ParaUni: Enhance Generation in Unified Multimodal Model with Reinforcement-driven Hierarchical Parallel Information Interaction

Supplementary Material

1. Details of Layer-wise Dynamic Adjustment Mechanism

In this section, we provide the implementation details of Layer-wise Dynamic Adjustment Mechanism(LDAM). We specify the choice of γ for the LDAM, which depends on the degree of reward decline and the degree of GradNorm increase. During the implementation process, we record the rewards and gradient norms of the most recent 5 epochs, and use the difference between the maximum and minimum values within this set as a guide for adjusting the value of γ, thereby adaptively adjusting the magnitude of perturbations., which is shown as follows:

$$G_{max} = max(g_{n-4}, g_{n-3}, g_{n-2}, g_{n-1}, g_n), \quad (1)$$
$$G_{min} = min(g_{n-4}, g_{n-3}, g_{n-2}, g_{n-1}, g_n), \quad (2)$$
$$R_{max} = max(r_{n-4}, r_{n-3}, r_{n-2}, r_{n-1}, r_n), \quad (3)$$
$$R_{min} = min(r_{n-4}, r_{n-3}, r_{n-2}, r_{n-1}, r_n), \quad (4)$$
$$\gamma(r, g) = \gamma_0(G_{max} - G_{min} + R_{max} - R_{min}), \quad (5)$$

where γ_0 is a control parameter set to 0.1. LDAM enhances reinforcement learning by detecting rewards and GradNorm in real-time and applying layer-wise perturbations when training gets stuck in local minima or unstable points, enabling the model to explore more possibilities and thus improving the overall performance.

2. Details of Training Recipe

In this section, we provide detailed experimental specifics in Tab. 1, covering Stages I to III. In Stage I, all image datasets are labeled by LLM. This forms a pre-trained corpus of approximately 23 million image-text pairs. In Stage II, the dataset contains 60,000 pairs of high-quality image-text pairs, generated by providing diverse descriptive prompts to GPT-4o and using models like DALL-E3 and Midjourney forimage synthesis. In Stage III, the dataset is inherited from FlowGRPO[4] with 50,000 prompts.

3. Training Comparison of Reinforcement Learning

In this section, we conduct comparative experiments to evaluate whether to use LDAM during the Reinforcement Learning(RL) phase. As shown in Fig. 1, Fig. 2 and Fig. 3, the red line represents the use of our LDAM, while the blue line represents the removal of LDAM, with all other settings remaining unchanged. We provide both the raw data and the data processed by Exponential Moving Average(EMA) with smoothing rate 0.1, including Pickscore, Aesthetic score, and CLIP score. As can be seen, the red line maintains a higher value throughout the RL process, demonstrating that using LDAM can effectively improve the reward score.

Table 1. Details of Training Recipe

Setting	Stage I	Stage II	Stage III
Diffusion Model	Frozen	Trainable	Trainable
Learning Rate	10^{-4}	10^{-5}	10^{-4}
Batch Size	512	512	9
Optimizer	AdamW		
Weight Decay	0.05		
Betas	(0.9, 0.95)		
Schedule	Cosine		
Training Steps	100,000	10,000	2,000
Warm-up Steps	1,000	100	0

4. Performance of Reconstruction

In this section, we test the image-to-image effect of ParaUni. Since our model is not trained on the editing task, we only test its image reconstruction effect. The input image is encoded by VAE and then fed into the VLM. This visual information was extracted by the learnable query and sent to the cross-attention of diffusion. The prompt was set to "Keep the image as it is." with no other inputs. We demonstrate the reconstruction effects of our method and the method using only the last layer features as conditions in Tab. 2. Note that our method was only train for text-to-image tasks. It can be seen that our method performs better in terms of CLIP feature similarity and PSNR, indicating that our method has great potential in detail recovery and semantic preservation.

Table 2. Performance of Reconstruction Task.

Setting	PSNR	SSIM	CLIP Similarity
Last layer interaction	13.5173	0.3640	0.9870
All layer interaction	13.6935	0.3722	0.9897

5. Performance of Understanding

We also present the performance of our method on the understanding tasks in Tab. 3. Since we freeze the parameters of the understanding module, our method inher-

its the performance of InternVL3 on these tasks. We report benchmarks include MMBench, SEED-Bench, MM-Vet, MMEPerception (MME-P), MMMU, RealWorldQA (RWQA) and TextVQA. Our approach is comparable in understanding performance to other methods.

6. More Qualitative Results

We provide more examples generated by ParaUni in Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8 and Fig. 9, which demonstrate that our method can produce a wide variety of vivid and realistic images, preserving both the details and semantics of the images, thus showcasing the superiority of our approach.

References

[1] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 3

[2] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025. 3

[3] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 3

[4] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. 1

[5] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. 3

[6] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024. 3

[7] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 3

[8] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 3

[9] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3

[10] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 3

[11] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[12] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 3

[13] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025. 3

[14] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 3

[15] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 3

[16] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 3

Table 3. Image understanding benchmark results. Since the parameters of MLLM are frozen, ParaUni still maintains excellent performance on the following benchmarks. We highlight the best results in **bold**.

Model	MMBench	SEED	MM-Vet	MME-P	MMMU	RWQA	TEXTVQA	POPE
EMU2 Chat [8]	-	62.8	48.5	-	34.1	-	66.6	-
Chameleon-7B [9]	19.8	27.2	8.3	202.7	22.4	39.0	0.0	-
Chameleon-34B [9]	32.7	-	9.7	604.5	38.8	39.2	0.0	-
Seed-X [3]	70.1	66.5	43.0	1457.0	35.6	-	-	-
VILA-U [14]	-	59.0	33.5	1401.8	-	46.6	48.3	85.8
LMFusion [7]	72.1	63.7	-	1603.7	41.7	60.0	-	-
Show-o-512 [15]	-	-	-	1097.2	26.7	-	-	73.8
EMU3 [11]	58.5	68.2	37.2	-	31.6	57.4	64.7	85.2
MetaMorph [10]	75.2	71.8	-	-	-	58.3	60.5	-
TokenFlow-XL [6]	76.8	72.6	48.2	1551.1	43.2	56.6	77.6	86.8
Janus-1.3B [12]	69.4	63.7	34.3	1338.0	30.5	-	-	87.0
Janus-Pro-7B [2]	79.2	72.1	50.0	1567.1	41.0	-	-	-
Harmon-0.5B [13]	59.8	62.5	-	1148.0	34.2	-	-	86.5
Harmon-1.5B [13]	65.5	67.1	-	1155.0	38.9	-	-	87.6
MetaQuery-B [5]	58.5	66.6	29.1	1238.0	31.4	-	-	-
MetaQuery-L [5]	78.6	73.8	63.2	1574.3	53.1	-	-	-
MetaQuery-XL [5]	**83.5**	76.9	**66.6**	**1685.2**	**58.6**	-	-	-
BLIP3-O-4B [1]	78.6	73.8	60.1	1527.7	46.6	60.4	78.0	-
BLIP3-O-8B [1]	**83.5**	**77.5**	**66.6**	1682.6	50.6	**69.0**	**83.1**	-
ParaUni [16])	81.1	64.6	62.2	1626.88	48.6	64.3	77.0	89.6

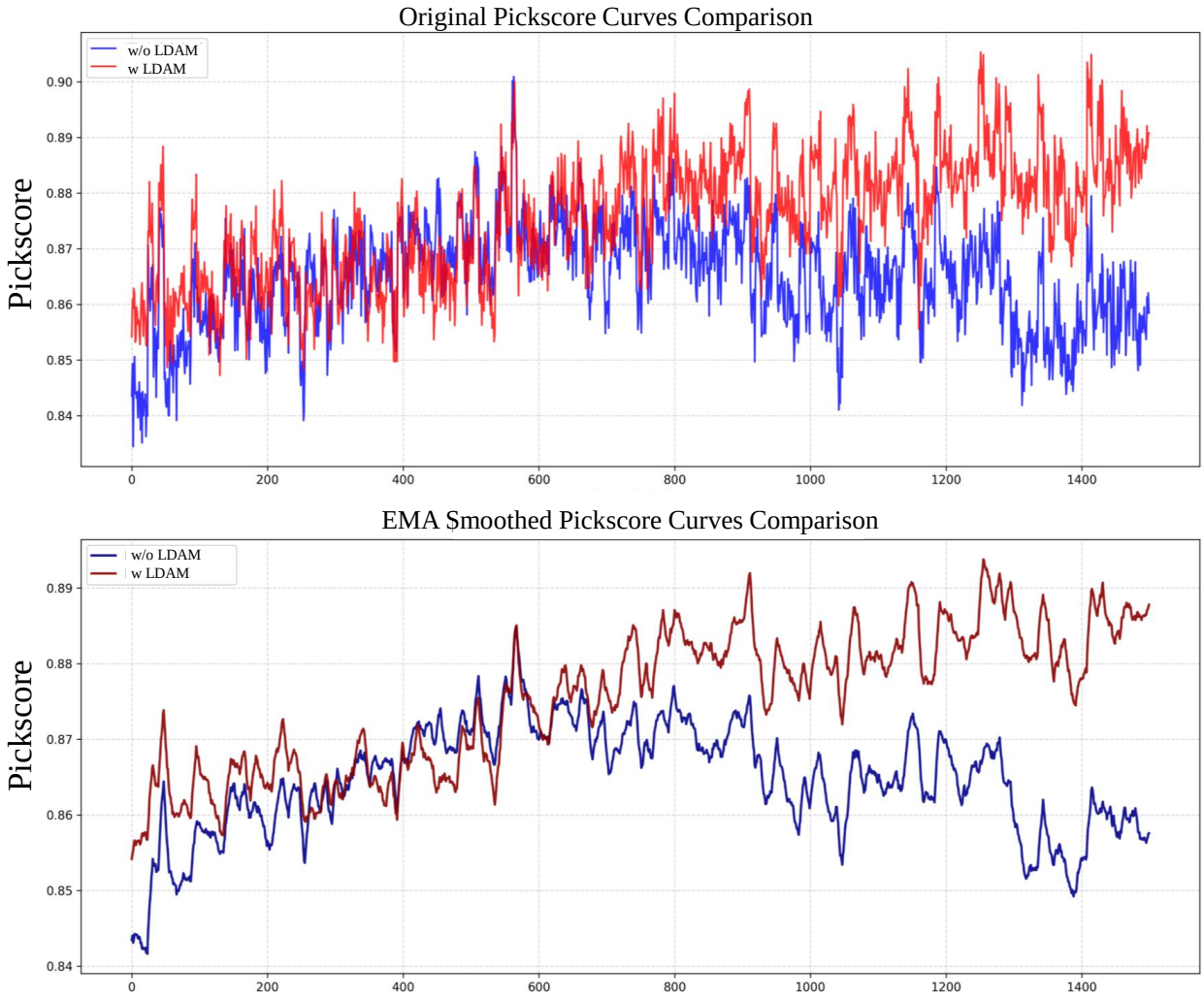


Figure 1. Pickscore comparison. Our approach(red line) outperforms the method that only uses the last layer(blue line).

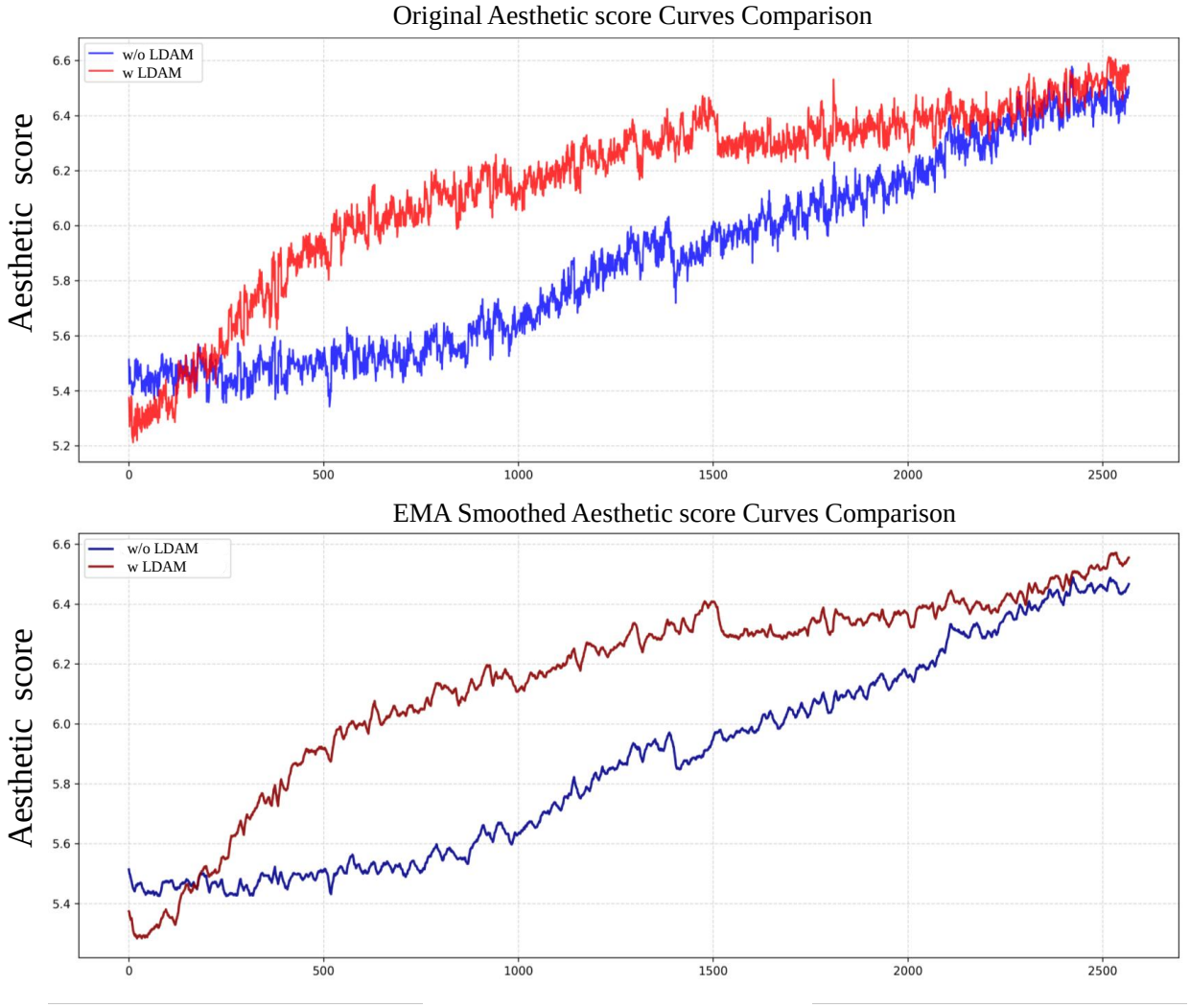


Figure 2. Aesthetic score comparison. Our approach(red line) outperforms the method that only uses the last layer(blue line).

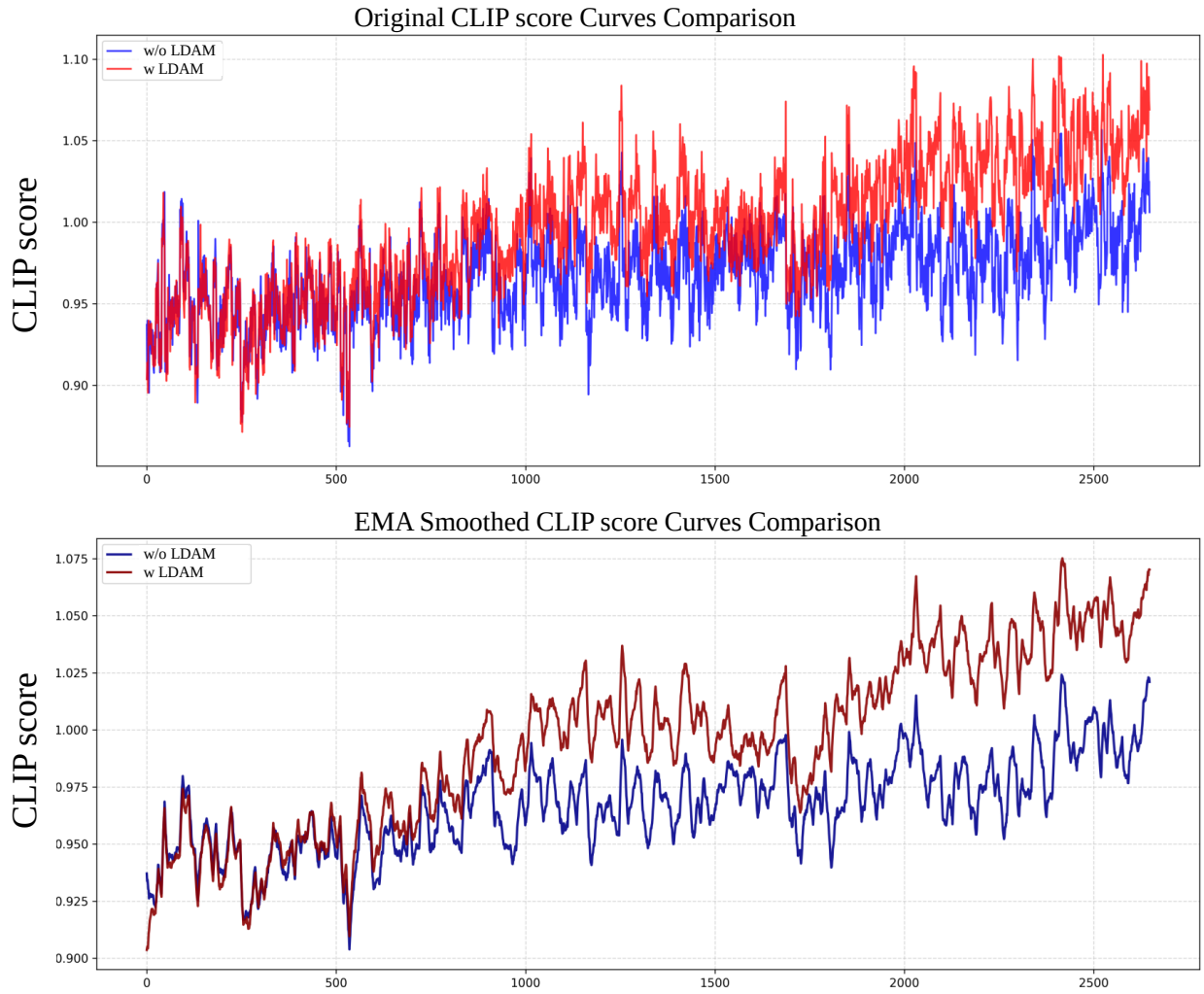
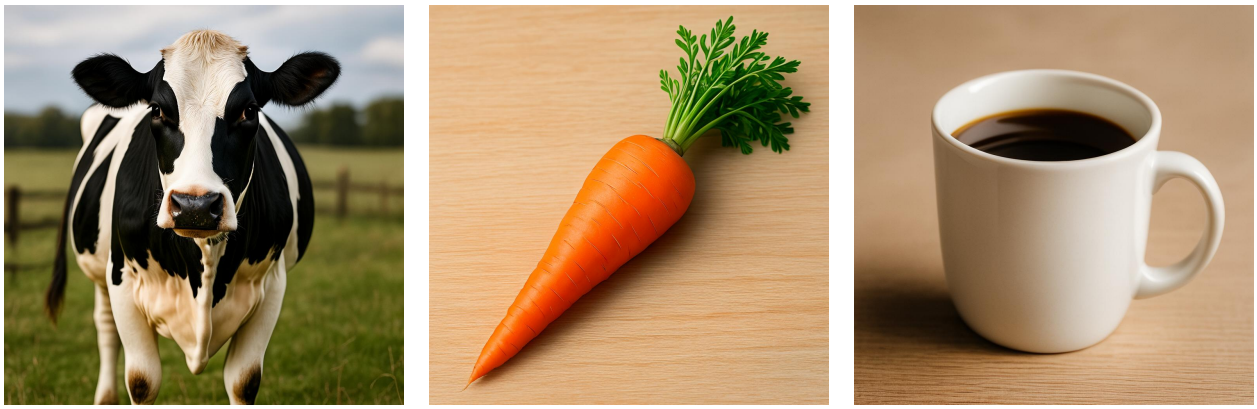


Figure 3. CLIP score comparison. Our approach(red line) outperforms the method that only uses the last layer(blue line).



a photo of a cow | a photo of a carrot | a photo of a cup

Figure 4. Qualitative Results from Geneval. Our approach can generate rich and vivid content.

a photo of a backpack

a photo of a person

a photo of a scissors and a bird

Figure 5. Qualitative Results from Geneval. Our approach can generate rich and vivid content.

a photo of two beds

a photo of a yellow airplane

a photo of a green microwave

Figure 6. Qualitative Results from Geneval. Our approach can generate rich and vivid content.

An ornate royal carriage, painted in deep red with golden trim, stands prominently against a landscape blanketed in pristine snow. Behind it, the silhouettes of tall pine trees dusted with white can be discerned through the soft haze of a winter's day. In front of the carriage, the snow-covered ground glistens under the subtle light of the afternoon sun.

In the midst of a vibrant garden, a cylindrical green cup stands alone on a stone path, its surface reflecting the bright afternoon sunlight. The cup, with a smooth finish, is surrounded by blossoming flowers and lush greenery. The shadows of nearby plants dance on the cup as gentle breezes sway their leaves.

A robust pigeon, with grey and white feathered plumage, sits comfortably on the sturdy branch of a venerable oak tree, replete with sprawling arms and knotted bark. Below, the mossy roots of the tree stretch out into the cobblestone paths of a charming village, where small, thatched-roof cottages neighbor each other. Sunlight dapples through the dense leaf canopy above, casting playful shadows on the scene below.

Figure 7. Qualitative Results from DPG-Bench. Our approach can generate rich content through long text.

A cluster of plump, purple grapes, their surface kissed by the morning's dew, reflects the soft, golden light of the rising sun. Each grape, tightly packed alongside its fellows, shows off a frosty sheen indicating the cool freshness of early day. They hang delicately from a green vine that's draped across a rustic, wooden trellis in a peaceful garden.

In a room bathed in the warm glow of the late afternoon sun, a single large golden camera sits prominently on a desk. This camera, with its polished metallic finish, outshines and is notably bigger than the two smaller silver monitors positioned on either side of it. The edges of the monitors are reflecting the soft light, creating a contrast with the camera's shining surface.

A rustic wooden table bathed in soft afternoon sunlight, showcasing a hearty, crusty loaf of freshly baked brown bread alongside a richly colored, firm purple eggplant. The textures of the bread's crackled crust juxtapose with the eggplant's smooth, glossy skin. Nearby, a folded linen napkin and an assortment of herbs suggest preparations for a savory meal.

Figure 8. Qualitative Results from DPG-Bench. Our approach can generate rich content through long text.

A woman dressed in a warm, stylish coat is seated at a small round table. She appears to be in a cozy indoor setting, possibly a café or a waiting area. The table is adorned with a vase containing a single flower, adding a touch of elegance to the scene.

A high-resolution portrait of the acclaimed actress Julianne Moore trending on Pinterest, captured by the lens of photographer Kyle Thompson. Her hair is styled in full platinum blonde waves that frame her intensely expressive, pale-skinned visage with high detail. The photograph, exuding realism and superior quality, showcases her piercing gaze that imparts a sense of subtle strength.

The image displays a vibrant array of multicolored flowers and lush green leaves clustered at the lower section. In the bottom right corner, a small, round, terracotta pot peeks into the frame, providing a contrast to the natural elements. The floral bouquet features petals ranging from deep purples to bright yellows, with a variety of leaf shapes and sizes nestled amongst them.

Figure 9. Qualitative Results from DPG-Bench. Our approach can generate rich content through long text.