CROSS-SCALE ALIGNED SUPERVISION FOR TRAINING GANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern GANs often use multi-scale adversarial supervision to provide direct realism feedback to intermediate generator outputs. However, we identify an important limitation of standard scale-wise supervision: each intermediate output is independently optimized for realism at its own resolution, without explicitly coordinating outputs across generator stages. As a result, realistic intermediate outputs need not remain aligned with the final output from the same generated sample. This motivates treating per-scale realism and cross-scale consistency as complementary objectives for multi-scale adversarial generation. Based on this observation, we propose CAT, a Cross-scale Aligned Transformer, which preserves scale-wise adversarial supervision for direct per-scale realism feedback while introducing a simple generator-side consistency regularization to align intermediate outputs with the final output. We further show that jointly exposing multiple scales to the discriminator improves cross-scale consistency but substantially degrades generation quality, supporting our design of enforcing cross-scale consistency separately on the generator side. In a matched 40-epoch CAT-H/2 comparison on ImageNet-256, the proposed consistency regularization improves FID-50K from 4.35 to 1.76. CAT-H/2 achieves an FID-50K of 1.56 with a single generator evaluation after only 60 training epochs.

1 INTRODUCTION

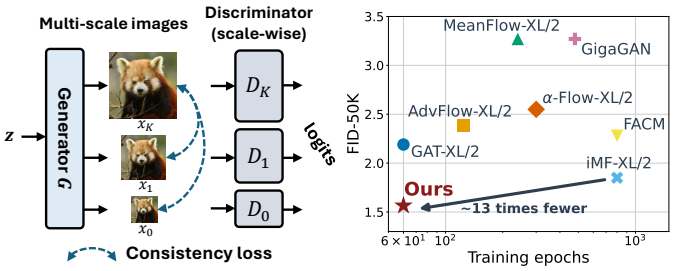
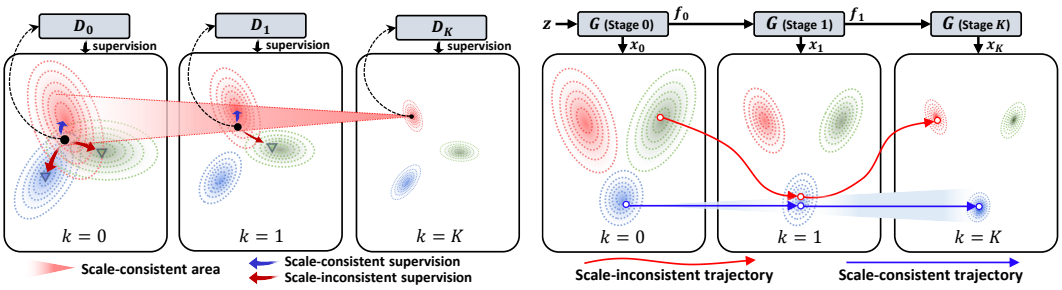


Figure 1: **Method overview and comparison with 1-step baselines.** (Left) Our method combines generator-side consistency regularization with a scale-wise discriminator. (Right) Our method achieves strong FID with substantially fewer training epochs in one-step generation in ImageNet-256.

Recent generative models have achieved remarkable progress in image synthesis. A common principle behind many of these advances is to decompose generation into a sequence of intermediate prediction steps rather than synthesizing the final image at once. For example, diffusion and flow-based models (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2023; Peebles & Xie, 2023) progressively transform an initial noise sample toward data, while autoregressive and masked prediction models (Esser et al., 2021; Chang et al., 2022; Tian et al., 2024; Li et al., 2024) generate an image through a sequence of predictions. In these paradigms, intermediate states are not merely auxiliary predictions; they actively participate in the generation process and guide the model toward the final sample.

Another representative family of generative models is Generative Adversarial Networks (Goodfellow et al., 2014) (GANs). GANs naturally generate images with a single generator forward pass, but they can also benefit from intermediate supervision inside the generator. Prior multi-stage GANs therefore expose intermediate generator outputs and provide adversarial feedback at multiple levels or image scales (Karnewar & Wang, 2020; Lin et al., 2021; Kang et al., 2023; Hyun et al., 2026). Importantly,

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.



(a) Independent per-scale objectives (b) Possible cross-depth trajectories

Figure 2: **Failure mode of scale-wise adversarial supervision: per-scale realism does not ensure cross-scale consistency.** (a) Each intermediate output is independently optimized for realism at its corresponding resolution, without explicitly constraining outputs across generator depths to represent the same generated sample. (b) Subsequent synthesis proceeds from the generator feature f_k, rather than explicitly refining the intermediate image x_k, allowing later stages to substantially revise earlier predictions. Together, these properties can produce individually realistic intermediate outputs that are not consistently aligned toward the final sample. The diagrams illustrate output relationships across generator depths, rather than a progressively growing spatial architecture.

how adversarial feedback is organized across generator stages determines how effectively multi-scale supervision guides the generation process. Some approaches jointly discriminate predictions from multiple stages (Karnewar & Wang, 2020; Hyun et al., 2026), whereas others preserve scale-specific feedback by evaluating generated outputs at different image scales separately (Lin et al., 2021; Kang et al., 2023).

Despite its effectiveness, we identify an important limitation of standard *scale-wise adversarial supervision*: direct realism feedback at individual scales does not ensure consistency across generator stages. Outputs from successive generator depths are independently evaluated at progressively finer image resolutions, encouraging a coarse-to-fine generation process. However, each output is optimized for realism at its own scale without explicitly constraining its relationship to outputs from other stages. Consequently, *scale-wise realism does not necessarily imply sample-wise cross-scale consistency*, allowing later stages to substantially rewrite the generated content rather than coherently progress toward the final output.

Fig. 2 illustrates why such misalignment can occur. First, scale-wise adversarial supervision independently compares each intermediate output with real images at its corresponding resolution, without explicitly coordinating outputs across generator stages. Second, subsequent generator blocks operate on hidden features f_k, and their output updates are not explicitly constrained to preserve earlier predicted content. Together, these properties allow substantial changes across stages despite direct per-scale realism feedback.

These observations motivate explicitly combining two complementary objectives for multi-scale adversarial generation: *per-scale realism* and *cross-scale consistency*. Standard scale-wise adversarial supervision directly provides the former, but leaves the latter largely unconstrained.

To handle this inconsistency, we propose CAT, a Cross-scale Aligned Transformer for multi-scale adversarial generation. Based on a modern transformer GAN (Hyun et al., 2026), CAT retains scale-wise adversarial supervision to preserve direct realism feedback at each image scale, while introducing a simple generator-side consistency regularization that encourages agreement between intermediate and final outputs from the same generated sample. This allows intermediate stages to benefit from scale-specific adversarial supervision while promoting a coherent generation trajectory. As qualitatively shown in Fig. 3, without $\mathcal{L}_{\text{cons}}$, intermediate outputs can exhibit noticeable changes in color and structure across stages, whereas $\mathcal{L}_{\text{cons}}$ encourages them to remain consistent with the final output.

We further examine whether per-scale realism and cross-scale consistency can be reconciled within the adversarial objective itself. To this end, we consider a scale-aggregated alternative that jointly exposes the full image pyramid to the discriminator, allowing it to explicitly model relationships across scales. While this formulation improves cross-scale consistency, it substantially degrades

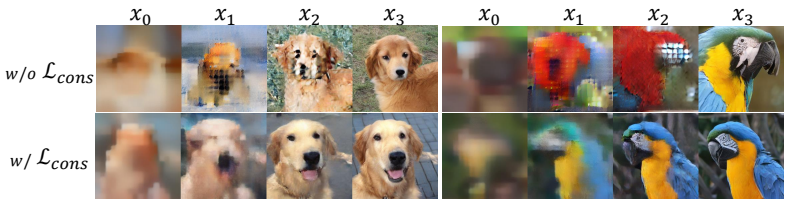


Figure 3: Comparison of multi-scale images generated without (top) and with (bottom) consistency regularization ($\mathcal{L}_{cons}$).

generation quality. This result supports our design choice to preserve direct scale-wise realism feedback in the discriminator while enforcing cross-scale consistency separately on the generator side.

Quantitative results further validate this design. The proposed generator-side consistency improves FID-50K from 4.35 to 1.76 in a matched CAT-H/2 experiment at 40 epochs. With a fixed Base discriminator, scaling the generator from Medium to Huge improves FID only when consistency regularization is applied in our tested settings. On ImageNet-256, CAT-H/2 achieves an FID-50K of 1.56 after 60 training epochs with a single generator evaluation, while requiring $2.2\times$ and $16.7\times$ lower estimated training compute than GAT-XL/2 and the one-step iMF-XL/2 baseline, respectively. These results demonstrate the effectiveness of combining per-scale realism with cross-scale consistency for multi-scale adversarial generation.

2 Related Work

One-step image generation. A major goal of recent generative modeling is to reduce sampling cost while preserving image quality. GANs (Goodfellow et al., 2014; Brock et al., 2019; Karras et al., 2020; Sauer et al., 2022; Kang et al., 2023; Hyun et al., 2026) naturally provide one-step generation, but their scalability has often lagged behind diffusion and autoregressive models. Recent diffusion and flow-based methods (Song et al., 2023; Song & Dhariwal, 2024; Frans et al., 2025; Geng et al., 2025; 2026) also pursue few-step or one-step generation by learning direct or averaged transport trajectories. While these methods improve inference efficiency, they often require long training schedules or specialized objectives. In contrast, our work revisits one-step generation from the adversarial learning perspective and shows that properly organized hierarchical supervision can make transformer-based GANs efficient and competitive.

Multi-scale adversarial supervision. Multi-scale supervision has been widely explored in GANs, ranging from progressive growing (Karras et al., 2018) to adversarial supervision of intermediate generator outputs (Karnewar & Wang, 2020; Lin et al., 2021; Kang et al., 2023; Hyun et al., 2026). These methods provide adversarial feedback at multiple generator stages or image scales, and are commonly associated with hierarchical or coarse-to-fine generation. Different approaches organize this feedback differently: some jointly expose multiple scales to the discriminator (Karnewar & Wang, 2020; Hyun et al., 2026), whereas others preserve scale-specific adversarial feedback (Lin et al., 2021; Kang et al., 2023). Our analysis shows that standard scale-wise supervision effectively provides *per-scale realism*, but does not explicitly enforce *cross-scale consistency* among outputs generated from the same latent sample. CAT preserves this scale-wise realism feedback and introduces generator-side *cross-scale alignment* to coordinate intermediate outputs with the final sample.

3 Preliminary

Multi-stage adversarial generation in GANs. Several GAN frameworks expose intermediate generator outputs and apply adversarial supervision at multiple stages or image scales (Karnewar & Wang, 2020; Kang et al., 2023; Hyun et al., 2026). Such intermediate supervision provides direct learning signals throughout the generator and is commonly associated with hierarchical or coarse-to-fine generation. It can be realized through multi-scale images (Karnewar & Wang, 2020; Kang et al., 2023) or multi-level noise perturbations (Hyun et al., 2026); in this work, we focus on the multi-scale image formulation.

We denote by x_k the image used for adversarial supervision at stage or scale k, where $k = 0, \ldots, K$ and x_K is the final output. In *scale-wise adversarial supervision* (Lin et al., 2021; Kang et al., 2023),

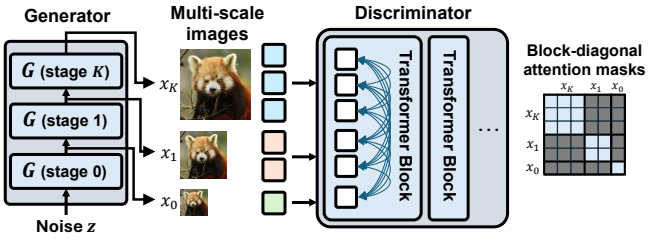


Figure 4: **Scale-wise supervision setup.** The generator produces scale-specific images $\{x_k\}_{k=0}^{K}$. A block-diagonal attention mask prevents cross-scale interaction in the shared discriminator, ensuring that each prediction d_k depends only on its corresponding image x_k.

each x_k is independently evaluated at its corresponding resolution. Let d_k denote the discriminator prediction for x_k. Since d_k is computed only from the corresponding image x_k, each generator-stage output receives direct per-scale realism feedback without cross-scale information exchange inside the discriminator.

The multi-scale adversarial objective is

$$\mathcal{L}_{\text{adv}} = \frac{1}{K+1} \sum_{k=0}^{K} \mathcal{L}_{\text{GAN}}(G, D, k), \tag{1}$$

where $\mathcal{L}_{\text{GAN}}(G, D, k)$ denotes the adversarial loss computed from the scale-k discriminator prediction.

4 Proposed Method

4.1 Cross-scale inconsistency in scale-wise supervision

Problem. Scale-wise adversarial supervision provides direct *per-scale realism* feedback, but does not explicitly enforce *cross-scale consistency* across generator stages. In a coherent coarse-to-fine generation process, intermediate predictions should remain aligned with the same final output while later stages introduce progressively finer content. However, because each intermediate image is independently evaluated against the real distribution at its corresponding resolution, the adversarial objective does not impose sample-wise correspondence across stages. We therefore examine whether standard scale-wise supervision produces intermediate outputs with the cross-scale consistency required for such generation.

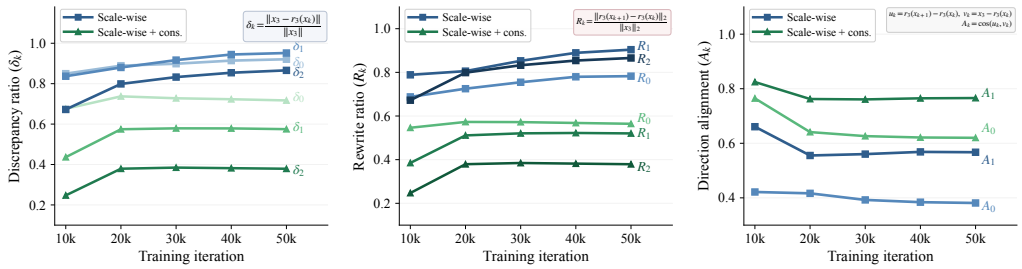
Scale-wise supervision setup. We analyze this behavior using a transformer generator (Hyun et al., 2026). The generator operates on a fixed latent grid and accumulates stage-wise image readouts through output skip connections. We denote the accumulated output after stage k by h_k. Since all h_k share the same latent resolution, the scale-specific image used for adversarial supervision is obtained by resizing:

$$x_k = r_k(h_k), \tag{2}$$

where $r_k(\cdot)$ denotes resizing to the resolution associated with scale k.

As illustrated in Fig. 4, we use a shared transformer discriminator with a block-diagonal attention mask across scales. Tokens from all scales are concatenated into a single sequence for efficient computation, while the mask prevents cross-scale information exchange. Consequently, each prediction d_k depends only on its corresponding image x_k, realizing scale-wise adversarial supervision within a shared discriminator.

Unless otherwise specified, our analyses use the ImageNet-256 latent-space setting with SD-VAE latents (Rombach et al., 2022). We use Base-scale generator and discriminator models with 12 layers and 768 hidden channels, trained for 20 epochs (50K iterations). For brevity, we refer to the latent outputs as images.



(a) Distance to highest-scale (b) Inter-stage rewrite magnitude (c) Rewrite direction alignment

Figure 5: **Cross-scale inconsistency analysis.** We evaluate whether intermediate outputs remain consistent with the final output under standard scale-wise supervision. (a) The normalized discrepancy (δ_k) remains large, indicating that intermediate outputs stay far from the final output. (b) The inter-stage rewrite magnitude (R_k) is also large, showing substantial changes between successive stage outputs. (c) The direction alignment (A_k) remains low, indicating that these changes are weakly aligned with the remaining direction toward the final output. The proposed consistency regularization reduces δ_k and R_k while increasing A_k, improving cross-scale consistency.

Metrics. We quantify cross-scale consistency by comparing stage-wise outputs after resizing them to the final resolution. Let x_K denote the final-stage output. We define

$$\begin{aligned} \delta_k &= \frac{\|x_K - r_K(x_k)\|_2}{\|x_K\|_2}, \\ R_k &= \frac{\|r_K(x_{k+1}) - r_K(x_k)\|_2}{\|x_K\|_2}, \\ A_k &= \cos\left(r_K(x_{k+1}) - r_K(x_k),\, x_K - r_K(x_k)\right). \end{aligned} \tag{3}$$

The discrepancy δ_k measures how far an intermediate output remains from the final output, while the rewrite magnitude R_k measures the amount of change between successive stages. The direction alignment A_k measures whether this stage-wise change points toward the remaining difference to the final output. Since large δ_k and R_k can also arise from resolution-dependent detail refinement, A_k provides a complementary measure of whether these changes are directed toward the same final output. Under a coherent coarse-to-fine process, we expect the remaining discrepancy and the magnitude of stage-wise rewriting to decrease toward finer stages, while the updates remain well aligned with the final output. We omit the final transition from the alignment plots because $A_{K-1} = 1$ whenever its update is nonzero.

Observation. As shown in Fig. 5, standard scale-wise supervision does not exhibit these behaviors. Both δ_k and R_k remain large throughout training, often exceeding 0.8, while A_k remains relatively low. Thus, intermediate outputs remain substantially different from the final output and undergo large stage-wise changes that are only weakly aligned with the remaining direction toward it. Moreover, δ_k and R_k tend to increase during training and do not systematically decrease toward finer stages. The examples without consistency regularization in Fig. 3 further show changes in color and structure across scales. Together, these observations suggest that direct per-scale realism feedback alone does not reliably establish cross-scale consistent coarse-to-fine generation in this setting.

4.2 Cross-scale aligned supervision

The analysis above suggests that scale-wise adversarial supervision provides direct *per-scale realism* feedback, but lacks an explicit constraint for *cross-scale consistency*. We therefore retain scale-wise discrimination and introduce generator-side consistency regularization that aligns intermediate outputs with the final output.

Generator-side consistency regularization. We implement this alignment as a consistency loss on the stage-wise outputs. The goal is not to introduce another realism objective, but to coordinate the intermediate outputs that already receive scale-wise adversarial supervision. We use the final-stage output as a common reference and penalize differences before resizing, where all stage-wise outputs

share the same latent resolution:

$$\mathcal{L}_{\text{cons}} = \frac{1}{K} \sum_{k=0}^{K-1} w_k \left\| h_k - h_K \right\|_2^2, \tag{4}$$

where h_K is the final-stage output and w_k is a scale weight. We use weaker weights for coarser stages because low-resolution outputs are inherently more ambiguous, as multiple high-resolution samples can share similar coarse structure. This relaxes the point-to-point constraint on early stages while still encouraging their outputs to remain consistent with the final output.

As shown in Fig. 5, the proposed consistency regularization directly mitigates the behaviors identified above, reducing both discrepancy and inter-stage rewriting while improving direction alignment.

Cross-scale Aligned Transformer. Based on this regularization, CAT combines scale-wise adversarial supervision with explicit generator-side cross-scale consistency. The adversarial objective provides direct realism feedback at each scale, while the consistency objective coordinates the corresponding intermediate outputs toward the same final output. The resulting generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}, \tag{5}$$

where λ_{cons} controls the strength of consistency regularization. The discriminator objective remains unchanged.

5 Experiments

Experimental settings. We evaluate class-conditional image generation on ImageNet-256 (Deng et al., 2009) at 256×256 resolution. Following prior latent-space one-step generators, we train all models in the latent space of SD-VAE (Rombach et al., 2022). Our implementation is largely based on GAT (Hyun et al., 2026): we adopt its generator and most training configurations, including the objective functions. For CAT, we use scale-wise adversarial supervision with generator-side consistency regularization. Following prior work (Geng et al., 2025; 2026), we report FID-50K (Heusel et al., 2017) using statistics computed from the full ImageNet training set.

Implementation details. Unless otherwise specified, we use $\lambda_{\text{cons}} = 0.1$. The scale weights w_k decrease toward lower resolutions: for $K = 3$, we set $w_0 = 1/3$, $w_1 = 1/2$, and $w_2 = 1$.

Table 1: Model configs.

Model	L	C	Params	GFLOPs
G-B/2	12	768	133M	23.0
G-M/2	24	768	261M	46.0
G-H/2	32	1280	960M	166.8
D-B/2	12	768	96M	33.9

We find that stable generator scaling can be achieved without increasing the discriminator capacity or reducing the generator learning rate, unlike the recipe used in GAT (Hyun et al., 2026). Thus, unless otherwise specified, we use a *Base* discriminator for all generator scales (*Base*, *Medium*, *Huge*) and use the same learning rate for the generator and discriminator, as summarized in Table 1. We use a batch size of 512, where 50K iterations correspond to 20 epochs, and a learning rate of 2×10^{-4}. For multi-scale adversarial supervision, we use token resolutions of $2^2, 4^2, 8^2$, and 16^2. Compared with the original GAT discriminator operating on 16^2 tokens, this increases the discriminator input token count from 256 to 340, or about 33%. This introduces only modest overhead, especially since we keep the discriminator at the *Base* scale for all generator sizes rather than scaling it with the generator. Further details are provided in Appendix A.1.

Comparison with prior work. As shown in Table 2, we compare CAT with prior work. Among 1-NFE models trained from scratch, CAT-H/2 achieves a state-of-the-art FID of 1.56. Notably, it improves over recent 1-NFE diffusion/flow models, including iMF-XL/2 (Geng et al., 2026), reducing FID from 1.72 to 1.56 with only 60 training epochs compared with 800. CAT-H/2 also establishes a new state of the art among GAN-based models, outperforming strong recent baselines such as GAT-XL/2 (Hyun et al., 2026). Although CAT-H/2 uses more parameters, this does not directly translate into higher practical cost, as analyzed in Table 5.

Table 2: **Class-conditional generation on ImageNet-256×256 (FID-50K).** All models use one function evaluation (1-NFE). GFLOPs denotes the generator's inference cost.

Method	Params	GFLOPs	Epochs	FID
1-NFE diffusion/flow from scratch				
iCT-XL/2 (Song & Dhariwal, 2024)	675M	119	-	34.24
Shortcut-XL/2 (Frans et al., 2025)	675M	119	250	10.60
MeanFlow-XL/2 (Geng et al., 2025)	676M	119	240	3.43
α-Flow-XL/2 (Zhang et al., 2026)	676M	119	300	2.58
FACM (Peng et al., 2026)	675M	119	800	2.27
iMF-XL/2 (Geng et al., 2026)	610M	175	800	1.72
1-NFE GANs from scratch				
BigGAN (Brock et al., 2019)	112M	59	-	6.95
GigaGAN (Kang et al., 2023)	569M	-	480	3.45
AdvFlow-XL/2 (Lin et al., 2026)	673M	-	125	2.38
StyleGAN-XL (Sauer et al., 2022)	166M	1574	-	2.30
GAT-XL/2 (Hyun et al., 2026)	602M	119	60	2.18
CAT-M/2 (Ours)	261M	46	**40**	**1.93**
CAT-H/2 (Ours)	960M	167	**60**	**1.56**

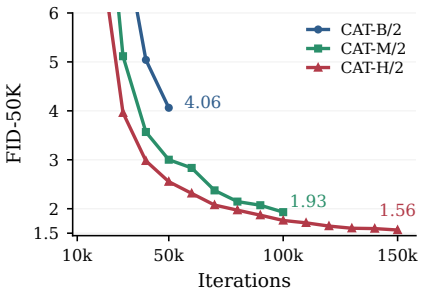


Figure 6: FID-50K training curve.

Model	$\mathcal{L}_{\text{cons}}$	FID-50K 20 epo	FID-50K 40 epo
G-B/2		5.43	–
G-B/2	✓	4.06	–
G-M/2		3.27	2.34
G-M/2	✓	**3.00**	1.93
G-H/2		6.86	4.35
G-H/2	✓	**2.55**	**1.76**

Table 3: Ablation study on $\mathcal{L}_{\text{cons}}$. Consistency regularization becomes effective across model sizes, with larger gains for larger models and longer training.

λ_{cons}	0.0	0.1	1.0
FID-50K	5.43	**4.06**	4.45

Table 4: Analysis on the strength of λ_{cons} (G-B/2, 20 epoch).

Training dynamics and consistency ablations. Fig. 6 shows the FID-50K training curves of CAT with different generator sizes. CAT consistently benefits from scaling the generator, and larger models continue to improve with longer training. In particular, CAT-H/2 steadily improves up to 150K iterations, reaching an FID of 1.56, while CAT-M/2 reaches 1.93 at 100K iterations. This suggests that the proposed method remains stable and scalable under longer training.

Table 3 further verifies the effect of the proposed consistency regularization. For G-B/2, adding $\mathcal{L}_{\text{cons}}$ improves FID from 5.43 to 4.06 at 20 epochs. For G-M/2, the gain becomes more pronounced with longer training: $\mathcal{L}_{\text{cons}}$ improves FID from 3.27 to 3.00 at 20 epochs and from 2.34 to 1.93 at 40 epochs. For G-H/2, consistency regularization improves FID from 4.35 to 1.76 at 40 epochs. At this training budget, scaling the generator from M/2 to H/2 with a fixed Base discriminator worsens FID without $\mathcal{L}_{\text{cons}}$ but improves it with regularization. Finally, Table 4 studies the strength of the consistency loss. A moderate weight of $\lambda_{\text{cons}} = 0.1$ gives the best result, showing that explicit cross-scale alignment is beneficial while overly strong consistency can over-constrain the generator.

Training and inference efficiency. Table 5 shows that CAT-H/2 is computationally efficient among strong one-step generators. Compared with the one-step diffusion/flow baseline iMF-XL/2 (Geng et al., 2026), CAT-H/2 achieves better FID with lower training and inference cost, while requiring over $16\times$ lower estimated total training compute. CAT-H/2 is also much cheaper to train than GAT-XL/2 (Hyun et al., 2026), reducing total training compute by about $2.2\times$. Further details of the compute comparison are provided in Appendix A.2.

Table 6 further shows that the gain does not simply come from increasing adversarial model capacity. GAT jointly processes multi-scale evidence inside the discriminator, whereas CAT keeps discriminator feedback scale-wise and imposes alignment through generator-side consistency. We observe that this design provides a substantially stronger capacity-performance trade-off. CAT-B/2 already achieves an

Table 5: Compute comparison (GFLOPs). Training and inference compute is measured per-iteration and sample.

Method	Train GFLOPs	Infer. GFLOPs	Epochs	Total ($\times 10^3$)	Rel. Total
iMF-XL/2	1,306.3	174.6	800	1,045.0	16.7×
GAT-XL/2	2,297.2	118.6	60	137.8	2.2×
CAT-H/2	1,040.2	166.7	60	62.4	1.0×

Table 6: Comparison with GAT across various model sizes (20 epochs).

Method	$(G+D)$ Params	FID↓	IS↑
GAT-B/2	116+104M	9.534	124.69
GAT-XL/2	602+467M	4.021	204.26
CAT-B/2	133+96M	4.063	174.54
CAT-H/2	960+96M	2.552	242.45

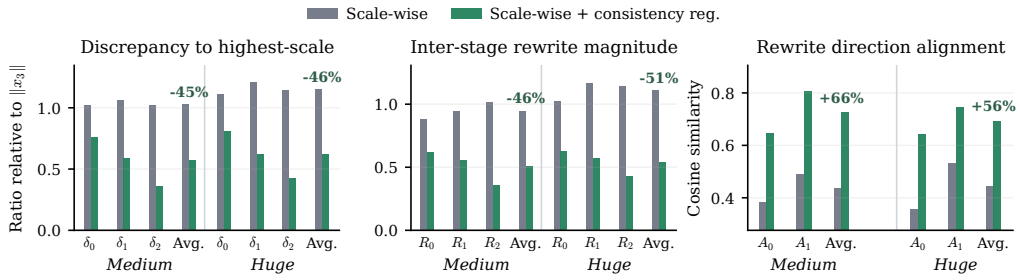


Figure 7: **Consistency effects across model sizes.** For both Medium and Huge models trained for 40 epochs, $\mathcal{L}_{\text{cons}}$ reduces discrepancy and rewrite magnitude while improving update-direction alignment at every measured stage.

FID comparable to GAT-XL/2 (4.063 vs. 4.021), despite using a *Base*-scale configuration. Moreover, CAT-H/2 substantially outperforms GAT-XL/2 (2.552 vs. 4.021 FID), even though the two models have comparable total $(G+D)$ parameters. These results suggest that the advantage of CAT comes from organizing hierarchical adversarial supervision to preserve direct scale-wise feedback while maintaining cross-scale consistency, rather than simply scaling both the generator and discriminator.

Cross-scale effect of consistency regularization. We further verify whether the proposed consistency loss improves the cross-scale consistency of scale-wise supervision. As shown in Fig. 7, adding $\mathcal{L}_{\text{cons}}$ consistently reduces the discrepancy to the final output δ_k and the inter-stage rewrite magnitude R_k, while improving the direction alignment A_k, across both *Medium* and *Huge* models trained for 40 epochs. On average, $\mathcal{L}_{\text{cons}}$ reduces δ_k by 45% and 46%, and R_k by 46% and 51%, for *Medium* and *Huge* models, respectively. It also improves A_k by 66% and 56%, showing that stage-wise updates become better aligned with the remaining direction toward the final output. These consistency improvements also hold in LPIPS and DINOv2-L feature spaces (Appendix A.6).

Effect of discriminator scaling. We further examine the effect of discriminator scaling under a fixed G-M/2 generator. To isolate the role of discriminator capacity, we change only the discriminator size while keeping the generator and training setup unchanged. As shown in Table 7, increasing the discriminator from D-B/2 to D-M/2 improves FID-50K from 5.11 to 4.28 after 12 epochs. This shows that CAT can still benefit from a stronger discriminator, although our main experiments intentionally use a *Base* discriminator to maintain training efficiency. These results suggest that further gains may be possible by studying the generator-discriminator capacity balance more systematically.

Discussion on scale-aggregated discrimination. As an additional diagnostic, we examine a natural alternative to our design: allowing the discriminator to observe all scale-specific images jointly, following prior work (Karnewar & Wang, 2020). This analysis addresses whether cross-scale consistency can be obtained simply by giving the discriminator access to the full image pyramid. We construct this scale-aggregated variant by removing the consistency regularization $\mathcal{L}_{cons}$ and the block-diagonal attention mask while keeping the other settings unchanged, using *Base*-sized models.

As shown in Fig. 8, the scale-aggregated variant performs substantially worse than its scale-wise counterpart. It also exhibits strong cross-scale interaction, with tokens from one resolution attending substantially to those from other resolutions. This suggests that the discriminator can rely on cross-scale evidence or compatibility rather than judging each scale through its own image-level realism, as noted in prior work (Kang et al., 2023; Lin et al., 2021). Thus, jointly exposing all

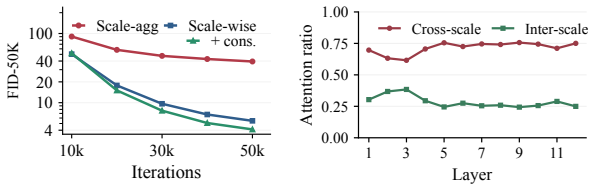
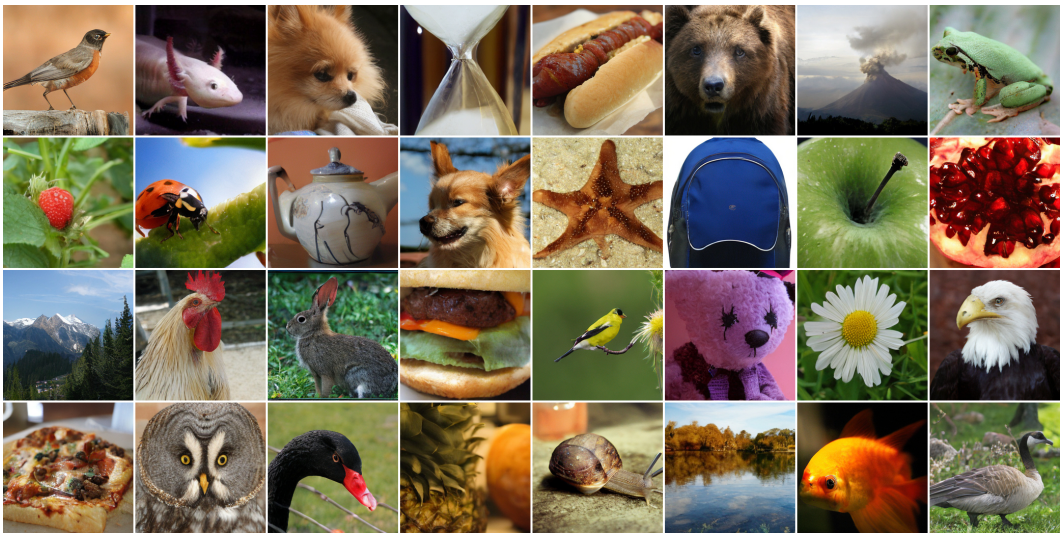


Figure 8: Analysis on scale-aggregated discriminator.

Model	FID-50K
G-M/2 / D-B/2	5.11
G-M/2 / D-M/2	**4.28**

Table 7: Discriminator scaling (12 epo).



(a) Generated examples from CAT-H/2 on ImageNet-256.



(b) class 90 to 140: lorikeet to red-backed sandpiper



(c) class 576 to 437: gondola to beacon

Figure 9: Generated examples and latent interpolation on ImageNet-256.

scales to the discriminator can entangle adversarial feedback across scales instead of preserving direct scale-wise supervision. This further supports our design of keeping the discriminator scale-wise while enforcing cross-scale consistency separately on the generator side. The consistency measurements in Appendix A.5 additionally show that scale-aggregated discrimination improves cross-scale consistency despite its worse FID.

Qualitative results. Fig. 9 shows samples generated by CAT-H/2 on ImageNet-256. The samples demonstrate that CAT produces diverse and high-fidelity images across different object categories with a single generator forward pass and also has semantic latent space by their interpolation. Additional uncurated samples and qualitative comparisons with iMF-XL/2 are provided in Appendix A.8.

6 Conclusion

We studied the limitations of standard scale-wise adversarial supervision in multi-stage GANs. While it provides direct per-scale realism feedback, our analysis shows that intermediate outputs do not necessarily remain consistent with the same final output, exhibiting large discrepancy and inter-stage rewriting with weak direction alignment. Based on this observation, we proposed CAT, which preserves scale-wise adversarial supervision while introducing generator-side consistency regularization to explicitly encourage cross-scale consistency. On ImageNet-256, CAT-H/2 achieves an FID-50K of 1.56 with a single generator evaluation after 60 training epochs. These results demonstrate that complementing per-scale realism with explicit cross-scale consistency provides an effective framework for scalable one-step adversarial generation.

AI USE STATEMENT

Generative AI tools were used to assist with writing experimental and visualization code and polishing the manuscript for clarity and readability. The core research ideas and methodology were developed by the authors. All AI-assisted code, analyses, and manuscript revisions were reviewed and verified by the authors. The authors take full responsibility for the final methodology, experiments, results, claims, code, and manuscript.

ETHICS STATEMENT

This work studies class-conditional image generation on datasets such as ImageNet, with a focus on improving training and generation efficiency. Although the experiments do not target potentially harmful applications, advances in generative image modeling may enable downstream misuse, including the creation of misleading content. Such risks should be considered when applying or deploying the proposed method.

REPRODUCIBILITY STATEMENT

All hyperparameter settings and training details are provided in the supplementary material. We also provide the code as supplementary material to facilitate reproduction of our results.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

Sangeek Hyun, MinKyu Lee, and Jae-Pil Heo. Scalable GANs with transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bTfHzEx9Xk.

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *International Conference on Learning Representations*, 2019.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10124–10134, 2023.

Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7799–7808, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14986–14996, 2021.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. In *International Conference on Machine Learning*, pp. 37959–37974. PMLR, 2025.

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Adversarial flow models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3GPPUHlJU9.

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

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WUK8JIeetF.

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. *Transactions on Machine Learning Research Journal*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. FACM: Flow-anchored consistency models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=k9BpW1c4in.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32211–32252, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16133–16142, 2023.

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 *The Thirteenth International Conference on Learning Representations*, 2025.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. Alphaflow: Understanding and improving meanflow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=adacb4JTIv.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

Table 8: **Model configurations and shared ImageNet-256 training hyperparameters.** The discriminator is Base-sized unless stated otherwise.

Configuration	Base/2	Medium/2	Huge/2
Depth	12	24	32
Hidden dimension	768	768	1280
Attention heads	12	12	16
Head dimension	64	64	80
Patch size	2×2	2×2	2×2
MLP ratio	4.0	4.0	4.0
Number of outputs	4	4	4
Output layers	(3,6,9,12)	(6,12,18,24)	(8,16,24,32)
Generator parameters	133M	261M	960M
Total G+D parameters	229M	357M	1056M

Hyperparameter	Value	Hyperparameter	Value
Image resolution	256×256	Batch size	512
VAE	SD-VAE	Optimizer	AdamW
D input resolutions	(32,16,8,4)	Adam (β_1, β_2)	(0.0,0.99)
G spatial tokens	256	Weight decay	0.0
D spatial tokens	256+64+16+4	Learning rate	2×10^{-4}
D class tokens	4	EMA decay	0.999
D depth / dim / heads	12 / 768 / 12	Mixed precision	bfloat16
REPA encoder	DINOv2-ViT-B	λ_{REPA}	1.0
λ_{cons}	0.1	(w_0, w_1, w_2)	$(1/3, 1/2, 1)$
R_1 weight / interval	1.0 / 1	R_2 weight / interval	1.0 / 1
Approximate GP ϵ	0.01		

Backbone and training details. We follow the implementation style of GAT (Hyun et al., 2026) and train all models in the SD-VAE latent space, where ImageNet-256 images are represented as $32 \times 32 \times 4$ latents. Both the generator and discriminator are implemented as ViT-style transformer networks. The generator starts from fixed 2D sinusoidal positional tokens and injects class and latent conditioning through a two-layer mapping network. Each generator block uses RMSNorm, multi-head self-attention with RoPE (Su et al., 2024) and qk-normalization, and a SwiGLU feed-forward network (Shazeer, 2020). We use AdaLN-style modulation in the generator, where the style vector predicts scale, shift, and residual gate parameters for both the attention and MLP branches. The generator produces four accumulated outputs from uniformly spaced transformer blocks through output skip connections. For quantitative evaluation, we do not use truncation trick, otherwise we denote it.

Unless otherwise specified, the discriminator is fixed to the Base configuration. We use the same transformer components, including RoPE, qk-normalization, and SwiGLU feed-forward layers. For the discriminator objective about representation alignment (Yu et al., 2025; Hyun et al., 2026), we use a frozen DINOv2-ViT-B encoder. All models are trained with AdamW, batch size 512, bfloat16 mixed precision, gradient clipping, and generator EMA for evaluation. We set the consistency weight to $\lambda_{\text{cons}} = 0.1$, apply R_1 and R_2 regularization at every iteration, and use an approximated gradient-penalty perturbation scale of 10^{-2}.

The detailed model configurations and shared training hyperparameters are summarized in Table 8.

Construction of multi-scale discriminator inputs. A key implementation detail is that the generator outputs are produced at the same latent resolution. Although we denote the hierarchical supervision by scale index k in the main text, the transformer generator maintains a fixed token grid across depth

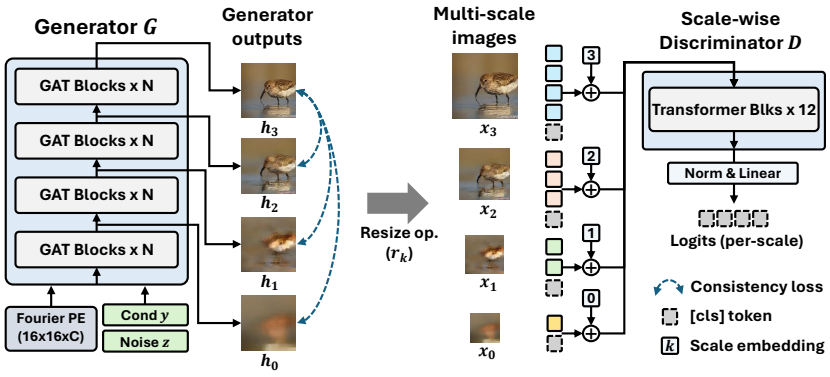


Figure 10: Overall pipeline of CAT. The generator produces accumulated intermediate outputs $\{h_k\}_{k=0}^{3}$ from uniformly spaced transformer blocks. Since the transformer generator operates on a fixed token grid, all h_k are synthesized at the same resolution. We construct multi-scale discriminator inputs by resizing each output with r_k, yielding $x_k = r_k(h_k)$. The discriminator receives the resulting multi-scale images with scale embeddings and per-scale [cls] tokens, while a block-diagonal attention mask enforces scale-wise discrimination without cross-scale token exchange. A generator-side consistency loss aligns the intermediate outputs, while the discriminator provides direct scale-wise adversarial feedback.

and therefore does not naturally produce different spatial resolutions. Let $h_k \in \mathbb{R}^{32\times32\times4}$ denote the k-th accumulated output produced by the generator. We then construct the scale-specific discriminator input by resizing h_k to the resolution associated with scale k:

$$x_k = r_k(h_k),$$

where $r_k(\cdot)$ denotes the resizing operator for the k-th scale. The real latent is resized in the same way to form the matching real input at each scale. Thus, the generator always synthesizes same-resolution latent outputs, while the multi-scale hierarchy used for adversarial supervision is constructed only when forming the discriminator inputs. This lets us control cross-scale interaction in the discriminator without changing the fixed-resolution transformer synthesis path.

Overall pipeline. Fig. 10 illustrates the overall pipeline of CAT. The generator G is divided into multiple stages, and each stage produces an accumulated output h_k through an output skip connection. Because the generator is transformer-based and maintains a fixed token grid, these intermediate outputs are generated at the identical resolution rather than at progressively different resolutions. To construct hierarchical adversarial supervision, we resize each h_k with a scale-specific resizing operator r_k and obtain the discriminator input

$$x_k = r_k(h_k).$$

The real latent is resized in the same way for the corresponding scale.

The discriminator D then receives the resulting multi-scale images $\{x_k\}$. For each scale, we append a separate [cls] token and add a learnable scale embedding. Although tokens from all scales are processed in a single discriminator for implementation efficiency, cross-scale attention is blocked with a block-diagonal attention mask, yielding a scale-wise discriminator in which each scale is discriminated using only same-scale tokens. This provides clean scale-wise adversarial feedback to the generator. At the same time, CAT applies a generator-side consistency loss to the intermediate outputs $\{h_k\}$, encouraging them to remain aligned with the final output. As a result, CAT combines scale-wise discrimination with explicit cross-scale alignment in the generator.

A.2 Details of GFLOPs computation

We provide details on how the training and inference compute in Table 5 is estimated. All numbers are analytical GFLOPs estimates. Inference compute is measured per generated sample. For GAN-based models, inference corresponds to a single generator forward pass, while for iMF it corresponds to

one evaluation-network path. Training compute is measured per sample per training iteration and includes both forward and backward computation. For training estimates, we follow the standard analytical convention that each trainable forward-backward computation costs approximately $3F$, while branches blocked by stop-gradient are counted only by their forward cost.

Table 9: **Analytical GFLOPs estimates.** Training compute is per sample and iteration; inference compute is per generated sample.

Method	Training GFLOPs	Inference GFLOPs
iMF-XL/2	1306.3	174.6
GAT-XL/2	2297.2	118.6
CAT-H/2	**1040.2**	166.7

Training computation cost. For GAN-based models, one training iteration consists of a discriminator step and a generator step. In the discriminator step, the generator first synthesizes fake samples without gradient update. The discriminator is then trained with a relativistic adversarial loss (Jolicoeur-Martineau, 2019) computed from real and fake logits, together with the approximated gradient-penalty terms. For CAT, we compute the approximated gradient penalty (Lin et al., 2025) using only a quarter of the batch, which reduces the discriminator-side training cost compared with a full-batch gradient-penalty computation. In the generator step, the generator synthesizes fake samples with gradients enabled, and the discriminator is used to provide the real/fake logits for the relativistic generator loss while its parameters are kept frozen. Thus, the discriminator still needs to propagate gradients to the generated samples, but not to its own parameters.

Let F_G and F_D denote one generator and discriminator forward pass, respectively. Under the above accounting rule, a CAT training iteration is approximated as

$$4F_G + 10.5F_D.$$

The generator term accounts for the no-gradient fake generation in the discriminator step and the trainable generator forward-backward computation in the generator step. The discriminator term accounts for adversarial real/fake discrimination, discriminator-side gradient penalties, and the frozen discriminator evaluation used for generator training. In the final reported numbers, we also include the small additional costs from discriminator-side auxiliary feature prediction and the frozen DINOv2 forward pass used for representation alignment.

The discriminator forward cost F_D of CAT-H/2 is much smaller than that of GAT-XL/2 because CAT keeps the discriminator at the *Base* scale while scaling the generator. This design is enabled by our scale-wise discriminator and generator-side consistency regularization: the discriminator can provide clean scale-wise adversarial feedback without requiring large cross-scale capacity, while consistency regularization maintains alignment across generated scales. As a result, CAT obtains strong scalability and generation quality with a relatively lightweight discriminator.

For GAT-XL/2, which scales both the generator and discriminator and uses a full-batch gradient-penalty setting, the corresponding estimate becomes

$$4F_G + 15F_D.$$

This leads to a substantially larger per-iteration training cost, despite its lower generator-only inference cost.

For iMF-XL/2, we use the code from official implementation[1] that contains velocity-guidance and meanflow training paths. During training, iMF first evaluates a velocity network for classifier-free guidance. This guidance branch consists of one doubled conditional/unconditional velocity evaluation and one additional conditional velocity evaluation, giving $3F_v$ in total, where F_v denotes one velocity-network forward pass. Because the guided velocity is blocked by stop-gradient before being used in the target, these guidance evaluations are counted as forward-only.

The average velocity branch then applies a JVP through the u network. Let F_u denote the forward cost of the u network used in this tangent path, and let F_{uv} denote the trainable joint path that returns

[1] https://github.com/Lyy-iiis/imeanflow

Table 10: **Preliminary pixel-space results.** We evaluate CAT in pixel space using Base-sized generator and discriminator backbones, with adapted patch sizes and a resolution hierarchy of $[16, 32, 64, 256]$. Following pMF, we additionally use a ConvNeXtV2-based vision-aided loss.

Method	Training epochs	FID-50K
pMF-B/16 (Lu et al., 2026)	160	3.53
CAT-B/16	40	3.54

the primal mean-flow output u together with the auxiliary velocity output v. The JVP tangent term is also stopped in the target construction, so this tangent path is counted as a forward-only F_u term. Reverse-mode gradients are counted only for the trainable primal u/v path, which is approximated as $3F_{uv}$. Thus, the iMF-XL/2 training step is estimated as

$$3F_v + F_u + 3F_{uv}.$$

The resulting call models and compute estimates are summarized in Table 9.

A.3 Preliminary Pixel-space Experiment

We further conduct a preliminary pixel-space experiment to examine whether the proposed supervision strategy is applicable beyond latent-space training. For this experiment, we use the same model configuration as our main setting, namely G-B/16 and D-B/16, while adapting the patch size for pixel-space training. We use a resolution hierarchy of $[16, 32, 64, 256]$ and keep the same token grid resolutions as in our latent-space model, i.e., $[2, 4, 8, 16]$. Accordingly, the patch sizes are set to $[8, 8, 8, 16]$ for the four resolutions, respectively. Following pMF (Lu et al., 2026), we additionally leverage a pretrained encoder, implemented as a ConvNeXtV2 (Woo et al., 2023)-based vision-aided loss.

Table 10 reports the results. Our method obtains an FID-50K of 3.54 after 40 epochs, which is comparable to the pMF baseline trained for 160 epochs. Although this experiment is less extensively tuned than our main latent-space setting, the result suggests that our method is compatible with pixel-space training and large-patch image-space generation. We leave more extensive tuning and a fully controlled pixel-space comparison to future work.

A.4 Unconditional generation on FFHQ-256

Table 11: **Unconditional generation on FFHQ-256.** All models are trained with a budget of 25M training images.

Model	FID↓
DiT-B/2	10.49
GAT-B/2	9.74
CAT-B/2 (Ours)	**7.76**

To evaluate CAT beyond class-conditional ImageNet generation, we conduct unconditional generation experiments on FFHQ at 256×256 resolution. We compare CAT-B/2 with GAT-B/2 and DiT-B/2, training all models with the same budget of 25M training images. CAT and GAT use the same generator architecture and nearly identical training configurations, differing only in multi-scale supervision and consistency regularization. As shown in Table 11, CAT achieves an FID of 7.76, improving over GAT (9.74) and DiT (10.49). These results demonstrate that CAT's benefits extend to other datasets beyond the class-conditional generation on ImageNet dataset. Figure 11 presents qualitative examples from CAT-B/2.

A.5 Consistency Regularization with Scale-Aggregated Discrimination

We further investigate whether generator-side consistency regularization also benefits scale-aggregated discrimination. Using the Base configuration on ImageNet-256, we compare scale-aggregated and



Figure 11: **Unconditional generation on FFHQ-256.** Images generated by CAT-B/2.

Table 12: **Consistency regularization with different discriminator designs.** We compare CAT-B/2 on ImageNet-256 after 12 training epochs. Cross-scale discrepancy, inter-stage rewriting, and alignment are averaged across scales.

Discrimination	$\mathcal{L}_{\text{cons}}$	Discrepancy↓	Rewriting↓	Alignment↑	FID↓
Scale-aggregated	No	0.537	0.436	0.773	47.25
Scale-aggregated	Yes	0.510	0.413	0.804	33.80
Scale-wise	No	0.883	0.813	0.476	9.60
Scale-wise	Yes	0.564	0.492	0.694	**7.61**

scale-wise discrimination with and without $\mathcal{L}_{\text{cons}}$ after 12 training epochs. Scale-aggregated discrimination allows the discriminator to use cross-scale relationships as cues for real/fake classification. Our regularizer provides an additional, explicit constraint that directly aligns intermediate generator outputs with the final output.

As shown in Table 12, adding $\mathcal{L}_{\text{cons}}$ to the scale-aggregated variant further reduces cross-scale discrepancy and inter-stage rewriting, increases alignment, and improves FID from 47.25 to 33.80. These results suggest that explicit generator-side alignment can complement the cross-scale relationships learned through joint discrimination, providing benefits beyond the scale-wise setting. Nevertheless, scale-wise discrimination achieves substantially better FID both with and without the regularizer. The best FID of 7.61 is obtained by combining scale-wise discrimination with $\mathcal{L}_{\text{cons}}$, supporting CAT's combination of direct scale-wise realism supervision and explicit cross-scale alignment.

A.6 Consistency Measurements in Perceptual and Semantic Spaces

Our original consistency diagnostics operate in latent space, where $\mathcal{L}_{\text{cons}}$ is also applied. Although these diagnostics quantify cross-stage alignment, improvements in the same representation space do not by themselves establish better perceptual or semantic consistency. We therefore extend our analysis to LPIPS (Zhang et al., 2018) and DINOv2-L (Oquab et al., 2024) feature spaces to examine whether the improvements are also reflected in these complementary representations.

For each generated sample, we decode the latent output of each generator stage into an RGB image using the VAE decoder. We then extract features from the decoded images using the pretrained LPIPS

network and DINOv2-L, and compute cross-scale discrepancy, inter-stage rewriting, and alignment in these feature spaces. Neither evaluation representation is used to define $\mathcal{L}_{\text{cons}}$. We evaluate B/2 at 20 epochs and M/2 and H/2 at 40 epochs, comparing models trained with and without $\mathcal{L}_{\text{cons}}$.

As shown in Table 13, consistency regularization reduces discrepancy and rewriting and increases alignment in both evaluation spaces across all three model sizes. These results provide complementary evidence that the improvements extend beyond latent-space alignment to perceptual and semantic consistency across generator stages.

Table 13: **Consistency measurements in perceptual and semantic spaces.** We compare scale-wise models trained with and without $\mathcal{L}_{\text{cons}}$ on ImageNet-256. Bold parenthesized values report relative changes from the corresponding model without consistency regularization.

Model	Epochs	$\mathcal{L}_{\text{cons}}$	Discrepancy↓	Rewriting↓	Alignment↑
LPIPS					
B/2	20	✗	0.5840	0.5479	0.6311
		✓	0.4936 **(–15.48%)**	0.4134 **(–24.55%)**	0.6803 **(+7.80%)**
M/2	40	✗	0.6212	0.5743	0.6075
		✓	0.4977 **(–19.88%)**	0.4267 **(–25.70%)**	0.6904 **(+13.65%)**
H/2	40	✗	0.7206	0.7013	0.5887
		✓	0.5635 **(–21.80%)**	0.4880 **(–30.42%)**	0.6479 **(+10.07%)**
DINOv2-L					
B/2	20	✗	0.7195	0.6224	0.5045
		✓	0.6431 **(–10.62%)**	0.5428 **(–12.79%)**	0.5672 **(+12.43%)**
M/2	40	✗	0.7537	0.6498	0.5012
		✓	0.6527 **(–13.40%)**	0.5558 **(–14.47%)**	0.5813 **(+15.98%)**
H/2	40	✗	0.7966	0.6845	0.4798
		✓	0.7179 **(–9.88%)**	0.6037 **(–11.81%)**	0.5363 **(+11.77%)**

A.7 Effect of Consistency Regularization on Recall

Enforcing cross-stage consistency could favor easily aligned samples, potentially reducing diversity. We therefore compare recall with and without $\mathcal{L}_{\text{cons}}$ under matched model sizes and training budgets. As shown in Table 14, recall remains nearly unchanged for B/2 and improves for M/2 and H/2, suggesting that consistency regularization does not systematically reduce coverage in these settings.

Table 14: **Effect of consistency regularization on recall.** We compare models of the same size trained for the same number of epochs on ImageNet-256. Δ Recall denotes the absolute change after adding $\mathcal{L}_{\text{cons}}$.

Model	Epochs	Recall↑		Δ Recall
		✗ $\mathcal{L}_{\text{cons}}$	✓ $\mathcal{L}_{\text{cons}}$	
B/2	20	0.5059	0.5024	**(–0.0035)**
M/2	40	0.5256	0.5494	**(+0.0238)**
H/2	40	0.4724	0.5493	**(+0.0769)**

A.8 QUALITATIVE COMPARISON WITH IMF.

We further compare the proposed CAT with iMF (Geng et al., 2026) using uncurated ImageNet-256 samples as belows. For sampling parameter, we use truncation $\psi = 0.85$ for CAT-H/2, while we follow the settings of iMF's the official implementation (interval $[0.42, 0.62]$ and $\omega = 8.0$).



(a) CAT-H/2 (Ours): class 277 (bee eater)
(b) iMF-XL/2: class 277 (bee eater)
(c) CAT-H/2 (Ours): class 349 (bighorn)
(d) iMF-XL/2: class 349 (bighorn)
(e) CAT-H/2 (Ours): class 933 (cheeseburger)
(f) iMF-XL/2: class 933 (cheeseburger)
(g) CAT-H/2 (Ours): class 84 (peacock)
(h) iMF-XL/2: class 84 (peacock)
(i) CAT-H/2 (Ours): class 270 (white wolf)
(j) iMF-XL/2: class 270 (white wolf)

Figure 12: **Uncurated sample comparison with iMF.** We compare randomly generated ImageNet-256 samples from our model (left) and iMF-XL/2 (right), without manual curation. Our model is trained for 60 epochs and uses 166.7 GFLOPs for one-step inference, whereas iMF-XL/2 is trained for 800 epochs and uses 174.6 GFLOPs. For sampling, our model uses truncation $\psi = 0.85$, and iMF-XL/2 follows the official configuration with interval $[0.42, 0.62]$ and $\omega = 8.0$.



(a) CAT-H/2 (Ours): class 562 (fountain)

(b) iMF-XL/2: class 562 (fountain)



(c) CAT-H/2 (Ours): class 17 (jay)



(d) iMF-XL/2: class 17 (jay)



(e) CAT-H/2 (Ours): class 107 (jellyfish)



(f) iMF-XL/2: class 107 (jellyfish)



(g) CAT-H/2 (Ours): class 145 (king penguin)



(h) iMF-XL/2: class 145 (king penguin)



(i) CAT-H/2 (Ours): class 323 (monarch)



(j) iMF-XL/2: class 323 (monarch)



(k) CAT-H/2 (Ours): class 35 (mud turtle)



(l) iMF-XL/2: class 35 (mud turtle)

Figure 13: **Uncurated sample comparison with iMF.** We compare randomly generated ImageNet-256 samples from our model (left) and iMF-XL/2 (right), without manual curation. CAT-H/2 requires 60 epochs training and 166.7 inference GFLOPs, while iMF-XL/2 needs 800 epochs and 174.6 GFLOPs. For sampling, our model uses truncation $\psi = 0.85$, and iMF-XL/2 follows the official configuration with interval $[0.42, 0.62]$ and $\omega = 8.0$.



(a) CAT-H/2 (Ours): class 9 (ostrich)

(b) iMF-XL/2: class 9 (ostrich)



(c) CAT-H/2 (Ours): class 963 (pizza)



(d) iMF-XL/2: class 963 (pizza)



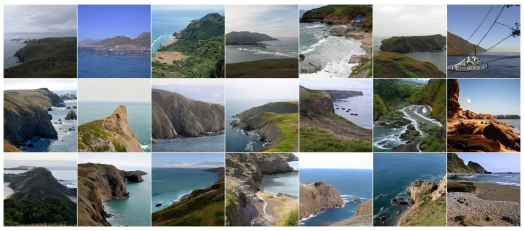
(e) CAT-H/2 (Ours): class 738 (pot)



(f) iMF-XL/2: class 738 (pot)



(g) CAT-H/2 (Ours): class 976 (promontory)



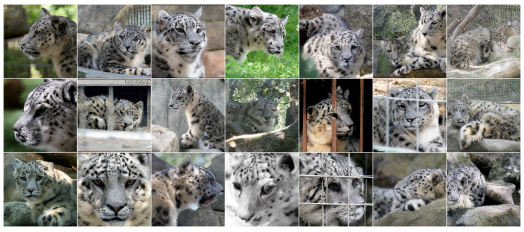
(h) iMF-XL/2: class 976 (promontory)



(i) CAT-H/2 (Ours): class 108 (sea anemone)



(j) iMF-XL/2: class 108 (sea anemone)



(k) CAT-H/2 (Ours): class 289 (snow leopard)



(l) iMF-XL/2: class 289 (snow leopard)

Figure 14: **Uncurated sample comparison with iMF.** We compare randomly generated ImageNet-256 samples from our model (left) and iMF-XL/2 (right), without manual curation. CAT-H/2 requires 60 epochs training and 166.7 inference GFLOPs, while iMF-XL/2 needs 800 epochs and 174.6 GFLOPs. For sampling, our model uses truncation $\psi = 0.85$, and iMF-XL/2 follows the official configuration with interval $[0.42, 0.62]$ and $\omega = 8.0$.