Extending One-Step Image Generation from Class Labels to Text via Discriminative Text Representation

Supplementary Material

1. Velocity Field Learning Challenges: Class-Label vs. Text Conditions

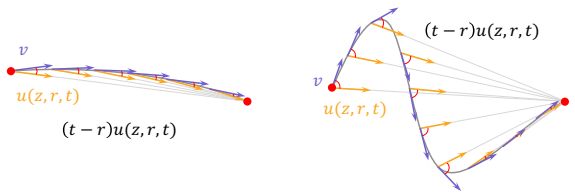

Figure 1. Denoising Trajectory Comparison. Simple class-label conditioning (left) yields a smooth path, whereas complex text conditioning (right) results in a tortuous path.

Extending MeanFlow from class-label conditioning to textual conditioning introduces fundamentally different challenges for velocity field learning.

Representation Separability. Class labels are discrete and well-separated in the embedding space, enabling the velocity field to maintain a stable direction. Consequently, the denoising trajectory is smooth, with the instantaneous velocity at each step closely aligning with the overall average velocity. This stability makes predicting the average velocity straightforward, ensuring high fidelity even in few-step generation. In contrast, textual embeddings form dense and continuous distributions where semantically related prompts (e.g. blue teapot vs. red teapot) occupy neighboring regions, reducing the *discriminability* of the representation. This density forces the velocity field to navigate fine-grained semantic distinctions, resulting in a more tortuous trajectory. The instantaneous velocity frequently diverges from the average, leading to semantic drift and necessitating additional corrective iterations to converge on the target concept.

Instruction Complexity. Class labels typically encapsulate a single semantic concept, whereas natural language prompts often bind multiple attributes, objects, and spatial relations (e.g., a blue ceramic teapot on a wooden table next to a vase of red tulips). In few-step regimes, the model has limited opportunities for correction. Therefore, inadequate *disentanglement* of these semantic components can easily lead to binding errors, missing objects, or incorrect attribute assignments.

The generation dynamics differ significantly between class-label and textual conditioning, a contrast visualized in Fig. 1. Under the simpler class-label conditioning, the denoising trajectory is relatively smooth. This smoothness indicates that the instantaneous velocity at each step closely aligns with the overall average velocity, making it straightforward for the model to predict this average. This stability is rooted in the embedding space, where class-label features form sparse clusters with large inter-class margins, ensuring category integrity and attribute accuracy even in single-step generation.

In stark contrast, the higher complexity and coupled nature of textual conditions lead to a more tortuous denoising trajectory. This winding path causes a significant divergence between the instantaneous and average velocities, often manifesting as early-stage semantic drift. Consequently, the model struggles to converge on the correct average velocity, necessitating additional corrective steps. This difficulty is exacerbated by the nature of textual embeddings, which reside in densely packed neighborhoods and inherently complicate the estimation of a stable average velocity.

These observations directly link the challenges of text-conditioned MeanFlow to the key properties of high-quality textual representations introduced in the main text: strong *discriminability* and *disentanglement* are essential for preserving semantic fidelity when the velocity field is learned under limited denoising steps.

2. Additional Experiment on text encoder

We conducted analyses of the post-trained SANA-1.5 and OpenUni text encoders, and ran mean-flow experiments on OpenUni. We chose OpenUni because it shares the SANA-1.5 diffusion backbone, but uses a InternVL3–based text encoder. Tab. 2 compares the two encoders. After training, Gemma becomes less discriminative but more disentangled, which helps 20-step generation by refining the language space. In contrast, mean-flow few-step generation needs strong image–text discriminability, so it still fails even after encoder training. We also train mean flow on OpenUni under the same setup (Tab. 3). OpenUni performs better than SANA-1.5, benefiting from stronger text-encoder representations, but it still falls short of the original model due to insufficient discriminability.

3. Inference Time Comparison.

When generating images from the same prompt and timing diffusion sampling only, BLIP3o-NEXT on H200 takes 1.24 s with 30 steps, while ours takes 0.22/0.12/0.08 s (4/2/1 steps). For end-to-end generation with different prompts,

Table 1. Quantitative evaluation results on DPG-Bench. Our method consistently outperforms distilled few-step models of comparable scale under the same denoising step settings.

Model	#Params	Steps	Global	Entity	Attribute	Relation	Other	Overall
Pretrained Models								
PixArt-α [1]	0.6B	20	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [2]	4B	20	82.82	88.65	86.44	80.53	81.82	74.63
Playground v2.5 [3]	/	/	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [4]	1.5B	50	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ [5]	0.6B	20	86.89	82.89	88.94	86.59	87.68	80.54
DALL-E 3 [6]	/	/	90.97	89.61	88.39	90.58	89.83	83.50
FLUX.1 [Dev] [7]	12.7B	50	74.35	90.00	88.96	90.87	88.33	83.84
SD3 Medium [8]	2B	50	87.90	91.01	88.83	80.70	88.68	84.08
HiDream-I1-Full [9]	3B	50	76.44	90.22	89.48	93.74	91.83	85.89
Lumina-Image 2.0 [10]	2.6B	50	-	91.97	90.20	94.85	-	87.20
Seedream 3.0 [11]	/	/	94.31	92.65	91.36	92.78	88.24	88.27
GPT Image 1 [High] [12]	/	/	88.89	88.94	89.84	92.63	90.96	85.15
Unified Models								
MetaQuery-L [13]	3B	30	-	-	-	-	-	81.10
BLIP3-o-8B [14]	8B	30	-	-	-	-	-	80.73
OpenUni-B-512 [15]	1.6B	20	85.87	87.33	86.54	86.91	89.43	80.29
Tar-7B [16]	9.6B	50	-	88.62	88.05	93.98	-	84.19
TBAC-UniImage-3B [17]	4.6B	30	83.52	87.94	87.80	87.17	87.02	80.97
Qwen-Image [18]	20B	50	91.32	91.56	92.02	94.31	92.73	88.32
Distilled Models								
SDXL-DMD2 [19]	2.6B	4	81.16	80.68	82.47	83.52	80.05	74.24
SD3.5-L-Turbo [8]	8B	4	90.99	87.43	87.42	87.81	86.10	81.97
SD3.5-Turbo [20]	8B	4	80.12	86.13	84.73	91.86	78.29	79.03
FLUX.1-schnell [7]	12B	4	86.62	90.82	88.35	93.45	82.00	84.94
SANA-Sprint [21]	1.6B	4	83.84	88.54	88.50	87.40	86.41	81.08
BLIP3o-NEXT and Ours under Few-Step Generation								
BLIP3o-NEXT	3B	1	73.60	69.10	73.48	79.92	69.09	57.05
	3B	2	82.32	79.35	79.16	77.71	79.66	67.38
	3B	4	88.53	85.99	85.04	87.78	86.44	78.15
	3B	30	86.21	88.55	86.82	90.14	88.01	82.05
EMF	3B	1	85.24	85.85	85.19	82.37	82.65	77.36
	3B	2	85.63	88.15	85.96	85.69	86.20	79.44
	3B	4	88.01	87.27	88.24	88.78	87.68	81.20
	3B	8	89.07	88.13	88.96	87.49	86.34	81.94

Table 2. Experiments on discriminability and disentanglement metrics for the trained SANA-1.5 and OpenUni text encoders.

Metric	Value
Disc. (Gemma-train)	0.694
Disc. (OpenUni)	0.724
Dise. (Gemma-train)	0.997
Dise. (OpenUni)	0.996

Table 3. Results of OpenUni trained on Mean Flow.

Steps	FM-GenEval	MF-Geneval
20	0.86	0.76
4	0.73	0.70
2	0.31	0.61
1	0.11	0.59

BLIP3o-NEXT (30 steps) takes 11.3 s, whereas our 4-step version takes 9.87 s. The remaining time is mostly spent on autoregressive text-embedding generation.

4. User Study and ImageReward Result

Considering instruction-following ability, we conducted PickScore and a user study on 50 prompts (similar to Fig.1 in our manuscript). We recruited 20 users, who compared images generated by five models for each prompt and answered: "Which result best matches the prompt?"

Table 4. Performance comparison across different models.

Model	PickScore	User Study
SDXL-DMD2	0.14	0.09
SD3.5-L-Turbo	0.16	0.13
FLUX.1-schnell	0.17	0.12
SANA-Sprint	0.25	0.16
Ours	**0.28**	**0.49**

All models use 4-step generation, and both experiments show that our method performs better.

5. Additional Quantitative and Qualitative Results

We provide supplementary quantitative and qualitative evaluations to further validate the effectiveness of our approach under limited denoising steps.

DPG-Bench evaluation. Generating high-fidelity images from complex and detail-rich textual prompts in a limited number of denoising iterations is a highly challenging task. To assess our model's capability in this regime, we conduct extensive tests on DPG-Bench, which focuses on long-form prompts with intricate attribute bindings and spatial relationships. As reported in Tab. 1, our method consistently outperforms equally sized distilled few-step models under the same step setting, despite the inherent difficulty of the benchmark. Notably, with only 8 sampling steps, our model delivers performance on par with the BLIP3o-NEXT baseline using 30 steps, and even under the challenging *1-step* regime, it surpasses widely-used distilled models such as SDXL-DMD2 and Playground v2.5 in overall score.

Vertical comparison across sampling steps. We additionally present the few-step generation results of our MeanFlow adaptation under *1-step*, *2-step*, *4-step*, and *8-step* settings, comparing them with the BLIP3o-NEXT baseline trained with standard Flow Matching under the same sampling step configurations.

As shown in Fig. 2, our method achieves an effective trade-off between inference speed and output quality: whereas the Flow Matching baseline exhibits noticeable blurring and loss of fine details when the number of sampling steps is reduced, our MeanFlow sampling retains salient object structures and fine-grained textures even at extremely low step counts, producing visually coherent and semantically faithful images at a fraction of the baseline's inference time.

Horizontal few-step comparison. We also present side-by-side comparisons between our model and other few-step approaches at same 4-step settings (Fig. 3). These results highlight our model's ability to preserve fine-grained details and adhere to textual instructions more faithfully than existing distilled models, across a diverse set of challenging prompts.

Figure 2. Representative visual results on DPG-Bench. Compared to the blurred outputs of few-step Flow Matching (FM) inference, our MeanFlow (MF) approach produces relatively sharp images even with a single sampling step, and with 8 sampling steps achieves visual quality comparable to Flow Matching using 30 steps, demonstrating a favorable trade-off between generation speed and visual fidelity.

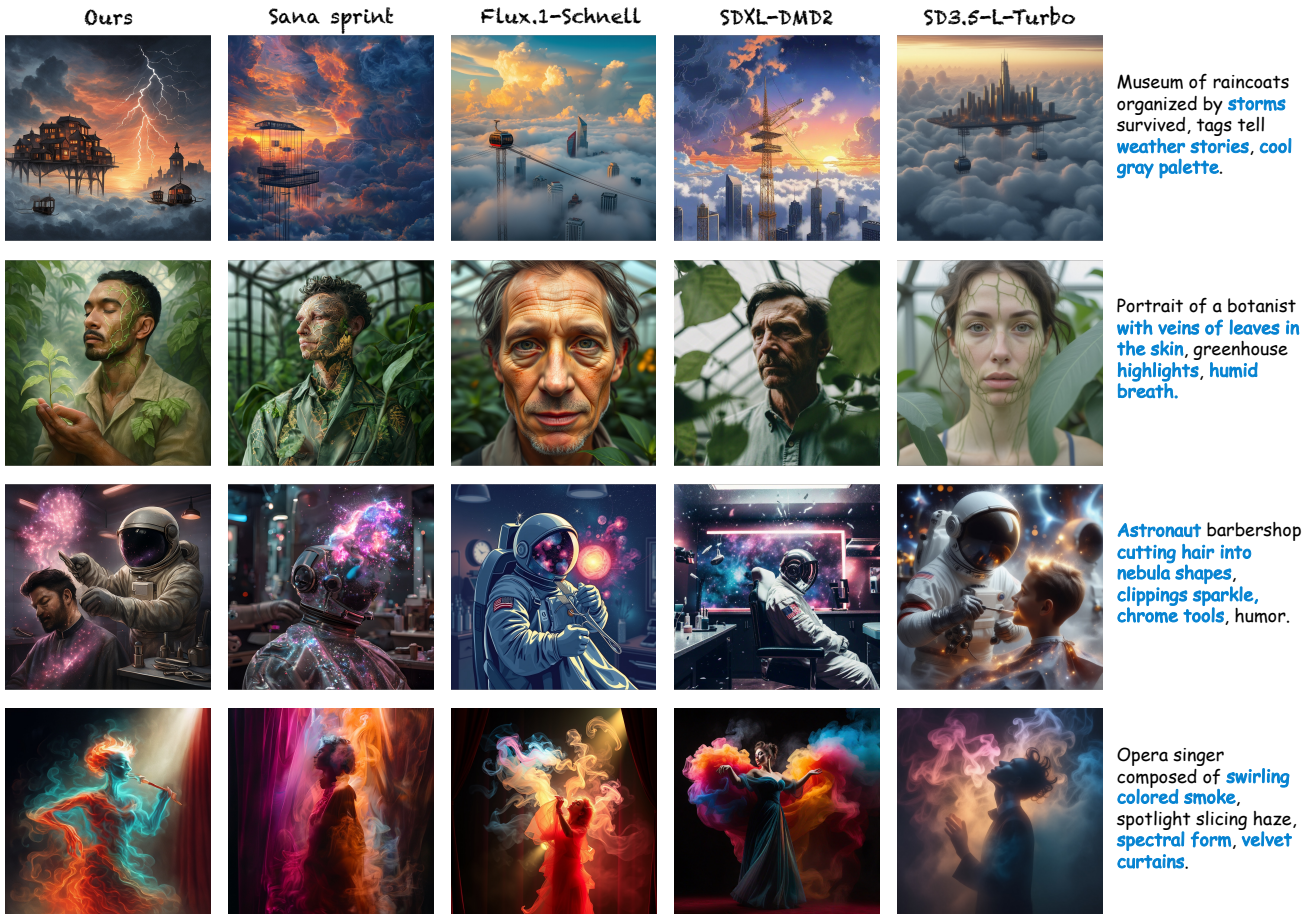

Figure 3. Additional comparisons under 4-step sampling between our method and existing distilled models. Our approach achieves higher semantic fidelity and richer visual details, closely adhering to complex text prompts. Blue text indicates cases where competing models fail to accurately render the described content.

References

[1] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 2

[2] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024. 2

[3] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 2

[4] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuandit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 2

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2

[6] OpenAI. DALL·E 3. https://openai.com/research/dall-e-3, September 2023. 2

[7] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[9] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. 2

[10] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. 2

[11] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025. 2

[12] OpenAI. Gpt-image-1, 2025. 2

[13] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 2

[14] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 2

[15] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025. 2

[16] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025. 2

[17] Junzhe Xu, Yuyang Yin, and Xi Chen. Tbac-uniimage: Unified understanding and generation by ladder-side diffusion tuning. *arXiv preprint arXiv:2508.08098*, 2025. 2

[18] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2

[19] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 2

[20] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2

[21] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. *arXiv preprint arXiv:2503.09641*, 2025. 2