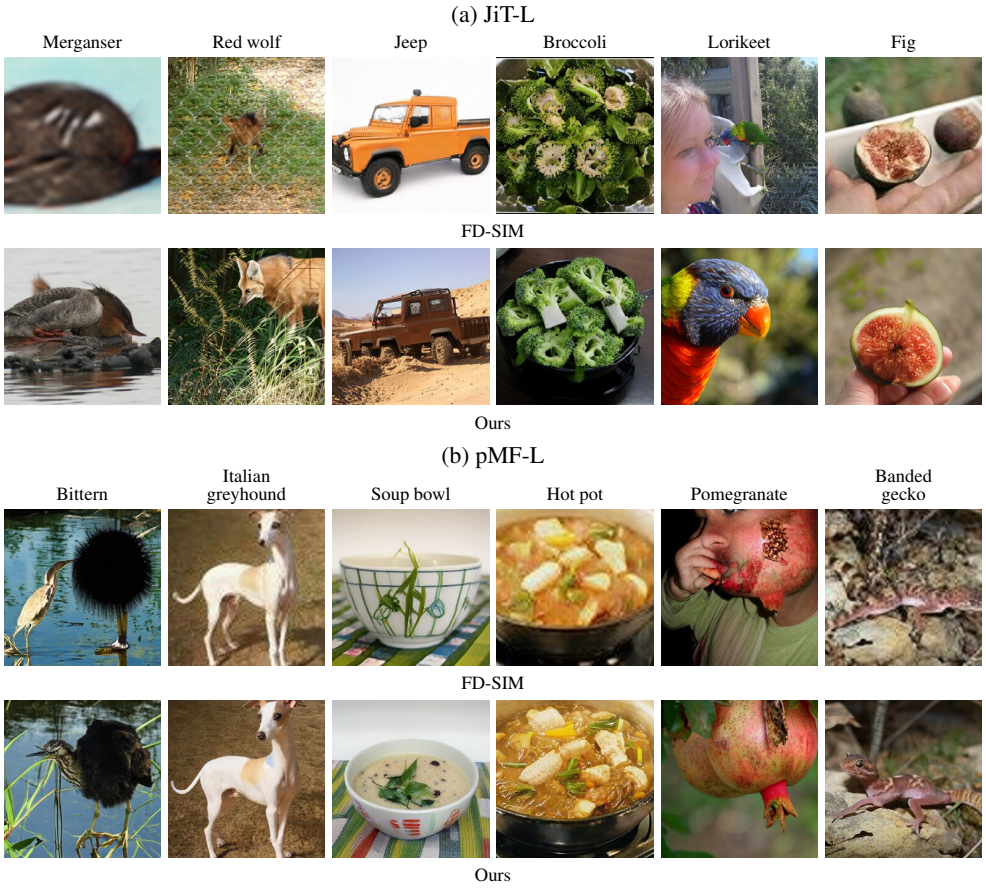


Figure 2: Qualitative comparison between IEMV-FD (Ours) and FD-SIM on ImageNet 256×256 using (a) JiT-L and (b) pMF-L. Each column shows samples generated from the same ImageNet class with identical initial noise.

Table 3: Training efficiency on one NVIDIA H200 under matched settings. Encoder GFLOPs represent the theoretical computational cost of the frozen supervision encoders, while peak GPU memory and step time are measured end to end across the entire training pipeline. Arrows indicate the preferred direction.

Method	# Visual Encoders ↓	Encoder GFLOPs/img ↓	Peak Memory (GiB) ↓	Step Time (s) ↓	Throughput (imgs/s) ↑
JiT-B/16					
FD-SIM	3	301.87	108.04	0.691	370.37
Ours	1	**66.65**	**54.57**	**0.321**	**797.65**
Reduction / Speedup	–	77.9% ↓	49.5% ↓	53.6% ↓	2.15×
pMF-B/16					
FD-SIM	3	301.87	136.05	0.753	339.88
Ours	1	**66.65**	**59.91**	**0.450**	**568.71**
Reduction / Speedup	–	77.9% ↓	56.0% ↓	40.2% ↓	1.67×

5.4 Ablation Studies

Token aggregation strategies. We conduct ablation studies on JiT-B/16 to investigate the effects of token aggregation strategies and network depth. As shown in Table 4(a), combining CLS-token

and average-pooled patch-token representations achieves lower FDr^6 and held-out FDr^6 than using either representation alone. With only the final layer, the combined configuration achieves a held-out FDr^6 of 11.24, compared with 15.01 for CLS-token supervision and 13.85 for average-pooled supervision. A similar pattern is observed with the four-layer configuration $\{4, 9, 10, 11\}$, where combining both representations achieves a held-out FDr^6 of 7.37, compared with 8.30 and 9.57 when using CLS-token or average-pooled supervision alone, respectively.

Layer selection. Table 4(b) further examines different layer selections while retaining both token aggregation strategies. Among the evaluated configurations, the selected layer set $\{4, 9, 10, 11\}$ achieves the lowest recorded FDr^6 and held-out FDr^6, with scores of 5.24 and 7.37, respectively. The alternative configuration $\{5, 8, 10, 11\}$ achieves comparable results of 5.43 and 7.50, while using only the final layer yields 12.06 and 11.24. These observations suggest that incorporating representations from multiple network depths can be beneficial and that the choice of layers affects the resulting supervision.

Table 4: Ablation study on JiT-B/16. Panel (a) varies token aggregation and depth jointly; panel (b) compares different layer sets while using both CLS and average-pooled tokens. Lower is better, and bold indicates the best result in each panel.

(a) Token aggregation

Layers (zero-indexed)	CLS	Avg. Token	$\mathrm{FDr}^6 \downarrow$	Held-out $\mathrm{FDr}^6 \downarrow$
$\{11\}$		✓	14.04	13.85
$\{11\}$	✓		20.42	15.01
$\{11\}$	✓	✓	12.06	11.24
$\{4, 9, 10, 11\}$		✓	7.37	9.57
$\{4, 9, 10, 11\}$	✓		6.05	8.30
$\{4, 9, 10, 11\}$	✓	✓	**5.24**	**7.37**

(b) Layer selection

CLS and Avg. Token are used for all rows.

Layers (zero-indexed)	$\mathrm{FDr}^6 \downarrow$	Held-out $\mathrm{FDr}^6 \downarrow$
$\{11\}$	12.06	11.24
$\{2, 5, 8, 11\}$	8.32	11.81
$\{4, 9, 10, 11\}$	**5.24**	**7.37**
$\{5, 8, 10, 11\}$	5.43	7.50
$\{8, 9, 10, 11\}$	9.17	10.58

6 Conclusion

We presented Intra-Encoder Multi-View Fréchet Supervision (IEMV-FD), which constructs diverse supervision views from different internal layers and token aggregation strategies of a single frozen visual encoder. By jointly optimizing the Fréchet objectives of these views through a shared encoder forward pass, IEMV-FD achieves multi-view distributional supervision without requiring multiple independent visual encoders. To assess generalization beyond directly supervised representations, we introduced held-out FDr^6, which evaluates performance across six representation spaces excluded from the training supervision of both methods. Experiments on ImageNet 256×256 demonstrate that IEMV-FD consistently achieves lower held-out FDr^6 than FD-SIM across all five evaluated generator configurations, while using only approximately 11.5% of its supervision-encoder parameters. Controlled efficiency benchmarks further demonstrate that this design substantially reduces computational costs, GPU memory consumption, and training time. These findings suggest that leveraging diverse internal representations of a single encoder provides an effective and computationally efficient alternative to multi-encoder Fréchet supervision.

Reproducibility Statement

To facilitate reproducibility, we provide a detailed description of our method in Section 4 and the experimental setup and evaluation protocols in Section 5. Additional implementation details, training configurations, sampling settings, and evaluation protocols are provided in the appendix. Our code and model checkpoints will be made publicly available.

AI Use Statement

Generative AI tools were used to improve the clarity and readability of the manuscript, assist with literature searches, and organize bibliographic references. All AI-assisted content, including references, was reviewed and verified by the authors. The authors take full responsibility for the final content of this paper.

References

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

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*, 2009.

Khoa D. Doan, Saurav Manchanda, Fengjiao Wang, Sathiya Keerthi, Avradeep Bhowmik, and Chandan K. Reddy. Image generation via minimizing Fréchet distance in discriminator feature space. *arXiv preprint arXiv:2003.11774*, 2020.

Cicero Nogueira dos Santos, Youssef Mroueh, Inkit Padhi, and Pierre Dognin. Learning implicit generative models by matching perceptual features. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for GAN training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of ImageNet classes in Fréchet inception distance. In *International Conference on Learning Representations*, 2023.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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*, 2023.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

Alexander Mathiasen and Frederik Hvilshøj. Backpropagating through Fréchet inception distance. *arXiv preprint arXiv:2009.14075*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *Advances in Neural Information Processing Systems*, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation Fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 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*, 2025.

A Post-Training Configuration

Table 5 summarizes the ImageNet 256×256 class-conditional post-training settings for the pMF, iMF, and JiT model families. Entries spanning all three model columns are shared settings. Initialization uses official pretrained base-model weights.

	pMF	iMF	JiT
base model			
Space	Pixel	Latent (SD-VAE)	Pixel
Sizes (256 px)	B, L	B	B, L
Patch size		16 px	
Initialization		Official pretrained weights	
representation models for FDr			
IEMV-FD		DINOv2-B layers $\{4, 9, 10, 11\}$, CLS + AVG	
Input resolution		256×256	
Normalization c		0.01	
Matrix square root		`torch.linalg.eigh`	
statistics estimator			
Default estimator		EMA, $\beta = 0.999$	
Warm-start samples		50,000 samples generated from the initialized base model	
training			
Epochs		50	
Optimizer		AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$	
Weight decay		0	
Learning rate	5×10^{-7}	5×10^{-7}	1×10^{-5}
LR schedule		Cosine	
Warmup epochs		5	
Global batch size	128	128	256
Precision		BF16	
Gradient clipping		None	
Model-weight EMA		0.9996	

Table 5: Configuration for IEMV-FD post-training. Layer indices are zero-based; AVG denotes average pooling over patch tokens. EMA for feature statistics and EMA for model weights are separate settings.

B Supplementary Quantitative Results

Tables 8–11 provide the detailed per-encoder results underlying the aggregate comparisons in the main paper. Official FDr^6 averages Inception, ConvNeXt, DINOv2, MAE, SigLIP, and CLIP. Strict FDr^4 averages DINOv3-L, PE-L, SwinV2-L, and EVA02-L. Held-out FDr^6 averages these four representations together with ConvNeXt and CLIP. Table 7 specifies the exact pretrained checkpoints and feature extraction settings used for these six held-out evaluation spaces. All values are rounded to two decimals, and lower is better.

B.1 Held-out Multi-Representation Evaluation

Following the normalization procedure of Representation Fréchet Loss (Yang et al., 2026), we introduce held-out FDr^6 to evaluate generative models in representation spaces that are not directly used for training supervision. Unlike the original FDr^6, which includes three encoders used by FD-SIM during training, held-out FDr^6 employs six evaluation encoders that are excluded from the supervision objectives of both IEMV-FD and FD-SIM.

Normalized Fréchet Distance Ratio. Following Yang et al. (2026), let $\mathcal{G}$, $\mathcal{T}$, and $\mathcal{V}$ denote the generated image set, ImageNet training set, and ImageNet validation set, respectively. For an eval-

uation encoder ϕ_i, the normalized Fréchet distance ratio is defined as

$$\mathrm{FDr}_{\phi_i}(\mathcal{G}) = \frac{\mathrm{FD}_{\phi_i}(\mathcal{G}, \mathcal{T})}{\mathrm{FD}_{\phi_i}(\mathcal{V}, \mathcal{T})}. \tag{B.1}$$

Here, $\mathrm{FD}_{\phi_i}(\mathcal{A}, \mathcal{B})$ denotes the Fréchet distance between the feature distributions of image sets $\mathcal{A}$ and $\mathcal{B}$, computed using encoder ϕ_i. The denominator provides a representation-specific normalization factor based on the distance between real training and validation images.

Held-out FDr^6. We define held-out FDr^6 as the arithmetic mean of the normalized Fréchet distance ratios across six evaluation encoders:

$$\mathrm{FDr}^6_{\text{held-out}}(\mathcal{G}) = \frac{1}{6} \sum_{\phi_i \in \mathcal{H}} \mathrm{FDr}_{\phi_i}(\mathcal{G}), \tag{B.2}$$

where the evaluation encoder set is

$$\mathcal{H} = \{\text{DINOv3-L},\ \text{PE-L},\ \text{SwinV2-L},\ \text{EVA02-L},\ \text{ConvNeXtV2-B},\ \text{CLIP ViT-L/14}\}. \tag{B.3}$$

The first four encoders provide additional evaluation representation spaces, while ConvNeXtV2-B and CLIP ViT-L/14 are retained from the original FDr^6 evaluation set. None of these six encoders is used for training supervision by either IEMV-FD or FD-SIM.

Evaluation Protocol. We follow a common evaluation protocol for both methods. For each generator checkpoint, we generate 50,000 class-balanced ImageNet images using fixed random seeds and matched sampling configurations within each generator architecture. The same generated samples are evaluated by all six encoders. We compute the normalized Fréchet distance ratio separately for each representation space and obtain the final held-out FDr^6 score using Eq. (B.2). The normalization constants for each encoder are shared across the compared methods.

Table 6: Sampling configurations for all evaluated models. Within each model, the baseline, FD-SIM, and our method share the same CFG scale and guidance interval.

Model	Method	Sampler	NFE	CFG	Guidance Interval
iMF-B	Baseline	Euler	1	8.0	$[0.4, 0.65]$
	+ FD-SIM	Euler	1		
	+ Ours	Euler	1		
pMF-B	Baseline	Euler	1	8.5	$[0.1, 0.7]$
	+ FD-SIM	Euler	1		
	+ Ours	Euler	1		
pMF-L	Baseline	Euler	1	7.0	$[0.2, 0.7]$
	+ FD-SIM	Euler	1		
	+ Ours	Euler	1		
JiT-B	Baseline	Heun	$50\times2\times2^\dagger$	3.0	$[0.1, 1.0]$
	+ FD-SIM	Euler	1		
	+ Ours	Euler	1		
JiT-L	Baseline	Heun	$50\times2\times2^\dagger$	2.4	$[0.1, 1.0]$
	+ FD-SIM	Euler	1		
	+ Ours	Euler	1		

† The JiT baseline NFE follows the source-paper Heun-sampler notation (Li & He, 2026), as in Table 1; the factors are not interpreted here as independent model calls.

This evaluation protocol allows us to examine whether improvements obtained through Fréchet-based training extend to representation spaces that are not directly optimized by either supervision objective.

B.2 HELD-OUT FDr^6 EVALUATION ENCODERS

Table 7: Visual encoders used to compute held-out FDr^6. "Dim." denotes the dimensionality of the extracted image representation, and the `timm` identifier specifies the exact pretrained checkpoint.

Model	`timm` identifier	Arch.	Dim.	Objective	Pooling
DINOv3-L	`vit_large_patch16_dinov3_qkvb.lvd1689m`	ViT	1024	contrastive	CLS token
PE-L	`vit_pe_core_large_patch14_336.fb`	ViT	1024	vision-language	MAP
SwinV2-L	`swinv2_large_window12to16_192to256.ms_in22k_ft_in1k`	Swin Transformer	1536	supervised	global avg. pool
EVA02-L	`eva02_large_patch14_448.mim_m38m_ft_in22k_in1k`	ViT	1024	masked image modeling	global avg. pool
ConvNeXtV2-B	`convnextv2_base.fcmae_ft_in22k_in1k`	CNN	1024	self-supervised	global avg. pool
CLIP ViT-L/14	`vit_large_patch14_clip_224.openai`	ViT	1024	vision-language	CLS token

Table 8: Per-encoder components of official FDr^6 for the main experiments. Lower is better, and bold marks the lowest value for each metric within each model.

Model	Method	Encoder-wise FDr ↓						FDr^6 ↓
		Inception	ConvNeXt	DINOv2	MAE	SigLIP	CLIP	
iMF-B	Baseline	2.07	3.27	14.49	12.14	32.51	28.06	15.42
	+ FD-SIM	**0.52**	**1.14**	6.31	**3.88**	9.08	12.77	5.61
	+ Ours	2.81	1.16	**3.15**	3.93	**8.81**	**10.32**	**5.03**
pMF-B	Baseline	1.95	1.89	10.11	14.76	31.01	23.26	13.83
	+ FD-SIM	**0.51**	**0.76**	4.11	**1.88**	**5.35**	8.71	3.55
	+ Ours	1.80	1.03	**2.60**	2.38	5.63	**7.76**	**3.53**
pMF-L	Baseline	1.60	1.34	6.72	10.43	20.24	14.69	9.17
	+ FD-SIM	**0.46**	**0.56**	2.24	**0.62**	**3.00**	5.85	2.12
	+ Ours	1.04	0.71	**1.52**	1.57	3.32	**4.52**	**2.11**
JiT-B	Baseline	2.20	4.48	10.69	12.99	30.21	34.28	15.81
	+ FD-SIM	**0.60**	**1.32**	5.14	**4.30**	**3.28**	19.32	5.66
	+ Ours	1.57	1.57	**2.82**	4.48	10.40	**10.61**	**5.24**
JiT-L	Baseline	1.54	3.46	6.05	8.66	19.27	25.82	10.80
	+ FD-SIM	**0.46**	**0.86**	2.64	**0.67**	**1.95**	13.27	3.31
	+ Ours	1.08	1.04	**1.50**	2.24	5.50	**6.71**	**3.01**

Table 9: Per-encoder components of held-out FDr^6 for the main experiments. Lower is better, and bold marks the lowest value for each metric within each model.

Model	Method	Encoder-wise FDr ↓						Held-out FDr^6 ↓
		DINOv3-L	PE-L	SwinV2-L	EVA02-L	ConvNeXt	CLIP-L	
iMF-B	Baseline	70.48	31.80	5.04	12.39	3.27	28.06	25.17
	+ FD-SIM	32.02	21.76	2.32	5.65	**1.14**	12.77	12.61
	+ Ours	**18.87**	**17.52**	**1.90**	**3.52**	1.16	**10.32**	**8.88**
pMF-B	Baseline	24.30	27.50	5.36	8.48	1.89	23.26	15.13
	+ FD-SIM	12.57	17.58	1.80	6.64	**0.76**	8.71	8.01
	+ Ours	**9.52**	**16.43**	**1.54**	**3.62**	1.03	**7.76**	**6.65**
pMF-L	Baseline	18.36	23.07	3.61	5.30	1.34	14.69	11.06
	+ FD-SIM	5.74	12.96	1.28	5.29	**0.56**	5.85	5.28
	+ Ours	**3.76**	**11.45**	**0.99**	**2.56**	0.71	**4.52**	**4.00**
JiT-B	Baseline	26.66	28.80	6.91	10.29	4.48	34.28	18.57
	+ FD-SIM	12.10	19.78	2.80	9.21	**1.32**	19.32	10.76
	+ Ours	**8.93**	**17.41**	**2.02**	**3.68**	1.57	**10.61**	**7.37**
JiT-L	Baseline	14.38	22.59	5.36	6.49	3.46	25.82	13.02
	+ FD-SIM	6.96	16.30	2.15	7.42	**0.86**	13.27	7.83
	+ Ours	**4.67**	**13.28**	**1.44**	**2.12**	1.04	**6.71**	**4.88**

C SUPPLEMENTARY ABLATION RESULTS

Table 10: Per-encoder components of official FDr^6 for the JiT-B ablations. Lower is better, and bold marks the lowest value in each metric column.

CLS layers	AVG layers	Encoder-wise FDr ↓						FDr^6 ↓
		Inception	ConvNeXt	DINOv2	MAE	SigLIP	CLIP	
–	{11}	3.56	3.56	4.97	22.56	22.60	27.00	14.04
{11}	–	3.90	3.98	3.71	46.21	35.38	29.37	20.42
{11}	{11}	2.63	2.53	3.08	21.37	21.53	21.21	12.06
–	{4, 9, 10, 11}	2.41	2.76	5.33	5.71	13.06	14.94	7.37
{4, 9, 10, 11}	–	2.45	1.87	3.11	5.49	12.06	11.31	6.05
{4, 9, 10, 11}	{4, 9, 10, 11}	**1.57**	**1.57**	**2.82**	**4.48**	**10.40**	**10.61**	**5.24**
{2, 5, 8, 11}	{2, 5, 8, 11}	3.25	3.44	6.40	6.31	16.15	14.36	8.32
{5, 8, 10, 11}	{5, 8, 10, 11}	1.78	1.60	2.89	4.61	10.52	11.15	5.43
{8, 9, 10, 11}	{8, 9, 10, 11}	2.88	2.35	3.45	9.58	16.39	20.37	9.17

Table 11: Per-encoder components of held-out FDr^6 for the JiT-B ablations. Lower is better, and bold marks the lowest value in each metric column.

CLS layers	AVG layers	Encoder-wise FDr ↓						Held-out FDr^6 ↓
		DINOv3-L	PE-L	SwinV2-L	EVA02-L	ConvNeXt	CLIP-L	
–	{11}	13.48	22.53	6.02	10.49	3.56	27.00	13.85
{11}	–	10.19	28.10	9.43	8.97	3.98	29.37	15.01
{11}	{11}	9.31	21.83	6.11	6.46	2.53	21.21	11.24
–	{4, 9, 10, 11}	10.83	19.96	3.05	5.85	2.76	14.94	9.57
{4, 9, 10, 11}	–	11.46	18.50	2.43	4.25	1.87	11.31	8.30
{4, 9, 10, 11}	{4, 9, 10, 11}	8.93	**17.41**	**2.02**	**3.68**	**1.57**	**10.61**	**7.37**
{2, 5, 8, 11}	{2, 5, 8, 11}	19.97	22.45	3.76	6.90	3.44	14.36	11.81
{5, 8, 10, 11}	{5, 8, 10, 11}	**8.50**	17.66	2.12	3.97	1.60	11.15	7.50
{8, 9, 10, 11}	{8, 9, 10, 11}	10.75	19.63	3.46	6.93	2.35	20.37	10.58

D Additional Qualitative Comparisons

Figures 3 and 4 show additional non-duplicate pairs from the author-provided JiT-L comparison set. Figures 5 and 6 provide broader paired comparisons for pMF-L. Within each comparison, corresponding samples use the same ImageNet class condition and evaluation sample ID.

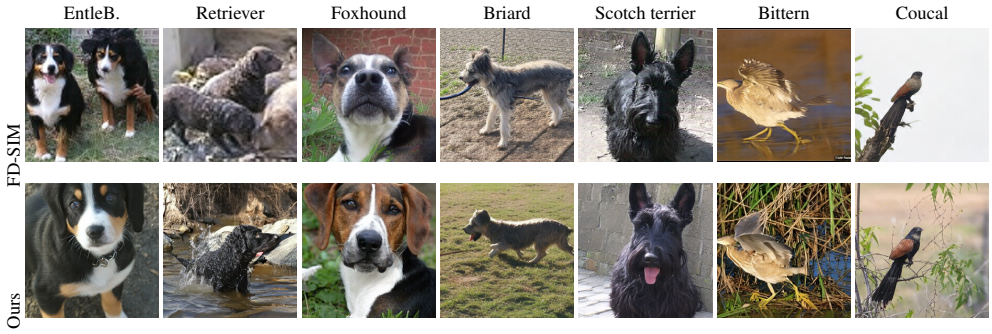


Figure 3: Additional paired JiT-L samples (part 1). Rows compare IEMV-FD with FD-SIM under matched class conditions and sample IDs.

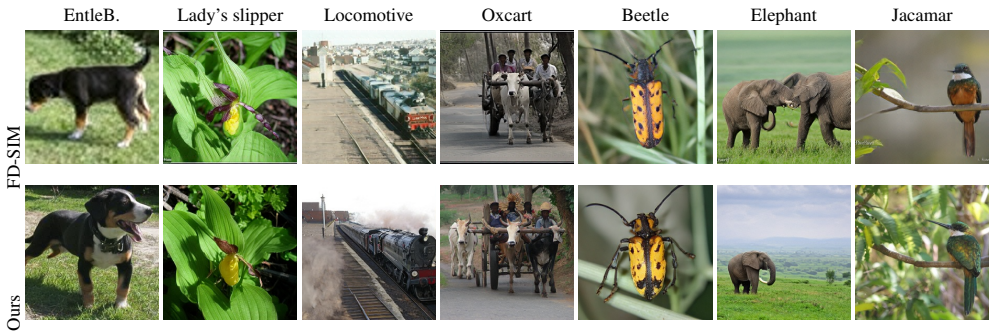


Figure 4: Additional paired JiT-L samples (part 2). Rows compare IEMV-FD with FD-SIM under matched class conditions and sample IDs.



Figure 5: Additional paired pMF-L samples (part 1). Each spatial position compares IEMV-FD (left) with FD-SIM (right) under matched class conditions and sample IDs.

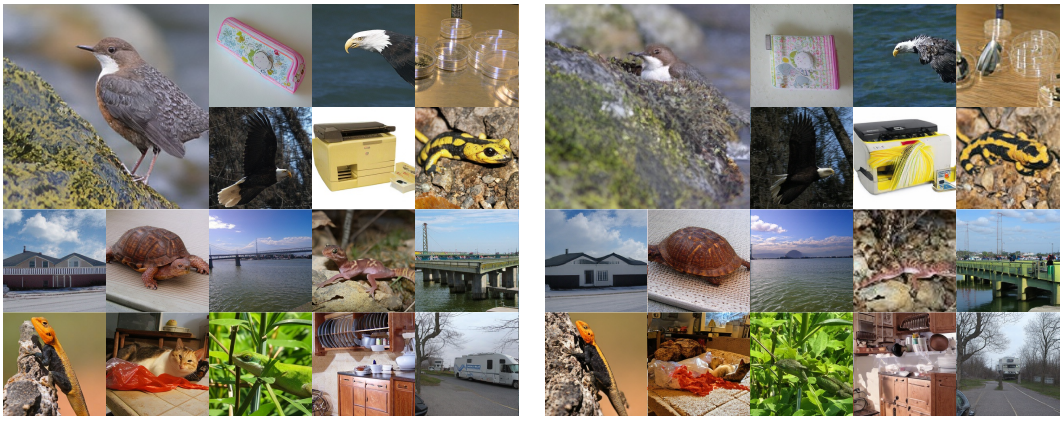


Figure 6: Additional paired pMF-L samples (part 2). Each spatial position compares IEMV-FD (left) with FD-SIM (right) under matched class conditions and sample IDs.

E TRAINING ALGORITHM

Algorithm 1 presents the detailed training procedure of IEMV-FD. Following FD-loss, we precompute the real-image reference statistics and maintain the generated-image statistics using an exponential moving average (EMA). The key difference lies in the construction and optimization of multiple supervision views. All selected representations are extracted through a single forward pass of the frozen visual encoder, and a separate Fréchet loss is computed for each view. These losses are averaged to jointly optimize the generator. Historical EMA statistics are detached from the computation graph, allowing gradients to propagate only through the current batch.

Algorithm 1 IEMV-FD Training

Require: Generator G_θ, frozen encoder ϕ, view set $\mathcal{S}$, dataset $\mathcal{D}_r$, EMA decay β
Ensure: Updated generator parameters θ

```
Compute real reference statistics {(μ_h^r, Σ_h^r)}_{h∈S} using D_r
Initialize generated EMA statistics {(μ_h^g, M_h^g)}_{h∈S}
for each training iteration do
    Sample generator inputs (z, y)
    Generate images x_g = G_θ(z, y)
    Extract all selected views {h(x_g)}_{h∈S} using one forward pass through φ
    for each view h ∈ S do
        Compute batch statistics (μ_{h,b}, M_{h,b}) using Eq. (2)
        Compute updated EMA statistics (μ̃_h^g, M̃_h^g) using Eqs. (3)–(4)
        Compute Fréchet loss L_h using Eq. (1)
    end for
    Compute joint loss L_IEMV-FD using Eq. (8)
    Update generator parameters θ
    Store updated EMA statistics with gradients detached
end for
```