Supplementary

A. Discussion

A.1. Computational Cost.

GroupDiff introduces a longer effective sequence length compared to conventional diffusion baselines, yet the computational cost does not scale linearly with sequence length. In practice, for conventional diffusion models, attention projections and MLPs dominate, while attention calculation cost is smaller. Table 5 reports FLOPS for forwarding cost with group sizes 1 and 4.

Group Size	GFLOPS				
	Attn. Proj	Attn. Score	Attn. Total	MLP	Total
1 (Baseline)	152.5	34.4	186.9	304.4	492.3
4 (Ours)	152.5	137.6	290.2	304.4	595.5 (+21%)

Table 5. Model forward Flops. Applying Group Attention with group size 4 increase the FLOPS by 21%.

Inference cost. With same inference steps, GroupDiff-*l*-4 requires 95%–110% FLOPs of the baseline methods (95% for REPA-SiT and 110% for DiT). At inference time,CFG requires both conditional and unconditional score predictions at each denoising step. Consequently, GroupDiff-l-4 incurs $492.25 + 595.48$ GFLOPs, compared with $492.25 + 492.25$ GFLOPs for the standard DiT baseline, introducing only about 10% additional FLOPs per step. Besides evaluating with the same number of inference steps, we additionally perform a fair-FLOPs comparison (within $\pm 2\%$), as shown in Table 6.

DiT. The DiT baseline uses 250 inference steps without guidance interval. Thus, our GroupDiff matches inference FLOPs with 225 steps and achieves an 28% improvement.

REPA-SiT. The REPA baseline uses 250 steps with guidance interval $[0, 0.7]$, where the unconditional score is required when t falls within the guidance interval. Meanwhile, ours obtain the best performance with interval $[0.25, 0.75]$, corresponding to 265 steps for matched FLOPs, and improves performance by 18%.

Overall, after strictly controlling inference FLOPs, GroupDiff consistently outperforms both baselines, confirming that the gains arise from learned cross-image correspondence rather than increased computation.

Setting	FID (DiT)	FID (REPA-SiT)
Baseline	2.27	1.42
Ours w/ same Steps (in paper)*	1.55 (32%)	1.14 (17%)
Ours w/ closest FLOPS**	1.63 (28%)	1.13 (18%)

Table 6. Comparison on ImageNet 256×256. * uses 250 steps. ** uses 225 (DiT) and 265 (REPA-SiT) steps to match FLOPS .

Training cost. GroupDiff-*l*-4 increases training cost by only approximately 2% FLOPs per step. During training, conditional and unconditional models account for approximately 90% and 10% of the data, respectively. In GroupDiff-*l*, the conditional branch uses group size 1, while only the unconditional model adopts group size 4 ($\sim 20\%$ higher cost). As a result, GroupDiff-*l*-4 introduces only 2% additional training FLOPs per step, ensuring a fair comparison with the baseline.

Limitations. While GroupDiff demonstrates strong improvements in generation quality, further extending already long sequence lengths still introduces non-negligible computational overhead, especially in high-resolution settings (e.g., 4K image generation), where the token sequence becomes substantially larger. When the group size is n, GroupDiff-*f* and GroupDiff-*l* require approximately $(n-1)\times$ and $(0.1n)\times$ longer training time in every iteration, and $(n-1)\times$ and $0.5(n-1)\times$ longer inference time, respectively. Nevertheless, (a) this design opens a new avenue for exploring the trade-off between computational cost and generation quality, and (b) a high-quality model can serve as a teacher to distill faster and lighter students. We leave the study for a more efficient method for future exploration.

A.2. In-Group Similarity

We measure in-group and cross-group diversity using average LPIPS [70] among same-class samples as suggested. Higher LPIPS [70] indicates greater diversity. As shown in Table 7, in-group samples consistently exhibit higher diversity than cross-group samples, demonstrating that GroupDiff improves image quality while encouraging more diverse generation.

We evaluate different vision encoders and observe that higher similarity thresholds encourage stronger cross-sample correspondence. We also report an ablation on τ in We choose $\tau = 0.7$ as the default in Table 8, balancing generation quality and retrieval time (1.65 ms per query). Similarity search is implemented with *faiss*, where lower sim-

Group Size	Diversity (LPIPS) ↑	
	In-group	Cross-group
1	–	0.7231
2	0.7361	0.7177
4	0.7430	0.7246
8	0.7443	0.7338

Table 7. Diversity analysis.

τ	FID-50K ↓	Avg. Search Time ↓
0.5	2.65	13.32ms
0.6	2.53	3.28ms
0.7	2.42	1.65ms
0.8	2.41	1.66ms

Table 8. Ablation on τ.

ilarity thresholds return more candidates and incur higher retrieval overhead.

B. Implementation Details

B.1. Baselines

We introduce the baselines of the leading generative systems as follows:

- **ADM** [6] leverages classifier for guiding diffusion sampling to improve generation.
- **LDM** [43] presents latent diffusion, enabling fast, high-resolution generation by training diffusion models in a latent space.
- **MDTv2** [11] combines masked token modeling with diffusion transformers to learn visual representations.
- **VAR** [55] introduces next-scale prediction to autoregressive generative models.
- **LlamaGen** [49] shows vanilla autoregressive models could achieve strong generation performance at scale, outperforming diffusion baselines.
- **RandAR** [39] proposes a decoder-only autoregressive model that utilizes position instruction tokens to generate image tokens in arbitrary orders.
- **MaskDiT** [71] uses masked input patches and an asymmetric encoder-decoder to achieve faster diffusion model training.
- **DiT** [40] proposes a scalable transformer architecture based on AdaIN-zero for diffusion model training.
- **SiT** [31] further improves the efficiency and scalability on DiT by introducing flow matching.
- **REPA** [67] analyzes the alignment between feature quality and generation fidelity of diffusion backbone and accelerates diffusion model training by aligning diffusion feature with pre-trained vision encoders.
- **REPA-E** [26] enables representation learning inside diffusion backbones by unlocking the latent encoder.
- **DDT** [61] proposes a diffusion architecture that separates semantic encoding from high-frequency decoding to accelerate convergence during training.
- **SRA** [20] introduces a simple approach to align cross-layer diffusion backbone features to improve training efficiency without a pre-trained vision encoder.
- **Dispersive Loss** [59] introduces a simple regularization loss that encourages internal representations to disperse in the hidden space to improve diffusion model training.

B.2. Evaluation Metric

We use the conventional evaluation pipelines for class-conditional generative models, following ADM [6]. Specifically, we introduce the focusing concept of each metric:

- **Fréchet Inception Distance (FID)** [14] evaluates the feature distance of generated images and the reference samples. Lower FID usually suggests better generation fidelity and diversity.
- **Inception Score (IS)** [46] measures image quality and diversity based on how confidently a classifier recognizes each image and how varied the generated classes are. A higher Inception Score indicates a more meaningful image within each class.
- **Precision and recall** [25]. Precision captures the realism of generated images, while recall captures their diversity relative to real data.

B.3. Hyperparameter

In Table 9,we introduce the hyperparameter setting for models reported at Table 3.

C. Experiment

C.1. Ablations

GroupDiff-*f*: group size. We additionally investigate into the group size in GroupDiff-*f* setting. Figure 9 shows the which images shares the same group during inference. We compare the uncurated samples from GroupDiff-*f*-{1,2,3,4} in Figure 10 and Figure 11. Our observation on GroupDiff-*f* aligns that of GroupDiff-*l*, where increasing the group size considerably improves the generation fidelity.

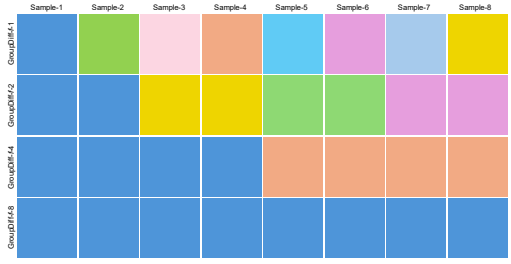

Figure 9. **Group attention illustration.** In each row, samples in the sample group shares the same color block.

GroupDiff-*l : query method**. Beyond training from scratch, resume from individual diffusion offer a efficient solution to adding GroupDiff over existing pipelines. Thus, we also explore different query method under this setting. Table 10 shows CLIP-L yields the optimality performance while the simplest GroupDiff-4* obtains a considerable improvement (14.5%) over the baseline, highlight the effectiveness of cross-sample attention.

	DiT-XL/2		SiT-XL/2		SiT-XL/2-Repa
	GroupDiff-4	GroupDiff-4*	GroupDiff-4	GroupDiff-4*	GroupDiff-4*
Architecture					
Input dim.	$32 \times 32 \times 4$	$32 \times 32 \times 4$	$32 \times 32 \times 4$	$32 \times 32 \times 4$	$32 \times 32 \times 4$
Num. layers	28	28	28	28	28
Hidden dim.	1,152	1,152	1,152	1,152	1,152
Num. heads	16	16	16	16	16
Optimization					
Resume	-	DiT-XL/2-7M	-	SiT-XL/2-7M	REPA-4M
Training Iteration	4M	500K	4M	500K	500K
Batch Size	256	256	256	256	256
Optimzier	AdamW	AdamW	AdamW	AdamW	AdamW
lr	0.0001	0.0001	0.0001	0.0001	0.0001
betas	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)	(0.9, 0.999)
weight decay	0.01	0.01	0.01	0.01	0.01
GroupDiff					
Mode	GroupDiff-*l*	GroupDiff-*l*	GroupDiff-*l*	GroupDiff-*l*	GroupDiff-*l*
Query Method	CLIP-L	CLIP-L	CLIP-L	CLIP-L	CLIP-L
τ_{img}	0.7	0.7	0.7	0.7	0.7
Group Size	4	4	4	4	4
Noise Var.	50	50	50	50	0
Inference					
Steps	250	250	250	250	250
Guidance Scale	1.70	1.60	2.35	1.85	2.45
Guidance Interval	(0,1)	(0,1)	(0.25,1.0)	(0.15,1.0)	(0.25,1.0)

Table 9. **Hyperparameter setup.**

Method	Query Method	FID ↓	IS ↑	Pre.↑	Rec.↑
SiT-XL/2	-	2.06	270.3	0.82	0.59
+ GroupDiff-4*	Class	1.76	283.5	0.81	0.61
+ GroupDiff-4*	CLIP-L	1.40	290.7	0.79	0.64

Table 10. **Ablation: query method**. *: continue training from pre-trained checkpoint for an additional 100 epochs.

C.2. Extending to Pixel Diffusion.

We further validate GroupDiff on pixel diffusion systems. As shown in Table 11, GroupDiff-4 with JiT-B/16 delivers a substantial 15.8% improvement with only 100 additional training steps when resumed from a pre-trained model. This again highlights the effectiveness of cross-sample collaboration in pixel diffusion and its strong potential for broader applicability.

C.3. Additional Qualitative Results.

We provide additional uncurated samples generated by GroupDiff-4 in Figures 14–26.

Method	params	FID	IS
ADM-G [6]	559M	7.72	172.7
RIN [19]	320M	3.95	216
SiD [74], UViT/2	2B	2.44	256.3
PixelFlow [3], XL/4	677M	1.98	282.1
PixNerd [60], XL/16	700M	2.15	297
JiT-H/16 [28]	953M	1.86	303.4
JiT-B/16 [28]	131M	3.66	275.1
+ **our** GroupDiff-4*	131M	3.08	245.6

Table 11. **System-level performance of pixel diffusion models** evaluated on ImageNet 256×256. *: continue training from pre-trained checkpoint for an additional 100 epochs.

Figure 10. **Uncurated generation results of GroupDiff-*f* without classifier-free guidance.** Examples of class-conditional generation on ImageNet 256×256. GroupDiff with a larger group size consistently obtains better generation fidelity.

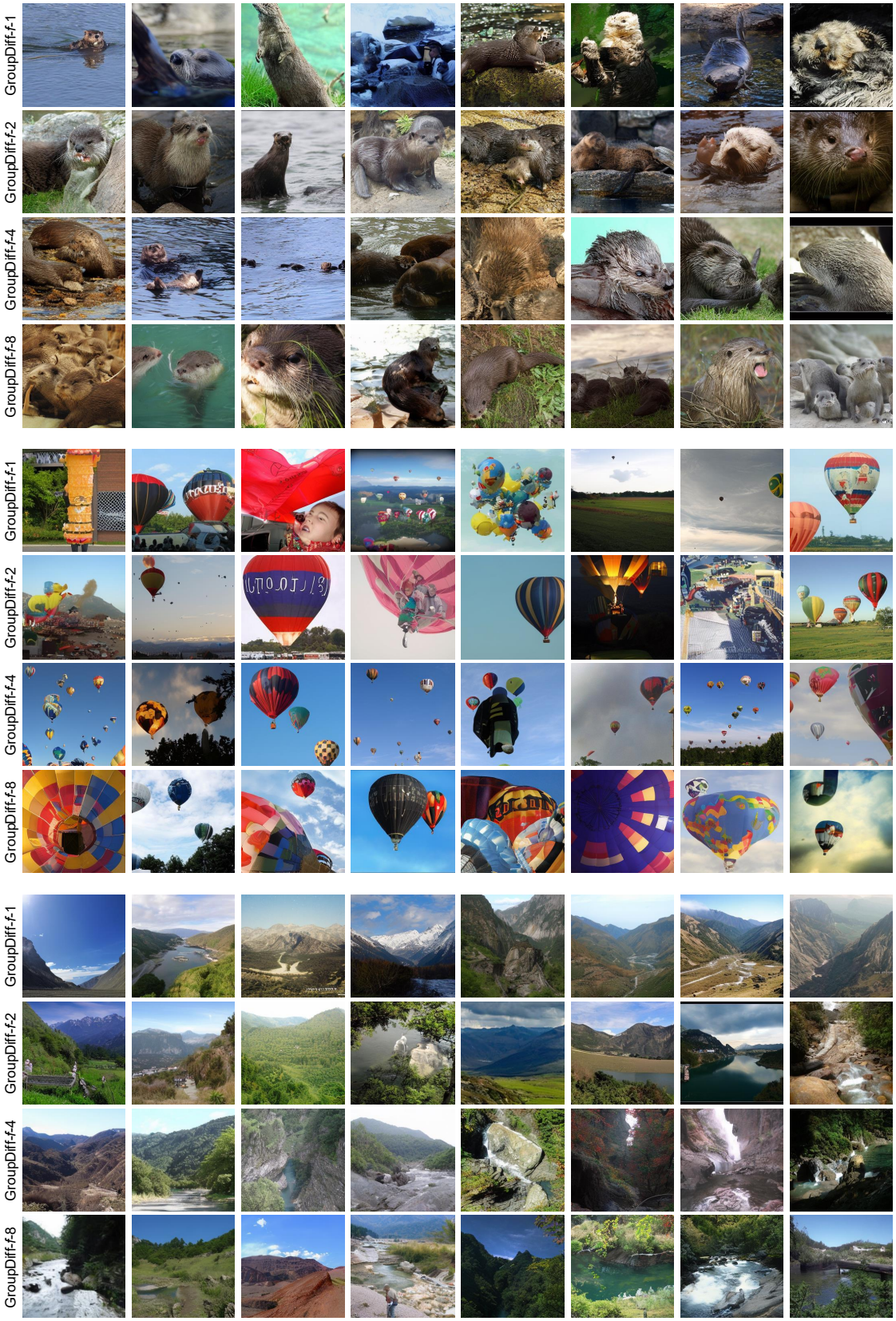

Figure 11. **Uncurated generation results of GroupDiff-*f* without classifier-free guidance.** Examples of class-conditional generation on ImageNet 256×256.GroupDiff with a larger group size consistently obtains better generation fidelity.

Figure 12. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “loggerhead sea turtle” (33).

Figure 13. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “macaw” (88).

Figure 14. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “sulphur-crested cockatoo, Kakatoe galerita, Cacatua galerita” (89).

Figure 15. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “golden retriever” (207).

Figure 16. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “Siberian husky” (250).

Figure 17. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “white wolf, Arctic wolf, Canis lupus tundrarum” (270).

Figure 18. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “Arctic fox, white fox, Alopex lagopus” (279).

Figure 19. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “acoustic guitar” (402).

Figure 20. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “balloon” (417).

Figure 21. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “baseball” (429).

Figure 22. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “fire engine, fire truck” (555).

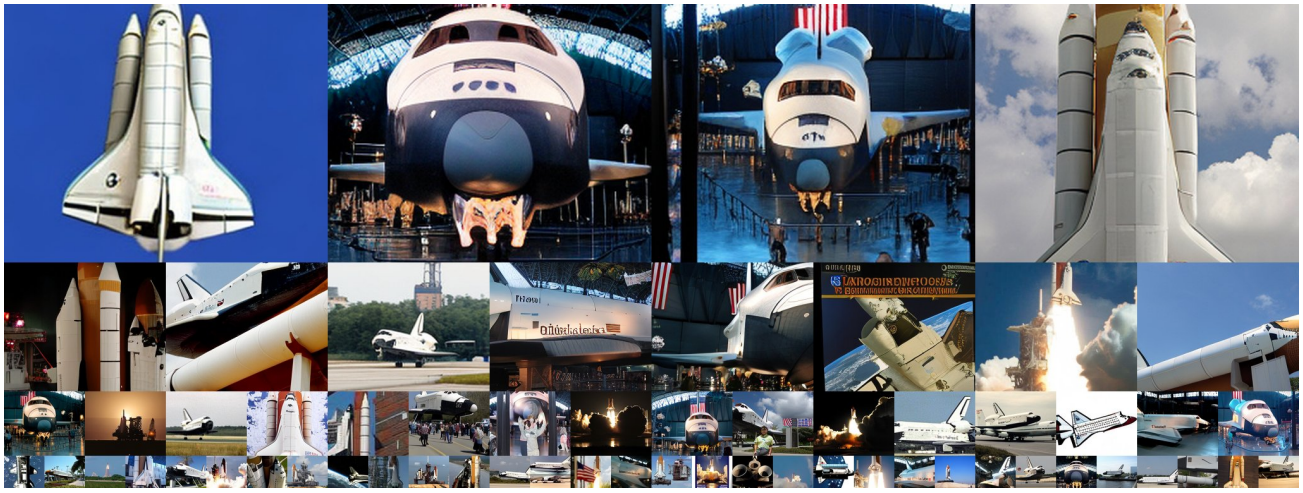

Figure 23. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “space shuttle” (812).

Figure 24. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “cheeseburger” (933).

Figure 25. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “coral reef” (973).

Figure 26. **Uncurated generation results of GroupDiff-*4***. We use classifier-free guidance with w= 3.5. Class label = “volcano” (980).

Method	Type	FID
AttnGAN [64]	GAN	35.49
DM-GAN [76]	GAN	32.64
VQ-Diffusion [13]	Diffusion	19.75
DF-GAN [52]	GAN	19.32
XMC-GAN [68]	GAN	9.33
Frido [10]	Diffusion	8.97
LAFITE [75]	GAN	8.12
U-Net [1]	Diffusion	7.32
U-ViT-S/2 [1]	Diffusion	5.95
U-ViT/S/2 (Deep) [1]	Diffusion	5.45
MMDiT [9]	Diffusion	5.3
DiT-XL/2 w/ Cross-Attention [40]	Diffusion	6.95
+ **our** GroupDiff-4	Diffusion	6.65

Table 12. **Quantitative comparison** on text-to-image generation (MS-COCO).

C.4. Text-to-Image Generation

We also validate GroupDiff in text-to-image generation. We mostly follow the experimental setup used in U-ViT [1] unless otherwise specified: we train the model from scratch on a train split of the MS-COCO dataset and use a validation split for evaluation. We use DiT-XL/2 with Cross-Attention and train it for 150K iterations with a batch size of 256. We use the frozen CLIP text encoder to extract text prompts from captions. Table 12 shows that GroupDiff remains effective in the T2I generation setting without bells and whistles, highlighting the importance of applying cross-sample attention even with text conditions.