Lumina-Image 2.0: A Unified and Efficient Image Generative Framework
Supplementary Material

A. Related Work

Recent advancements in text-to-image generation have been remarkable. Diffusion-based models have progressively transitioned from U-Net architectures [29] to Diffusion Transformers [28], as demonstrated by models such as PixArt [4, 5], FLUX [17], SD3 [9], Lumina [10, 47], and SANA [44]. These Diffusion Transformers exhibit superior scalability and are progressively evolving toward a unified multimodal representation [43]. Regarding text encoders, early approaches [32] employed CLIP [30], while subsequent works [9, 17, 18] additionally adopted T5-XXL [31]. More recently, SANA [44], Lumina [10, 47] and our Lumina-Image 2.0 have incorporated Gemma [36] as the text encoder. Furthermore, the latest models leverage flow-based parameterizations [19, 25], which enhance both training and inference efficiency compared to conventional diffusion methods. In parallel, a range of advanced autoregressive and hybrid text-to-image models have emerged [7, 12, 21, 35, 41, 45], achieving performance on par with their diffusion-based counterparts. However, the sampling speed of these autoregressive models remains significantly slower than that of diffusion-based approaches, posing a critical challenge for their practical deployment.

Meanwhile, the advancement of text-to-image models has been significantly shaped by the evolution of vision-language models (VLMs) [6, 8, 20, 24], where the quality of image captions plays a critical role in both model performance [3, 9]. Currently, the most commonly employed image captioners in text-to-image research include LLaVA [23], CogVLM [40], ShareGPT-4 [6], and Qwen-VL [1, 2, 39], all of which are general-purpose vision-language models (VLMs). However, there is a significant lack of research focused on developing captioner models specifically tailored for the text-to-image task, which may impede the further advancement of text-to-image models.

B. More Detailes for Efficient Inference

Flow-DPM-Solver (FDPM). Lumina-Next supports a range of ODE solvers, such as Midpoint and Euler method. While these solvers ensure stability, they are relatively slow since they are not designed for flow models, requiring a large number of function evaluations (NFE) for convergence. To improve this, we integrate FDPM [27, 44], which adapts DPM-Solver++ [27] to flow models, into Lumina-Image 2.0. FDPM achieves convergence in just 14-20 NFEs, providing a faster and more efficient solution. However, we find that FDPM sometimes suffers from poor stability in practice.

Timestep Embedding Aware Cache (TeaCache). TeaCache [22] is designed to selectively cache informative intermediate results during the inference, thereby accelerating diffusion models. TeaCache has successfully accelerated various mainstream image and video generation models, including FLUX [17], HunyuanVideo [16], as well as Lumina-Next. Building on its success, we integrate TeaCache into Lumina-Image 2.0. Similar to FDPM, our experiments show that TeaCache also introduces visual quality degradations when combined with the above techniques.

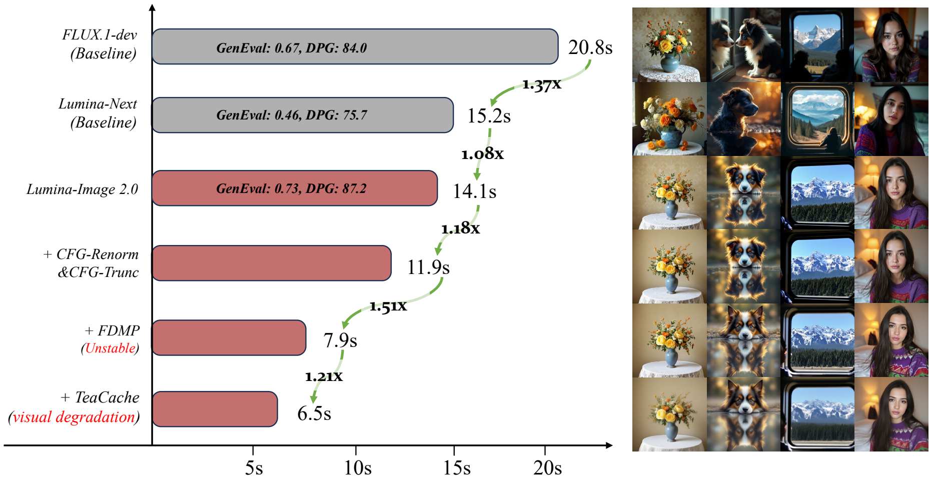


Figure 1. Ablation study on efficient inference strategy. The performances are measured on a single A100 GPU with batch size 1.

C. More Implement Details

Training Dataset. Following the methods in [4, 9, 15, 34, 41, 43, 45], we constructed a dataset combining both real and synthetic data, and performed data filtering based on the techniques outlined in [4, 16, 18], resulting in total 110M samples. This dataset is reorganized into three training phases, with 100M, 10M, and 1M samples used for each phase. As the dataset size decreased, the quality of the data progressively improved.

Architecture and Training Setups. The architecture configurations of our Unified Next-DiT model, along with a comparison to Lumina-Next [47], are summarized in Tab. 1. We employed 32 A100 GPUs across all three stages to optimize our Unified Next-DiT. The corresponding training configurations are detailed in Tab. 2. In addition, for multi-image generation task, we introduce an extra fine-tuning phase, where we consolidate different visual tasks into image grids and generate captions for these concatenated grids to form image-pair pairs.

Model	**Params**	**Patch Size**	**Dimension**	**Heads**	**KV Heads**	**Layers**	**RMSNorm ϵ [46]**	**Pos. Emb.**
Lumina-Next	1.7B	2	2304	16	8	24	$1e^{-5}$	2D-RoPE
Lumina-Image 2.0	2.6B	2	2304	24	8	26	$1e^{-5}$	M-RoPE

Table 1. Comparison of configuration between Lumina-Next and Lumina-Image 2.0.

Stage	**Image Resolution**	**#Images**	**Training Steps (K)**	**Batch Size**	**Learning Rate**	**GPU Days (A100)**	**Optimizer**
Low Res. Stage	256×256	100M	144	1024	2×10^{-4}	191	AdamW [26]
High Res. Stage	1024×1024	10M	40	512	2×10^{-4}	176	
HQ Tuning Stage	1024×1024	1M HQ	15	512	2×10^{-4}	224	

Table 2. Training configuration across different stages.

D. More Details for ELO Scores

Model	FLUX1.1 [pro] [17]	FLUX.1 [dev] [17]	**Lumina-Image 2.0**	Kolors [38]	HunyuanDiT [18]	Lumina-Next [47]
Score	0.4859	0.4712	0.4545	0.3924	0.3920	0.3229

Table 3. Comparison of ELO scores evaluated in text-to-image arena from AGI-Eval (as of February 23, 2025).

E. Prompt Template

Template A	You are an assistant designed to generate high-quality images based on user prompts. <Prompt Start> <Image Prompt>
Template B	You are an assistant designed to generate superior images with the superior degree of image-text alignment based on textual prompts or user prompts. <Prompt Start> <Image Prompt>
Template C	Generate a dual-panel image where the <lower half> displays a <depth map>, while the <upper half> retains the original image for direct visual comparison. <Prompt Start> <Image Prompt>

Table 4. Prompt template for Lumina-Image 2.0. <Image Prompt> will be replaced with the user specific image description. <lower half> and <upper half> will be replaced with the specific spatial relationships. <depth map> will be replaced with the target image type.

F. Experiments on Model Training with various Caption Length

During the training of Lumina-Image 2.0, we specifically observed that the length and quality of image captions have a significant impact on the model's convergence speed. As shown in Fig. 2, we train the model using three versions of image captions: (1) short captions generated by Florence, (2) short but precise captions generated by UniCap, and (3) long and detailed captions, also generated by UniCap. We observe that as captions became more precise and detailed, as well as longer and richer, the model's convergence speed significantly improved. This phenomenon motivates us to rethink the role of caption embeddings in the model capacity.

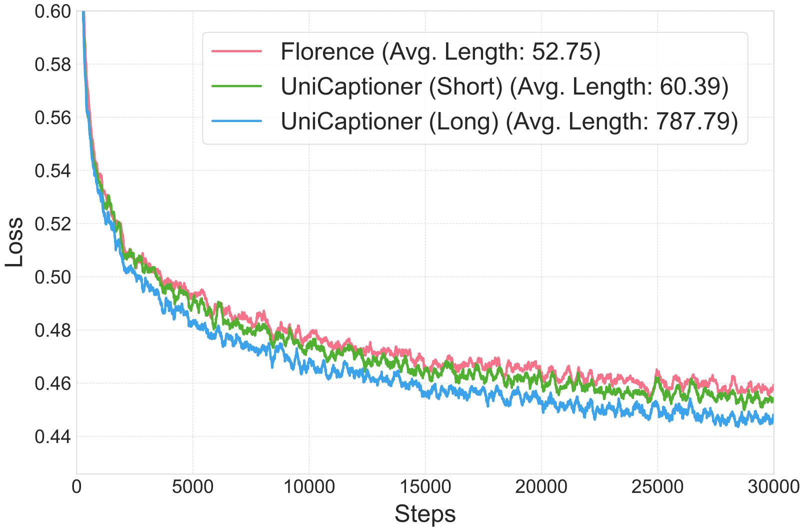


Figure 2. The training loss curve with respect to captions with different lengthes. The "Avg. Length" represents the average character number.

G. More Details for Multi-Stage Training Strategy

Stage	Steps (K)	DPG	GenEval
Low Res. Stage	15	84.5	0.63
High Res. Stage	38	85.7	0.67
HQ Tuning Stage	1	86.6	0.71
HQ Tuning Stage	5	87.2	0.73
HQ Tuning Stage	11	87.6	0.72

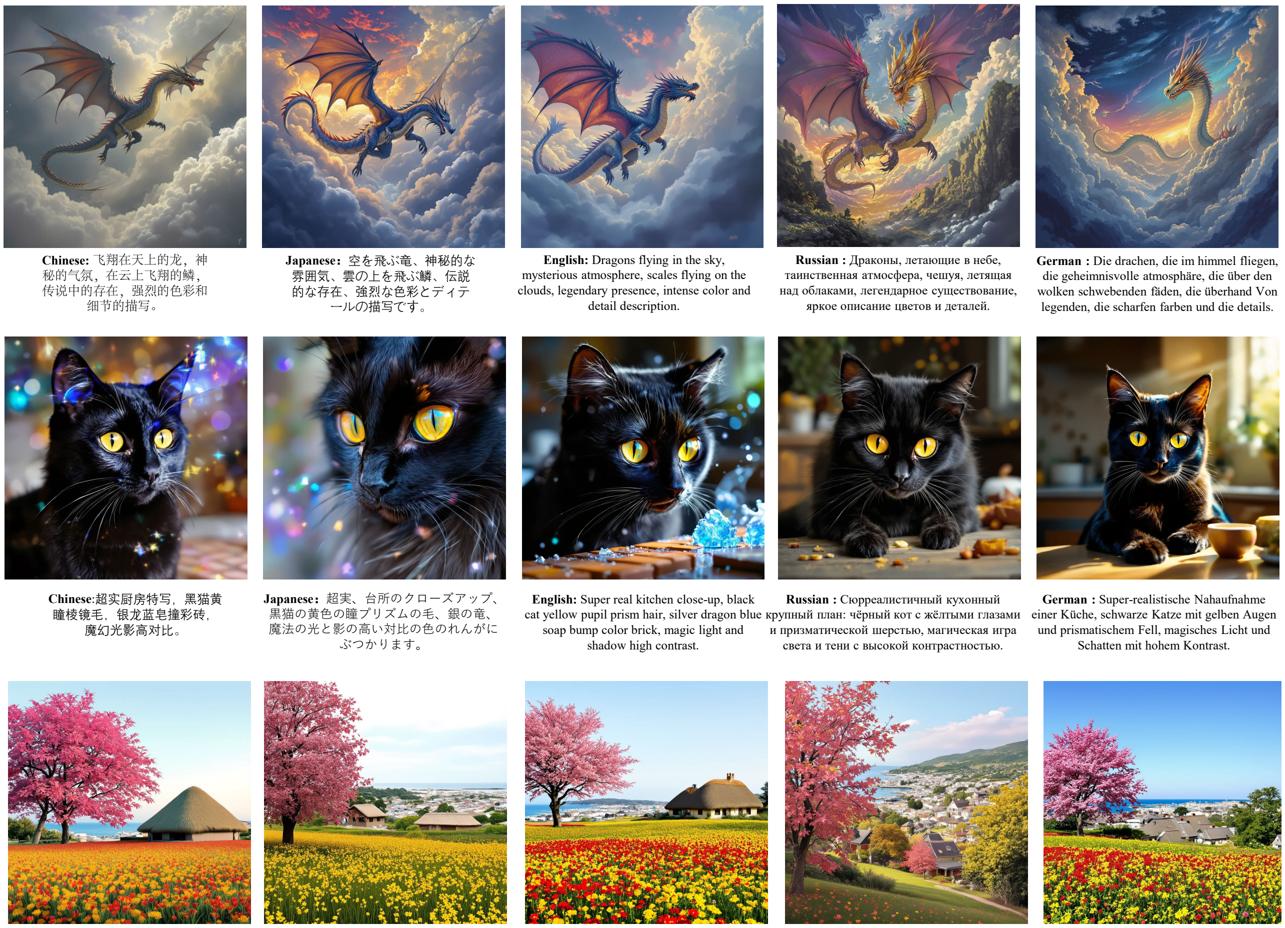
Table 5. Performance Comparison Across Stages on DPG [13] and GenEval [11] Benchmarks.

H. More Examples: Multi-lingual Generation

Compared to previous T2I models [4, 33] that use CLIP [30] and T5 [31] as text encoders, we employ Gemma2-2b [37] as the text encoder, enabling our model to understand multiple languages. It naturally exhibits zero-shot capability in languages such as German, Japanese, and Russian. As shown in Fig. 3, we present the generation results in five different languages.

I. More Examples: High-quality Image Generation

In Fig. 4, we present additional generation results of Lumina-Image 2.0. These results illustrate that our model is capable of producing images in various resolutions that are remarkably realistic, aesthetically refined, and creatively imaginative. Furthermore, extensive experiments with both Chinese and English prompts of different lengths demonstrate robust text-image alignment.



Chinese: 飞翔在天上的龙，神秘的气氛，在云上飞翔的鳞，传说中的存在，强烈的色彩和细节的描写。

Japanese: 空を飛ぶ竜、神秘的な雰囲気、雲の上を飛ぶ鱗、伝説的な存在、強烈な色彩とディテールの描写です。

English: Dragons flying in the sky, mysterious atmosphere, scales flying on the clouds, legendary presence, intense color and detail description.

Russian : Драконы, летающие в небе, таинственная атмосфера, чешуя, летящая над облаками, легендарное существование, яркое описание цветов и деталей.

German : Die drachen, die im himmel fliegen, die geheimnisvolle atmosphäre, die über den wolken schwebenden fäden, die überhand Von legenden, die scharfen farben und die details.

Chinese:超实厨房特写，黑猫黄瞳棱镜毛，银龙蓝皂撞彩砖，魔幻光影高对比。

Japanese: 超実、台所のクローズアップ、黒猫の黄色の瞳プリズムの毛、銀の竜、魔法の光と影の高い対比の色のれんがにぶつかります。

English: Super real kitchen close-up, black cat yellow pupil prism hair, silver dragon blue soap bump color brick, magic light and shadow high contrast.

Russian : Сюрреалистичный кухонный крупный план: чёрный кот с жёлтыми глазами и призматической шерстью, магическая игра света и тени с высокой контрастностью.

German : Super-realistische Nahaufnahme einer Küche, schwarze Katze mit gelben Augen und prismatischem Fell, magisches Licht und Schatten mit hohem Kontrast.

Chinese:红黄野花田野，左侧粉红叶大树，中央茅草屋顶的乡村建筑，背景是沿海小镇。柔和蓝天，平衡构图，宁静田园氛围。

Japanese: 赤と黄色の野花が咲く野原、左にピンクの葉の木、中央に茅葺き屋根の村、背景に沿岸の町。柔らかな青空、静かな田園の雰囲気。

English: Red and yellow wildflower field, pink-leaved tree, thatched cottages, coastal town in the background. Soft blue sky, serene atmosphere.

Russian : Поле с красными и желтыми цветами, дерево с розовыми листьями слева, деревенские дома и прибрежный город на фоне. Мягкое голубое небо, спокойная атмосфера.

German: Ein Feld mit roten und gelben Blumen, ein Baum mit rosa Blättern links, Reetdachhäuser und eine Küstenstadt im Hintergrund. Sanfter blauer Himmel, ruhige Atmosphäre.

Figure 3. Visualization results of multilingual text-to-image generation by our Lumina-Image 2.0, covering five languages: Chinese, Japanese, English, Russian, and German.

J. More Examples: Captioning Everything With UniCap

We evaluate the advantages of our proposed UniCap over existing captioners, such as ShareGPT4V [6] and Florence [42], from four dimensions: complex scenes, dense text, visual understanding, and spatial relationships. UniCap supports multilingual annotations, including both Chinese and English, and can generate captions of varying lengths to accommodate diverse user needs. As shown in the comparisons in Fig. 5 and Fig. 6, UniCap delivers highly detailed and accurate descriptions, significantly outperforming the other two methods.

K. Limitation

Although we have followed previous works [4, 7, 12, 43, 44] to evaluate our method on benchmarks such as GenEval [11] and T2ICompBench [14], achieving comparable performance with state-of-the-art models, we argue that these academic benchmarks are not comprehensive and may sometimes fail to accurately assess image quality in alignment with human perception. To illustrate this point, Fig. 7 highlights several limitations of Lumina-Image 2.0. First, for complex and diverse structures (e.g., human bodies) and for rare concepts in the training data (e.g., handguns), our model struggles to consistently generate correct results. Second, when handling images with intricate textures, such as densely crowded scenes or tire spokes, our model frequently generates disordered details. Finally, our model still needs substantial improvements in accurately rendering long and complex text.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 3, 4

[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. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 4, 8, 9

[7] 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. 1, 4

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[9] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1, 2

[10] Peng Gao, Le Zhuo, Chris Liu, , Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 1

[11] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 3, 4

[12] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 1, 4

[13] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 3

[14] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2

[17] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2023. 1, 2

[18] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1, 2

[19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[20] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 1

[21] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 1

[22] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024. 1

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1

[26] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 1

[28] William S Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2022. 1

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 1, 3

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020. 1, 3

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[34] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[35] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 1

[36] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 1

[37] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 3

[38] Kuaishou Technology. Kolors. `https://github.com/Kwai-Kolors/Kolors`, 2024. 2

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[40] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[41] 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. 1, 2

[42] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4, 8, 9

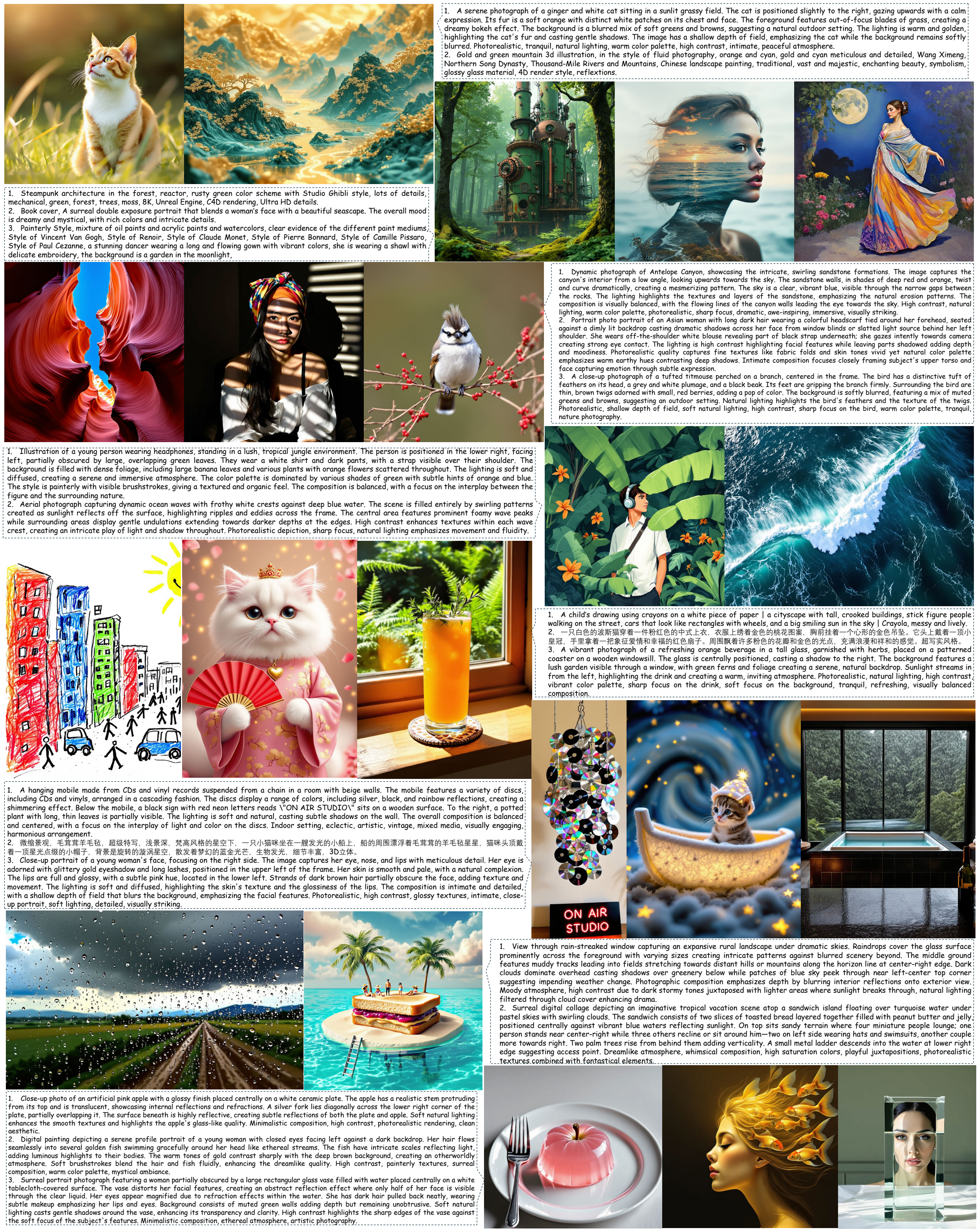
[43] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 4

[44] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 1, 4

[45] 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. 1, 2

[46] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[47] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2



1. A serene photograph of a ginger and white cat sitting in a sunlit grassy field. The cat is positioned slightly to the right, gazing upwards with a calm expression. Its fur is a soft orange with distinct white patches on its chest and face. The foreground features out-of-focus blades of grass, creating a dreamy bokeh effect. The background is a blurred mix of soft greens and browns, suggesting a natural outdoor setting. The lighting is warm and golden, highlighting the cat's fur and casting gentle shadows. The image has a shallow depth of field, emphasizing the cat while the background remains softly blurred. Photorealistic, tranquil, natural lighting, warm color palette, high contrast, intimate, peaceful atmosphere.
2. Gold and green mountain 3d illustration, in the style of fluid photography, orange and cyan, gold and cyan meticulous and detailed, Wang Ximeng, Northern Song Dynasty, Thousand-Mile Rivers and Mountains, Chinese landscape painting, traditional, vast and majestic, enchanting beauty, symbolism, glossy glass material, 4D render style, reflextions.

1. Steampunk architecture in the forest, reactor, rusty green color scheme with Studio Ghibli style, lots of details, mechanical, green, forest, trees, moss, 8K, Unreal Engine, C4D rendering, Ultra HD details.
2. Book cover, A surreal double exposure portrait that blends a woman’s face with a beautiful seascape. The overall mood is dreamy and mystical, with rich colors and intricate details.
3. Painterly Style, mixture of oil paints and acrylic paints and watercolors, clear evidence of the different paint mediums, Style of Vincent Van Gogh, Style of Renoir, Style of Claude Monet, Style of Pierre Bonnard, Style of Camille Pissaro, Style of Paul Cezanne, a stunning dancer wearing a long and flowing gown with vibrant colors, she is wearing a shawl with delicate embroidery, the background is a garden in the moonlight,

1. Dynamic photograph of Antelope Canyon, showcasing the intricate, swirling sandstone formations. The image captures the canyon's interior from a low angle, looking upwards towards the sky. The sandstone walls, in shades of deep red and orange, twist and curve dramatically, creating a mesmerizing pattern. The sky is a clear, vibrant blue, visible through the narrow gaps between the rocks. The lighting highlights the textures and layers of the sandstone, emphasizing the natural erosion patterns. The composition is visually balanced, with the flowing lines of the canyon walls leading the eye towards the sky. High contrast, natural lighting, warm color palette, photorealistic, sharp focus, dramatic, awe-inspiring, immersive, visually striking.
2. Portrait photo portrait of an Asian woman with long dark hair wearing a colorful headscarf tied around her forehead, seated against a dimly lit backdrop casting dramatic shadows across her face from window blinds or slatted light source behind her left shoulder. She wears off-the-shoulder white blouse revealing part of black strap underneath; she gazes intently towards camera creating strong eye contact. The lighting is high contrast highlighting facial features while leaving parts shadowed adding depth and moodiness. Photorealistic quality captures fine textures like fabric folds and skin tones vivid yet natural color palette emphasizes warm earthy hues contrasting deep shadows. Intimate composition focuses closely framing subject's upper torso and face capturing emotion through subtle expression.
3. A close-up photograph of a tufted titmouse perched on a branch, centered in the frame. The bird has a distinctive tuft of feathers on its head, a grey and white plumage, and a black beak. Its feet are gripping the branch firmly. Surrounding the bird are thin, brown twigs adorned with small, red berries, adding a pop of color. The background is softly blurred, featuring a mix of muted greens and browns, suggesting an outdoor setting. Natural lighting highlights the bird's feathers and the texture of the twigs. Photorealistic, shallow depth of field, soft natural lighting, high contrast, sharp focus on the bird, warm color palette, tranquil, nature photography.

1. Illustration of a young person wearing headphones, standing in a lush, tropical jungle environment. The person is positioned in the lower right, facing left, partially obscured by large, overlapping green leaves. They wear a white shirt and dark pants, with a strap visible over their shoulder. The background is filled with dense foliage, including large banana leaves and various plants with orange flowers scattered throughout. The lighting is soft and diffused, creating a serene and immersive atmosphere. The color palette is dominated by various shades of green with subtle hints of orange and blue. The style is painterly with visible brushstrokes, giving a textured and organic feel. The composition is balanced, with a focus on the interplay between the figure and the surrounding nature.
2. Aerial photograph capturing dynamic ocean waves with frothy white crests against deep blue water. The scene is filled entirely by swirling patterns created as sunlight reflects off the surface, highlighting ripples and eddies across the frame. The central area features prominent foamy wave peaks while surrounding areas display gentle undulations extending towards darker depths at the edges. High contrast enhances textures within each wave crest, creating an intricate play of light and shadow throughout. Photorealistic depiction, sharp focus, natural lighting emphasizes movement and fluidity.

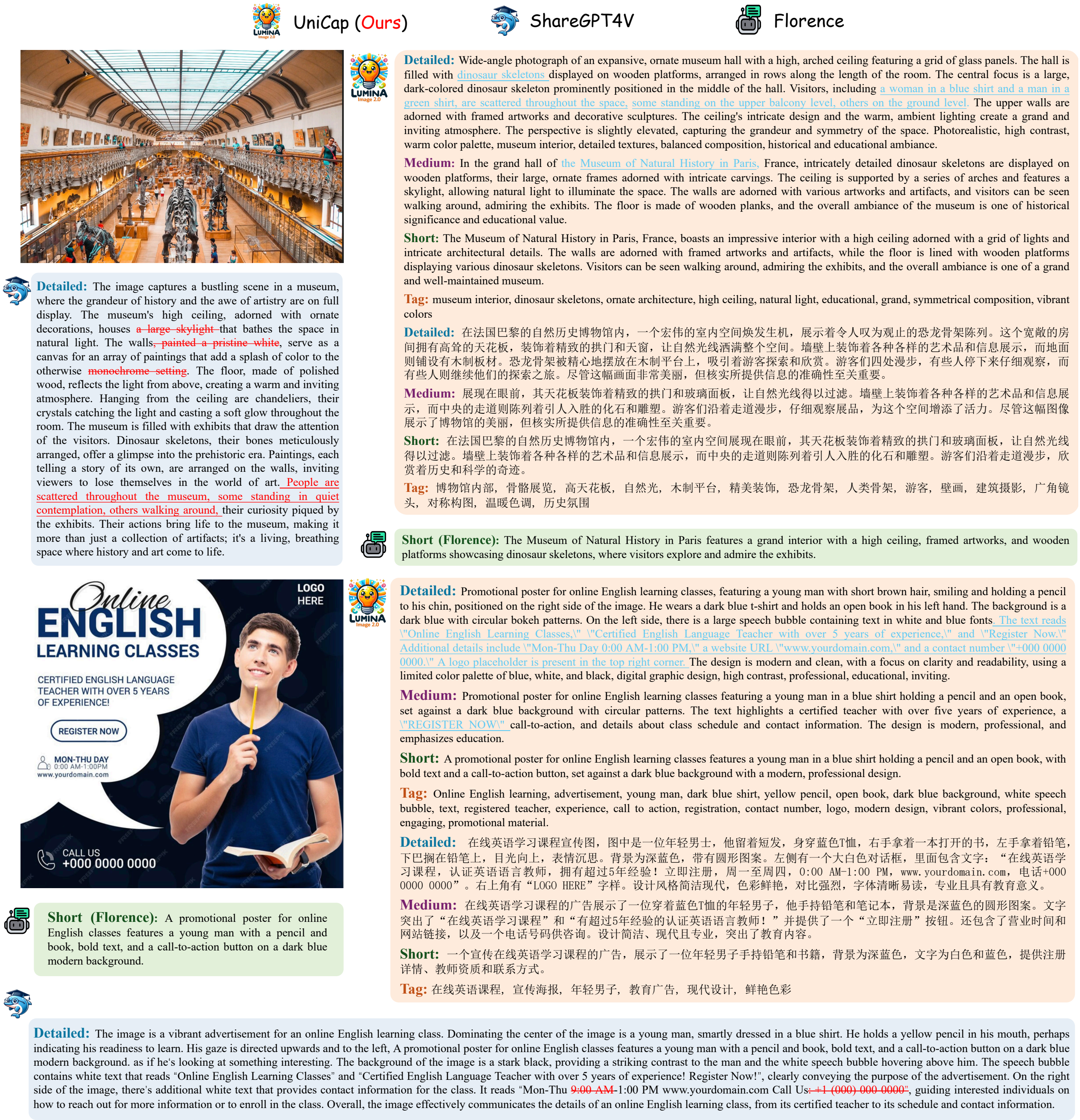
1. A child’s drawing using crayons on a white piece of paper | a cityscape with tall, crooked buildings, stick figure people walking on the street, cars that look like rectangles with wheels, and a big smiling sun in the sky | Crayola, messy and lively.
2. 一只白色的波斯猫穿着一件粉红色的中式上衣，衣服上绣着金色的桃花图案，胸前挂着一个心形的金色吊坠。它头上戴着一顶小皇冠，手里拿着一把象征爱情和幸福的红色扇子。周围飘着许多粉色的花瓣和金色的光点，充满浪漫和祥和的感觉。超写实风格。
3. A vibrant photograph of a refreshing orange beverage in a tall glass, garnished with herbs, placed on a patterned coaster on a wooden windowsill. The glass is centrally positioned, casting a shadow to the right. The background features a lush garden visible through a window, with green ferns and foliage creating a serene, natural backdrop. Sunlight streams in from the left, highlighting the drink and creating a warm, inviting atmosphere. Photorealistic, natural lighting, high contrast, vibrant color palette, sharp focus on the drink, soft focus on the background, tranquil, refreshing, visually balanced composition.

1. A hanging mobile made from CDs and vinyl records suspended from a chain in a room with beige walls. The mobile features a variety of discs, including CDs and vinyls, arranged in a cascading fashion. The discs display a range of colors, including silver, black, and rainbow reflections, creating a shimmering effect. Below the mobile, a black sign with red neon letters reads \"ON AIR STUDIO\" sits on a wooden surface. To the right, a potted plant with long, thin leaves is partially visible. The lighting is soft and natural, casting subtle shadows on the wall. The overall composition is balanced and centered, with a focus on the interplay of light and color on the discs. Indoor setting, eclectic, artistic, vintage, mixed media, visually engaging, harmonious arrangement.
2. 微缩景观，毛茸茸羊毛毡，超级特写，浅景深，梵高风格的星空下，一只小猫咪坐在一艘发光的小船上，船的周围漂浮着毛茸茸的羊毛毡星星，猫咪头顶戴着一顶星光点缀的小帽子，背景是旋转的漩涡星空，散发着梦幻的蓝金光芒，生物发光，细节丰富，3D立体。
3. Close-up portrait of a young woman's face, focusing on the right side. The image captures her eye, nose, and lips with meticulous detail. Her eye is adorned with glittery gold eyeshadow and long lashes, positioned in the upper left of the frame. Her skin is smooth and pale, with a natural complexion. The lips are full and glossy, with a subtle pink hue, located in the lower left. Strands of dark brown hair partially obscure the face, adding texture and movement. The lighting is soft and diffused, highlighting the skin's texture and the glossiness of the lips. The composition is intimate and detailed, with a shallow depth of field that blurs the background, emphasizing the facial features. Photorealistic, high contrast, glossy textures, intimate, close-up portrait, soft lighting, detailed, visually striking.

1. View through rain-streaked window capturing an expansive rural landscape under dramatic skies. Raindrops cover the glass surface prominently across the foreground with varying sizes creating intricate patterns against blurred scenery beyond. The middle ground features muddy tracks leading into fields stretching towards distant hills or mountains along the horizon line at center-right edge. Dark clouds dominate overhead casting shadows over greenery below while patches of blue sky peek through near left-center top corner suggesting impending weather change. Photographic composition emphasizes depth by blurring interior reflections onto exterior view. Moody atmosphere, high contrast due to dark stormy tones juxtaposed with lighter areas where sunlight breaks through, natural lighting filtered through cloud cover enhancing drama.
2. Surreal digital collage depicting an imaginative tropical vacation scene atop a sandwich island floating over turquoise water under pastel skies with swirling clouds. The sandwich consists of two slices of toasted bread layered together filled with peanut butter and jelly, positioned centrally against vibrant blue waters reflecting sunlight. On top sits sandy terrain where four miniature people lounge; one person stands near center-right while three others recline or sit around him—two on left side wearing hats and swimsuits, another couple more towards right. Two palm trees rise from behind them adding verticality. A small metal ladder descends into the water at lower right edge suggesting access point. Dreamlike atmosphere, whimsical composition, high saturation colors, playful juxtapositions, photorealistic textures combined with fantastical elements.

1. Close-up photo of an artificial pink apple with a glossy finish placed centrally on a white ceramic plate. The apple has a realistic stem protruding from its top and is translucent, showcasing internal reflections and refractions. A silver fork lies diagonally across the lower right corner of the plate, partially overlapping it. The surface beneath is highly reflective, creating subtle reflections of both the plate and apple. Soft natural lighting enhances the smooth textures and highlights the apple's glass-like quality. Minimalistic composition, high contrast, photorealistic rendering, clean aesthetic.
2. Digital painting depicting a serene profile portrait of a young woman with closed eyes facing left against a dark backdrop. Her hair flows seamlessly into several golden fish swimming gracefully around her head like ethereal streams. The fish have intricate scales reflecting light, adding luminous highlights to their bodies. The warm tones of gold contrast sharply with the deep brown background, creating an otherworldly atmosphere. Soft brushstrokes blend the hair and fish fluidly, enhancing the dreamlike quality. High contrast, painterly textures, surreal composition, warm color palette, mystical ambiance.
3. Surreal portrait photograph featuring a woman partially obscured by a large rectangular glass vase filled with water placed centrally on a white tablecloth-covered surface. The vase distorts her facial features, creating an abstract reflection effect where only half of her face is visible through the clear liquid. Her eyes appear magnified due to refraction effects within the water. She has dark hair pulled back neatly, wearing subtle makeup emphasizing her lips and eyes. Background consists of muted green walls adding depth but remaining unobtrusive. Soft natural lighting casts gentle shadows around the vase, enhancing its transparency and clarity. High contrast highlights the sharp edges of the vase against the soft focus of the subject's features. Minimalistic composition, ethereal atmosphere, artistic photography.

Figure 4. High-quality image generation examples from Lumina-Image 2.0, showcasing its precise prompt-following ability and its capability to generate highly aesthetic and realistic images across different resolutions.



Detailed: Wide-angle photograph of an expansive, ornate museum hall with a high, arched ceiling featuring a grid of glass panels. The hall is filled with dinosaur skeletons displayed on wooden platforms, arranged in rows along the length of the room. The central focus is a large, dark-colored dinosaur skeleton prominently positioned in the middle of the hall. Visitors, including a woman in a blue shirt and a man in a green shirt, are scattered throughout the space, some standing on the upper balcony level, others on the ground level. The upper walls are adorned with framed artworks and decorative sculptures. The ceiling's intricate design and the warm, ambient lighting create a grand and inviting atmosphere. The perspective is slightly elevated, capturing the grandeur and symmetry of the space. Photorealistic, high contrast, warm color palette, museum interior, detailed textures, balanced composition, historical and educational ambiance.

Medium: In the grand hall of the Museum of Natural History in Paris, France, intricately detailed dinosaur skeletons are displayed on wooden platforms, their large, ornate frames adorned with intricate carvings. The ceiling is supported by a series of arches and features a skylight, allowing natural light to illuminate the space. The walls are adorned with various artworks and artifacts, and visitors can be seen walking around, admiring the exhibits. The floor is made of wooden planks, and the overall ambiance of the museum is one of historical significance and educational value.

Short: The Museum of Natural History in Paris, France, boasts an impressive interior with a high ceiling adorned with a grid of lights and intricate architectural details. The walls are adorned with framed artworks and artifacts, while the floor is lined with wooden platforms displaying various dinosaur skeletons. Visitors can be seen walking around, admiring the exhibits, and the overall ambiance is one of a grand and well-maintained museum.

Tag: museum interior, dinosaur skeletons, ornate architecture, high ceiling, natural light, educational, grand, symmetrical composition, vibrant colors

Detailed: 在法国巴黎的自然历史博物馆内，一个宏伟的室内空间焕发生机，展示着令人叹为观止的恐龙骨架陈列。这个宽敞的房间拥有高耸的天花板，装饰着精致的拱门和天窗，让自然光线洒满整个空间。墙壁上装饰着各种各样的艺术品和信息展示，而地面则铺设有木制板材。恐龙骨架被精心地摆放在木制平台上，吸引着游客探索和欣赏。游客们四处漫步，有些人停下来仔细观察，而有些人则继续他们的探索之旅。尽管这幅画面非常美丽，但核实所提供信息的准确性至关重要。

Medium: 展现在眼前，其天花板装饰着精致的拱门和玻璃面板，让自然光线得以过滤。墙壁上装饰着各种各样的艺术品和信息展示，而中央的走道则陈列着引人入胜的化石和雕塑。游客们沿着走道漫步，仔细观察展品，为这个空间增添了活力。尽管这幅图像展示了博物馆的美丽，但核实所提供信息的准确性至关重要。

Short: 在法国巴黎的自然历史博物馆内，一个宏伟的室内空间展现在眼前，其天花板装饰着精致的拱门和玻璃面板，让自然光线得以过滤。墙壁上装饰着各种各样的艺术品和信息展示，而中央的走道则陈列着引人入胜的化石和雕塑。游客们沿着走道漫步，欣赏着历史和科学的奇迹。

Tag: 博物馆内部，骨骼展览，高天花板，自然光，木制平台，精美装饰，恐龙骨架，人类骨架，游客，壁画，建筑摄影，广角镜头，对称构图，温暖色调，历史氛围

Detailed: The image captures a bustling scene in a museum, where the grandeur of history and the awe of artistry are on full display. The museum's high ceiling, adorned with ornate decorations, houses ~~a large skylight~~ that bathes the space in natural light. The walls~~, painted a pristine white~~, serve as a canvas for an array of paintings that add a splash of color to the otherwise ~~monochrome setting~~. The floor, made of polished wood, reflects the light from above, creating a warm and inviting atmosphere. Hanging from the ceiling are chandeliers, their crystals catching the light and casting a soft glow throughout the room. The museum is filled with exhibits that draw the attention of the visitors. Dinosaur skeletons, their bones meticulously arranged, offer a glimpse into the prehistoric era. Paintings, each telling a story of its own, are arranged on the walls, inviting viewers to lose themselves in the world of art. People are scattered throughout the museum, some standing in quiet contemplation, others walking around, their curiosity piqued by the exhibits. Their actions bring life to the museum, making it more than just a collection of artifacts; it's a living, breathing space where history and art come to life.

Short (Florence): The Museum of Natural History in Paris features a grand interior with a high ceiling, framed artworks, and wooden platforms showcasing dinosaur skeletons, where visitors explore and admire the exhibits.

Detailed: Promotional poster for online English learning classes, featuring a young man with short brown hair, smiling and holding a pencil to his chin, positioned on the right side of the image. He wears a dark blue t-shirt and holds an open book in his left hand. The background is a dark blue with circular bokeh patterns. On the left side, there is a large speech bubble containing text in white and blue fonts. The text reads \"Online English Learning Classes,\" \"Certified English Language Teacher with over 5 years of experience,\" and \"Register Now.\" Additional details include \"Mon-Thu Day 0:00 AM-1:00 PM,\" a website URL \"www.yourdomain.com,\" and a contact number \"+000 0000 0000.\" A logo placeholder is present in the top right corner. The design is modern and clean, with a focus on clarity and readability, using a limited color palette of blue, white, and black, digital graphic design, high contrast, professional, educational, inviting.

Medium: Promotional poster for online English learning classes featuring a young man in a blue shirt holding a pencil and an open book, set against a dark blue background with circular patterns. The text highlights a certified teacher with over five years of experience, a \"REGISTER NOW\" call-to-action, and details about class schedule and contact information. The design is modern, professional, and emphasizes education.

Short: A promotional poster for online English learning classes features a young man in a blue shirt holding a pencil and an open book, with bold text and a call-to-action button, set against a dark blue background with a modern, professional design.

Tag: Online English learning, advertisement, young man, dark blue shirt, yellow pencil, open book, dark blue background, white speech bubble, text, registered teacher, experience, call to action, registration, contact number, logo, modern design, vibrant colors, professional, engaging, promotional material.

Detailed: 在线英语学习课程宣传图，图中是一位年轻男士，他留着短发，身穿蓝色T恤，右手拿着一本打开的书，左手拿着铅笔，下巴搁在铅笔上，目光向上，表情沉思。背景为深蓝色，带有圆形图案。左侧有一个大白色对话框，里面包含文字：“在线英语学习课程，认证英语语言教师，拥有超过5年经验！立即注册，周一至周四，0:00 AM-1:00 PM，www.yourdomain.com，电话+000 0000 0000”。右上角有“LOGO HERE”字样。设计风格简洁现代，色彩鲜艳，对比强烈，字体清晰易读，专业且具有教育意义。

Medium: 在线英语学习课程的广告展示了一位穿着蓝色T恤的年轻男子，他手持铅笔和笔记本，背景是深蓝色的圆形图案。文字突出了“在线英语学习课程”和“有超过5年经验的认证英语语言教师！”并提供了一个“立即注册”按钮。还包含了营业时间和网站链接，以及一个电话号码供咨询。设计简洁、现代且专业，突出了教育内容。

Short: 一个宣传在线英语学习课程的广告，展示了一位年轻男子手持铅笔和书籍，背景为深蓝色，文字为白色和蓝色，提供注册详情、教师资质和联系方式。

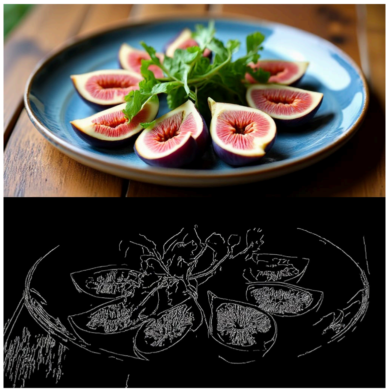
Tag: 在线英语课程，宣传海报，年轻男子，教育广告，现代设计，鲜艳色彩

Short (Florence): A promotional poster for online English classes features a young man with a pencil and book, bold text, and a call-to-action button on a dark blue modern background.

Detailed: The image is a vibrant advertisement for an online English learning class. Dominating the center of the image is a young man, smartly dressed in a blue shirt. He holds a yellow pencil in his mouth, perhaps indicating his readiness to learn. His gaze is directed upwards and to the left, A promotional poster for online English classes features a young man with a pencil and book, bold text, and a call-to-action button on a dark blue modern background. as if he's looking at something interesting. The background of the image is a stark black, providing a striking contrast to the man and the white speech bubble hovering above him. The speech bubble contains white text that reads "Online English Learning Classes" and "Certified English Language Teacher with over 5 years of experience! Register Now!", clearly conveying the purpose of the advertisement. On the right side of the image, there's additional white text that provides contact information for the class. It reads "Mon-Thu ~~9:00 AM~~-1:00 PM www.yourdomain.com Call Us~~: +1 (000) 000-0000"~~, guiding interested individuals on how to reach out for more information or to enroll in the class. Overall, the image effectively communicates the details of an online English learning class, from its certified teacher to its schedule and contact information.

Figure 5. Comparison with ShareGPT4V [6] and Florence [42] in complex scenes and dense text for caption generation. The blue underline correspond to areas with more detailed and accurate descriptions, while red underline and red strikethrough represent the incorrect and insufficient descriptions respectively.

UniCap (Ours)   ShareGPT4V   Florence



Detailed: Top image: A close-up photograph of a plate with sliced figs arranged in a circular pattern on a wooden table. The figs are halved, revealing their vibrant pink and red interiors with a purple outer skin. A small bunch of fresh green leaves, possibly arugula, is placed in the center. The plate is a rustic blue ceramic with a brown rim. The lighting is natural, coming from the left, casting soft shadows and highlighting the textures of the figs and leaves. Photorealistic, shallow depth of field, warm color palette, high contrast, natural lighting, rustic and organic feel, visually balanced composition. Bottom image: A black and white line drawing of figs on a plate, featuring the same arrangement as the top image. The figs are depicted with detailed line work, showing the intricate textures of the fruit's interior and skin. The plate is outlined with a simple, thin line. The drawing has a sketch-like quality, emphasizing the organic shapes and forms of the figs. Minimalistic, monochrome, high contrast, stylized illustration, focus on line art and texture, artistic representation of natural elements.

Medium: A top-down view of a blue ceramic plate with halved figs in a circular pattern, adorned with fresh green leaves, sits on a wooden table. The figs display vibrant pink insides and purple skin. Below, a detailed black and white sketch of figs and leaves on a black background provides a minimalist contrast. The scene features natural lighting, a warm color palette, and a balanced composition, highlighting the contrast between the photorealistic image and the sketch.

Short: A top-down view of a plate with halved figs and fresh greens on a wooden table, accompanied by a detailed black and white line drawing of figs and branches, showcasing a contrast between realistic photography and minimalist illustration.

Tag: Fig, blue ceramic plate, wooden table, green leaves, vibrant, natural lighting, shallow depth of field, rustic, artistic, minimalistic, monochrome sketch, botanical illustration, contrast, texture, organic, elegant, visually balanced

Detailed: 俯视图，一个陶瓷盘子装满了切片无花果，放在木桌上。盘子位于中心，上面有六个无花果，每个都展示了鲜艳的粉红色果肉和紫色外皮。新鲜的绿色叶子，可能是芝麻菜，被放在无花果中间。图片的下半部分是一个黑白的无花果切片和叶子的线条画，背景是黑色的。这幅画捕捉了无花果的复杂细节和叶子的纹理。自然光线，浅景深，温暖的色彩调色板，摄影写实，高对比度，乡村风格，有机感觉，艺术构图，视觉平衡。

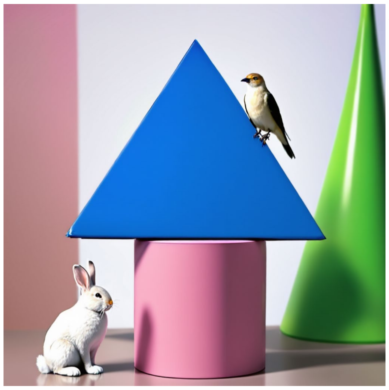
Medium:这张图片分为两部分：上半部分展示了一个充满活力的静物画，画中切开的无花果放在一个蓝色的陶瓷盘子上，摆放在一张木制桌子上，中间点缀着新鲜的绿叶；下半部分则是一个黑白的无花果切片插图，背景为黑色，突出了无花果的纹理和形状。整个构图将现实主义与插画相结合，通过对比鲜明的色彩和单色元素，创造出一个视觉上引人入胜且平衡的画面。

Short: 一个充满活力的无花果切片在蓝色陶瓷盘子上的俯视图，配有新鲜的绿叶，放置在木制桌子上，下方有黑白素描，营造出一种视觉上平衡且对比强烈的构图。

Tag: 无花果，食物摄影，木桌，自然光，高对比度，乡村风格，极简主义，艺术对比

Short (Florence): A top-down view of a plate with halved figs and greens on a wooden table, alongside a black and white line drawing, highlighting the contrast between realism and minimalism.

Detailed: The image presents a striking contrast between two different perspectives of a plate of figs. In the top photo, a blue plate cradles ~~six figs~~, their vibrant red and purple hues standing out against the cool blue of the plate. The figs are arranged in a circular pattern, creating a sense of symmetry and balance. ~~A sprig of parsley~~ rests atop the figs, adding a touch of green to the otherwise ~~monochromatic palette~~. The bottom photo is a black and white sketch of the same plate of figs. The absence of color focuses the viewer's attention on the shapes and forms within the composition. The figs, still life, and parsley are sketched with varying degrees of detail, creating a layered effect that adds depth to the image. ~~Despite the absence of color, the~~ image is rich in detail and contrast, offering a unique perspective on the simple yet captivating subject of a plate of figs.



Detailed: Minimalist still life composition featuring a white rabbit, a blue triangular prism, a pink cylindrical base, and a small bird. The rabbit is positioned in the lower left foreground, sitting on a reflective surface, facing right. The blue prism is centrally placed, balanced on the pink cylinder. A small bird with a yellow and black head perches on the right edge of the blue prism. In the background, a large green cone stands to the right, partially out of focus. The backdrop is a soft gradient of pink and white, creating a serene and balanced atmosphere. High-key lighting, clean lines, smooth textures, pastel color palette, modern and abstract, visually balanced, playful yet sophisticated.

Medium: A minimalist still life with a white rabbit on a reflective surface, a blue triangular prism on a pink cylinder, and a small bird perched on the prism. A large green cone stands in the background. The scene features a pastel color palette, high contrast, and a clean, modern aesthetic with a balanced composition.

Short: A minimalist still life featuring a blue triangular prism on a pink cylinder with a perched bird and a white rabbit, set against a pastel background with a green cone, creating a balanced and whimsical composition.

Tag: 3D rendering, surreal composition, geometric shapes, pastel colors, minimalist design, whimsical, high contrast, glossy textures, playful, visually balanced.

Detailed: 超现实主义的静物构图，以几何图形和动物为主题，背景为柔和的色彩。一只白色兔子坐在左前景，面向右侧，耳朵竖起。兔子右侧是一个粉色圆柱形底座，支撑着一个巨大的蓝色三角形。一只小而多彩的鸟儿，羽毛为黄色和黑色，停在三角形的右上角。背景中，一个绿色的圆锥体直立着，部分可见于右侧。背景由柔和的粉红色和白色垂直条纹组成，营造出一种诙谐和极简的氛围。高调照明，柔和阴影，鲜艳的色彩搭配，光滑的质感，现代艺术风格，平衡的构图，既有趣又宁静的氛围。

Medium: 一个极简主义的3D场景展示了一只白色兔子坐在粉色圆柱上，旁边是一个蓝色三角形棱镜，上面停着一只小黄鸟。背景中有一个绿色圆锥体，背景是柔和的粉红色和白色渐变。设计强调干净的线条、鲜艳的色彩和现代、有趣的审美。

Short: 一个极简主义的3D渲染场景，展示了一个蓝色三角形棱镜置于粉色圆柱体上，一只黄色和黑色的鸟儿停歇在棱镜之上，一只白色兔子坐在前景，背景是柔和的粉红色和白色，营造出一个充满活力、超现实且和谐的构图。

Tag: 极简主义，几何形状，鲜艳色彩，光滑质感，现代艺术，奇幻，鸟，兔子，三角形，圆柱形，绿锥体，柔和灯光，高对比度，抽象，当代艺术

Short (Florence): The image shows ~~a white rabbit sitting on top of a pink and blue pyramid-shaped object~~, with a bird perched on top. To the right of the rabbit is a green object, and in the background is a wall.

Detailed: In the image, a white rabbit is sitting on a pink pedestal. The rabbit is positioned on the left side of the pedestal. On the right side of the pedestal, there is a bird perched on a blue triangle. The bird is facing the rabbit, creating a sense of interaction between the two animals. In the background, there is a green cone ~~and a purple cylinder~~, adding a touch of color to the scene. The objects are arranged in a way that draws attention to the interaction between the rabbit and the bird, making it the focal point of the image. The overall scene is a blend of colors and shapes, creating a visually appealing composition.

Figure 6. Comparison with ShareGPT4V [6] and Florence [42] in visual understanding and spatial relationships. The blue underline correspond to areas with more detailed and accurate descriptions, while red underline and red strikethrough represent the incorrect and insufficient descriptions respectively.

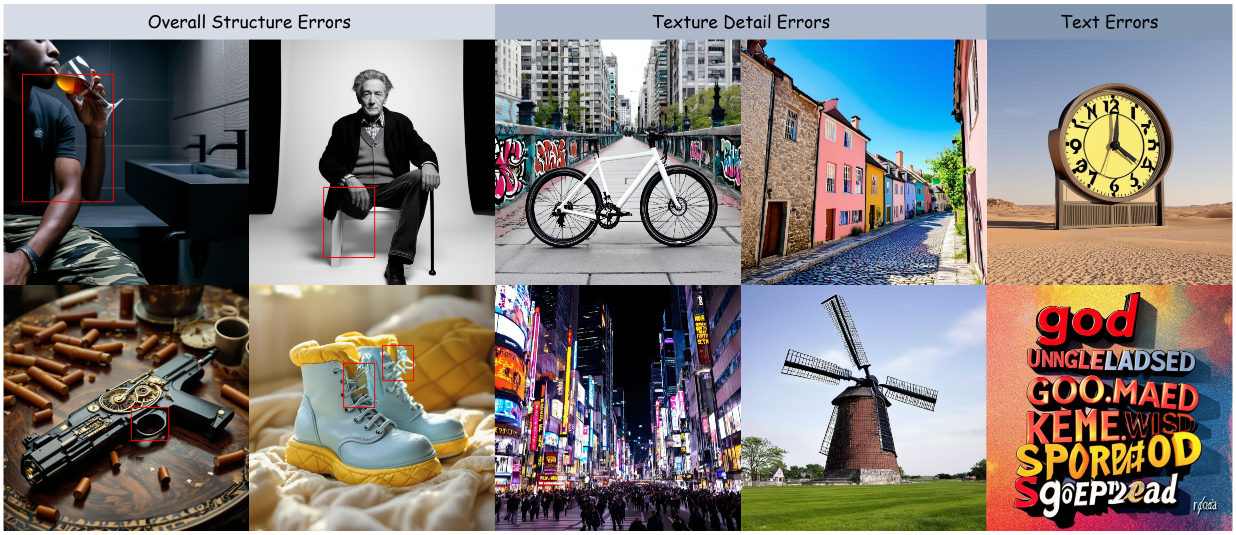


Figure 7. Generation defects of Lumina-Image 2.0, categorized into overall structural errors, texture detail errors, and text errors.