Flash-Unified: A Training-Free and Task-Aware Acceleration Framework for Native Unified Models

Supplementary Material

A. Experimental Setup

We evaluate FlashU against state-of-the-art (SOTA) unified and specialized models across multimodal understanding and generation tasks.

A.1. Baselines

Following the taxonomy proposed in Show-o2 [23], we classify baselines into three categories:

- **Native Unified Models:** Architectures that integrate understanding and generation within a single framework. We compare against *Show-o2* [23], *VILA-U* [22], *Emu3* [20], *Transfusion* [26], *SynerGen-VL* [12], *Liquid* [21], *D-DiT* [13], and *MUSE-VL* [24].
- **Assembling Tailored Models:** Systems that achieve unification by pipelining specialized experts (e.g., separate LLMs and diffusion models). Baselines include *MetaMorph* [18], *ILLUME* [19], *JanusFlow* [16], *SEED-X* [6], and *TokenFlow-XL* [17].
- **Specialized Generation Models:** To benchmark generation fidelity against domain experts, we include *Hunyuan-DiT* [14], *PixArt-Σ* [2], *SD3-Medium* [4], *Playground v2.5* [11], and *DALL-E 3* [1].

Our primary latency-performance comparison is conducted against open-source native unified models that share architectural similarities with our work, namely VILA-U [22], Show-o2 [23], and Emu3 [20]. To facilitate a more comprehensive analysis, we present an extended latency-performance comparison with *Liquid* [21], *TokenFlow-XL* [17], and *JanusFlow* [16] in Table 1 and Figure 1 of the Appendix.

A.2. Benchmarks and Datasets

Multimodal Understanding We utilize the `lmms-eval` suite to assess performance across six standard benchmarks:

- **MME** [5]: Evaluates perception and cognition using a strict Yes/No format to decouple reasoning capabilities from instruction-following bias.
- **GQA** [9]: Tests compositional reasoning via scene graph-based queries, requiring the parsing of multi-hop spatial and semantic relationships.
- **MMBench** [15]: Addresses evaluation robustness. It employs a "Circular Evaluation" strategy (shuffling options) to penalize inconsistency and random guessing.
- **MMMU** [25]: Assesses expert-level reasoning across diverse disciplines, utilizing college-level problems that demand subject-matter expertise.

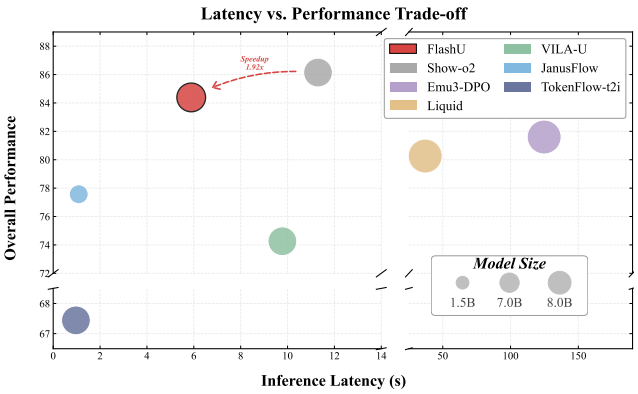


Figure 1. **Latency vs. Performance Trade-off.** Comparison of FlashU against native unified models (Show-o2, VILA-U, Emu3-DPO, Liquid) and tailored models (JanusFlow, TokenFlow-t2i). The x-axis represents inference latency (lower is better), and the y-axis represents overall performance on DPG-Bench(higher is better). Circle sizes denote model parameter counts (1.5B, 7B, 8B).

- **MMStar** [3]: Focuses on visual dependency by filtering out samples answerable via text priors or world knowledge alone ("blind answering").
- **AI2D** [10]: Examines diagrammatic reasoning, requiring the interpretation of synthetic layouts, symbolic pointers, and process flows in scientific textbook illustrations.

Visual Generation We evaluate alignment and fidelity using two specialized benchmarks designed to test prompt adherence beyond standard metrics:

- **GenEval** [7]: An object-focused evaluation framework that treats generation assessment as a detection problem. It utilizes a fixed set of prompts and pre-trained object detectors to verify specific compositional properties, including *Single/Two Object* presence, *Counting* precision, *Color* attribute binding, and *Spatial Position*.
- **DPG-Bench** [8]: Introduced in ELLA [8], this benchmark employs over 1000 dense, highly descriptive prompts to assess the model's ability to handle high information loads. It evaluates fine-grained alignment across four specific dimensions: *Global* scene consistency, *Entity* presence, *Attribute* accuracy, and *Relation* (interaction between objects).

B. Implementation Details

In this section, we provide a comprehensive overview of the FlashU inference protocol. We first visually detail the four

Table 1. Latency Comparison on DPG-Bench [8]. Results in gray correspond to vanilla Show-o2. Subscripts indicate the speedup of FlashU. Latency is reported in seconds (s), averaged per sample.

Method	# Params.	Global	Entity	Attribute	Relation	**Overall↑**	**Latency (s)↓**
Hunyuan-DiT [14]	1.5B	84.59	80.59	88.01	74.36	78.87	20.1
Playground v2.5 [11]	-	83.06	82.59	81.20	84.08	75.47	7.67
TokenFlow-t2i [17]	7B	72.64	74.72	77.67	83.80	67.44	0.97
JanusFlow [16]	1.5B	81.16	85.16	85.32	91.05	77.57	1.09
VILA-U [22]	7B	89.92	82.73	80.80	87.83	74.26	9.78
Liquid [21]	8B	79.64	87.61	85.24	91.01	80.26	37.05
Emu3-DPO [20]	8B	-	-	-	-	81.60	124.8
Show-o2 [23]	1.5B	87.53	90.38	91.34	90.30	85.02	5.23
+ FlashU (Ours)	1.5B	84.19	89.48	90.98	90.38	84.12	$2.82_{+1.85\times}$
Show-o2 [23]	7B	89.00	91.78	89.96	91.81	86.14	11.30
+ FlashU (Ours)	7B	87.79	90.30	89.29	90.30	84.39	$5.89_{+1.92\times}$

core acceleration method introduced in our framework, followed by the complete pseudocode governing the dynamic dispatching of these mechanisms and the specific hyperparameter configurations.

B.1. Pseudocode for Unified FlashU Inference

We present the unified inference protocol of FlashU in Algorithm 1. This algorithm demonstrates how the framework dynamically dispatches input queries to task-specific acceleration pathways—switching between iterative generation optimization and understanding optimization within a single unified architecture.

B.2. Hyperparameter Settings

In this section, we provide the specific hyperparameter values used for Generation and Multimodal Understanding (MMU) tasks in Table 2 and Table 3, respectively. Unless otherwise stated, all notation aligns with the definitions in the main text. Baseline and limit constants are defined as follows: T_0 denotes the total inference steps in the baseline Show-o2 model; s_0 represents the base guidance scale; and N_{max} indicates the maximum number of generated tokens for Question Answering (QA) tasks.

Table 2. Hyperparameters for Generation Tasks.

Hyperparameter	**Symbol**	**Value**
Total Inference Steps	T	$0.7T_0$
Hybrid Threshold	τ	0.3
FFN Pruning Ratio	r_p	0.20
Layer Skipping Ratio	r_{LS}	0.20
Recalculation Interval	T_{LS}	10
Diffusion Head Cache Interval	$\mathcal{T}_{cache}$	5
Switching Timestep	t_{switch}	$0.3T$
Low Guidance Scale	s_{low}	$0.75s_0$
High Guidance Scale	s_{high}	s_0

Table 3. Hyperparameters for Multimodal Understanding Tasks.

Hyperparameter	**Symbol**	**Value**
Max Generated Tokens	N_{max}	128
Hybrid Threshold	τ	0.3
FFN Pruning Ratio	r_p	0.10
Layer Skipping Ratio	r_{LS}	0.10
Calibration Samples	N_{calib}	20
Recalculation Interval	T_{LS}	10
Shallow layer to prune	L_{prune}	2
Token Pruning Ratio	ρ	0.50

C. Parameter Sensitivity Analysis

In this section, we conduct a comprehensive sensitivity analysis to evaluate the impact of key hyperparameters on both inference latency and generation quality. Specifically, we investigate the FFN Pruning Ratio (r_p), Layer Skipping Ratio (r_{LS}), and the Hybrid Threshold (τ) under different guidance strategies. We performed this analysis on the 1.5B model of Show-o2, reporting the DPG-Bench score and predicted speedup relative to the baseline.

C.1. Impact of FFN Pruning Ratio

As shown in Figure 2, we examine the model's sensitivity to the FFN Pruning Ratio (r_p) under a fixed configuration of $\tau = 0.3$ and $r_{LS} = 0.2$, utilizing the adaptive guidance strategy.

As r_p increases from 0.1 to 0.5, the speedup exhibits a moderate monotonic increase, rising from $1.82\times$ to $1.94\times$. However, the generation quality, measured by the DPG-Bench score, reveals a distinct non-linear trend. The score

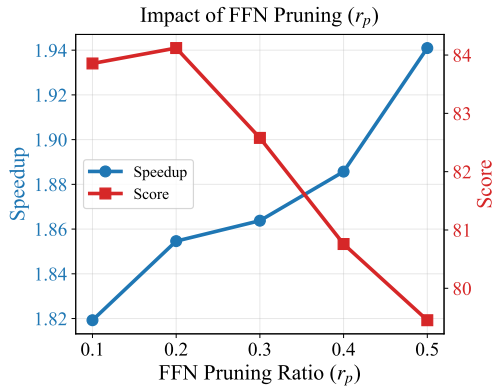


Figure 2. **Sensitivity Analysis of FFN Pruning Ratio.** The trade-off between generation quality (DPG-Bench score) and inference speedup is shown as the FFN pruning ratio (r_p) varies from 0.1 to 0.5.

peaks at 84.12 when $r_p = 0.2$, defining the optimal performance point. Beyond this point, increasing r_p to 0.4 and 0.5 leads to a noticeable degradation in quality, with the score dropping to 79.45 at $r_p = 0.5$. Further aggressive pruning yields diminishing returns in acceleration while significantly compromising visual fidelity.

C.2. Impact of Layer Skipping Ratio

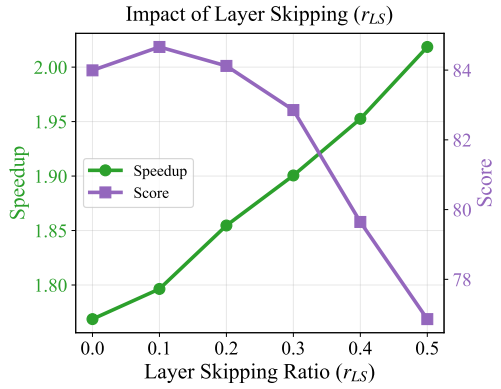


Figure 3. **Sensitivity Analysis of Layer Skipping Ratio.** The relationship between generation quality and speedup across different layer skipping ratios (r_{LS}).

We further analyze the effect of r_{LS} by varying it from 0 to 0.5, while maintaining $r_p = 0.2$ and $\tau = 0.3$.

As shown in Figure 3, the speedup demonstrates a strong positive correlation with r_{LS}, improving significantly from $1.77\times$ achieved at $r_{LS} = 0$ to $2.02\times$ at $r_{LS} = 0.5$. In terms of quality, the model exhibits robustness when $r_{LS} \leq 0.2$, with the DPG-Bench score remaining stable and peaking at 84.66 when $r_{LS} = 0.1$. However, a sharp performance collapse is observed when r_{LS} exceeds 0.3; specifically, at $r_{LS} = 0.5$, the score plummets to 76.86.

Setting r_{LS} to 0.2 establishes the upper bound for layer skipping. Although this setting incurs a negligible quality trade-off compared to the peak at $r_{LS} = 0.1$, it guarantees a substantial $1.85\times$ acceleration while maintaining robust generative capability.

C.3. Trade-off between Hybrid Threshold and Adaptive Guidance Strategy

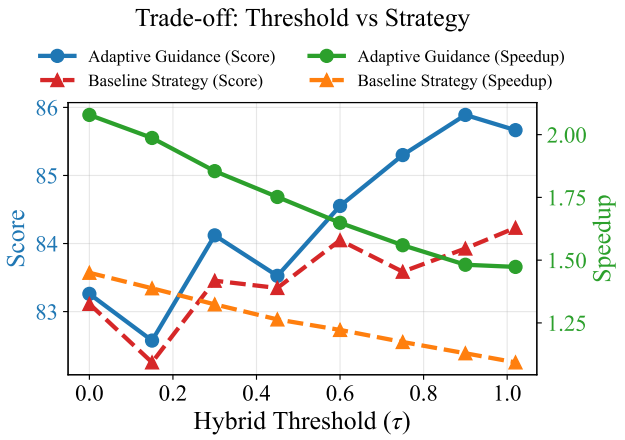


Figure 4. **Comparative Analysis of Hybrid Threshold Strategies.** The efficiency-quality trade-off curves for Adaptive Guidance versus Baseline strategies across different hybrid threshold (τ) values.

We also investigate the trade-off between efficiency and quality by varying the Hybrid Threshold (τ) across two strategies: the proposed Adaptive Guidance with reduced steps and the Baseline as shown in Figure 4.

Both strategies follow a consistent trade-off pattern: as τ increases (retaining more computation), the speedup decreases while the generation quality improves.

The Adaptive Guidance demonstrates a superior Pareto frontier compared to the Baseline. At comparable quality levels, Adaptive Guidance consistently achieves higher speedups. For instance, at $\tau \approx 0.3$, Adaptive Guidance achieves $1.85\times$ speedup versus $1.32\times$ for the Baseline, achieving a score of 84.12 compared to the Baseline's 83.45. Adaptive Guidance not only accelerates inference but also achieves a higher peak quality. At higher thresholds, such as $\tau \approx 0.9$, Adaptive Guidance reaches a score of 85.89, significantly outperforming the Baseline's peak performance of roughly 84.23. The combination of Adaptive Guidance and reduced sampling steps proves to be a robust strategy, effectively expanding the efficiency-quality operating envelope beyond the standard baseline.

Based on the analysis above, we adopt the configuration of $r_p = 0.2$, $r_{LS} = 0.2$, and $\tau = 0.3$ with Adaptive Guidance as our default setting. This combination strikes an optimal balance, delivering a $1.85\times$ speedup while maintaining a competitive DPG-Bench score of 84.12.

Algorithm 1 Unified FlashU Inference Protocol

Input: Task Indicator $\mathcal{T} \in \{\text{GEN}, \text{UND}\}$, Input Query $\mathcal{Q}$ (Text/Image), Unified Model Weights θ.
Output: Generated Response $\mathcal{R}$ (Image or Text).
Initialize:
Apply Task-Specific Network Pruning.
$\mathcal{L}_{skip} \leftarrow \emptyset$; $H_{cache} \leftarrow$ None;
$\mathcal{R} \leftarrow \emptyset$.
if $\mathcal{T} ==$ GEN **then** ▷ **Pathway A: Image Generation**
  Parse $\mathcal{Q}$ to initial noise z_T and text condition c.
  for $t = T$ **to** 1 **do** ▷ Iterative Denoising Process
    ▷ *Adaptive Guidance*
    $s_t \leftarrow (t > t_{switch})?s_{low} : s_{high}$
    $RecalcFlag \leftarrow (t \pmod{K_{skip}} == 0)$
    $h \leftarrow \text{Embed}(z_t, c)$
    for $l = 1$ **to** L **do**
      if $l \in \mathcal{L}_{skip}$ **and not** $RecalcFlag$ **then**
        continue
      $h_{in} \leftarrow h$; $h \leftarrow \text{SelfAttention}(h, \text{Mask}_{gen})$
      $h \leftarrow (t > \tau T)?h+\text{FFN}_p(h) : h+\text{FFN}_f(h)$
      if $RecalcFlag$ **and** $S(h_{in}, h) > \delta_{th}$ **then**
        $\mathcal{L}_{skip}$.add(l)
    ▷ *Diffusion Head Cache*
    if $t \pmod{K_{cache}} == 0$ **then**
      $H_{cache} \leftarrow \text{DiffHead}(h, t)$
    $z_{t-1} \leftarrow \text{SchedulerStep}(z_t, H_{cache}, s_t)$
  $\mathcal{R} \leftarrow \text{VAE-Decode}(z_0)$
else if $\mathcal{T} ==$ UND **then** ▷ **Pathway B: MMU**
  Parse $\mathcal{Q}$ to visual tokens X_v and text tokens X_t.
  for $t = 1$ **to** N_{max} **do** ▷ Autoregressive Generation
    $H \leftarrow \text{Embed}(X_v, X_t, \mathcal{R})$
    for $l = 1$ **to** L **do**
      if $l \in \mathcal{L}_{skip}$ **and not** $RecalcFlag$ **then**
        continue
      ▷ *Dynamic Visual Token Pruning*
      if $l == L_{prune}$ **then**
        $Q, K, V \leftarrow \text{Project}(H)$
        $S_{imp} \leftarrow \|V\|_2$ ▷ V-Norm proxy
        $\mathcal{I}_{keep} \leftarrow \text{TopK}(S_{imp}, (1-\rho)|X_v|)$
        $H \leftarrow \text{Gather}(\mathcal{I}_{keep} \cup X_v, X_t, \mathcal{R})$
      $H \leftarrow \text{TransformerBlock}(H)$
    $w_n \leftarrow \text{Sample}(\text{LM-Head}(H[-1]))$
    break if $w_n ==$ EOS ▷ Stopping Criterion
    $\mathcal{R}$.append(w_{next}) ▷ Update context
return $\mathcal{R}$

D. Component-wise Acceleration Analysis

To understand where FlashU's acceleration comes from, we conduct fine-grained profiling of the inference time breakdown by component on the 1.5B model. As shown in Figure 5, the analysis reveals heterogeneous acceleration across components: the Diffusion Head achieves the highest relative speedup (2.82×), primarily from iterative caching and adaptive guidance. The LLM forward pass, being the largest component (49% of total time), achieves 1.70× speedup through FFN pruning and dynamic layer skipping, contributing the most absolute time savings. Feature Extraction shows moderate acceleration (1.49×) via the Hybrid FFN's lightweight path. VAE decoding remains unchanged, confirming our optimizations introduce no overhead. This component-wise analysis validates that our multi-component optimization strategy effectively eliminates bottlenecks across the entire inference pipeline.

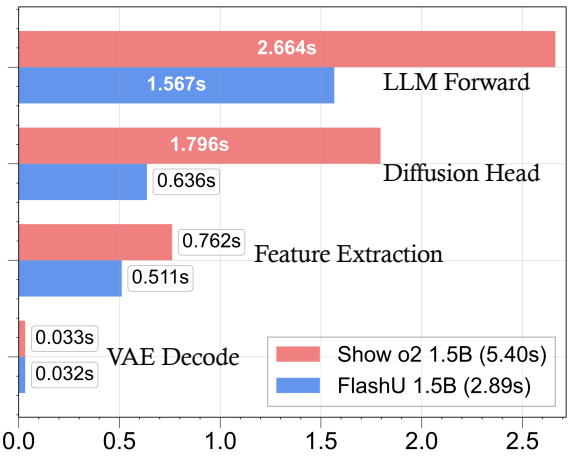


Figure 5. **Component-wise Latency Profiling Analysis.** Inference time breakdown by component for Show o2 and FlashU on the 1.5B model.

E. Additional Qualitative Results

We present additional qualitative results in Figure 6 to further demonstrate the generation capabilities of FlashU. As shown, our method generates high-fidelity and photorealistic images that accurately align with the given text prompts, effectively preserving the visual quality of the baseline model while significantly accelerating the inference process.

References

[1] 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

[2] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. 1

Prompt: An elegant and modern bathroom featuring a sleek, white rectangular bathtub filled with a **froth of soap bubbles**. The bathtub rests upon a **floor of gray**, matte tiles that complement the room's **minimalistic design**. Against the room's far wall stands a large window that frames the **warm, amber hues of a sunset**, casting a **tranquil glow** throughout the space.

Prompt: During the warm glow of a dwindling summer evening, a **particular fussy feline** with distinctive calico markings is perched atop a garden table. The cat, seemingly indifferent to its surroundings, sports a pair of **large, reflective aviator sunglasse**s that sit comically upon its small, furry face. Around the cat, there are scattered pots of **blooming flowers**, contributing to the charm of the scene, and in the background, **hints of orange and pink skies** are visible through the foliage.

Prompt: On a rustic wooden table, **three ripe eggplants** with a **glossy royal purple skin** are carefully arranged in a **neat row**. Their plump, oblong shapes complement the table's textured surface, and they **cast soft shadows** in the warm, ambient light. **Nearby**, the **woven pattern** of a tan-colored napkin **peeks out from beneath** the vibrant, richly colored vegetables.

Prompt: A curious vessel, with an **architecture reminiscent of a giant green broccoli**, basks in the **bright sunlight**, casting a shimmer across its intricate, leaf-like structures. It **floats serenely in the midst of a vast ocean**, the **water around it sparkling as if sprinkled with diamonds** due to the sun's reflection. The horizon stretches endlessly, with **the clear blue sky meeting the deep azure of the sea** at a distant line.

Prompt: An **ornate silver cosmetics mirror** gracefully positions itself upon a pristine **white marble** vanity top. **Surrounding the mirror** are various high-end makeup products and delicate perfume bottles, **each catching the room's natural light in their uniquely colored glass**. The vanity itself, sleek in design with clean lines, is nestled in a space that feels both modern and timeless.

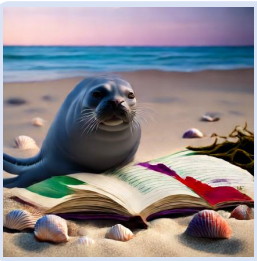
Prompt: A deep red rose with plush petals sits elegantly **coiled atop** an **ivory, intricately patterned lace napkin**. The napkin rests on a rustic wooden table that contributes to the charming garden setting. As the late evening sun casts a **warm golden hue** over the area, the **shadows of surrounding foliage dance gently around the rose**, enhancing the romantic ambiance. Nearby, the **green leaves** of the garden plants provide a fresh and verdant backdrop to the scene.

Prompt: Two multicolored butterflies with **delicate, veined wings** gently balance atop a vibrant, orange tangerine in a bustling garden. The tangerine, with its **glossy, dimpled texture**, is situated on a wooden table, contrasting with the **greenery of the surrounding foliage and flowers**. The butterflies, appearing nearly small in comparison, add a touch of grace to the scene, complementing the natural colors of the **verdant backdrop**.

Prompt: On the soft, warm sand of the beach, a fluffy white rabbit with rounded ears is caught in a curious moment, gently placing its paw on the **ribbed surface of a pink scallop shell**. The scallop, slightly open, reveals its smooth interior contrasting with its coarse outer texture, while hues of pink and orange from the setting sun reflect off its surface. There's a **tranquil ocean backdrop** with the gentle ebbing of the tide, and the **fading daylight casts a golden glow** over the scene, highlighting the rabbit's soft fur and the shell's subtle color.

Prompt: A traditional Venetian mask with intricate designs and feather embellishments is placed on a polished wooden table. Next to the mask, there sits **a substantial golden trophy**, shining with **reflected light**, its **surface etched with small, detailed engravings**. The table and its prestigious items are set against a **deep sepia-toned backdrop** that enhances the objects' visual appeal.

Prompt: In the depths of a **vibrant underwater scene**, **a large, dark red fish** glides gracefully through the water, its scales glistening with the filtered sunlight from above. Surrounding the fish is a **lively coral reef**, bustling with an array of corals in **striking hues of purple, yellow, and green**, their unique and intricate forms providing a stunning backdrop. **Tiny, iridescent fish** dart around the nooks of the corals, adding to the dynamic and rich tapestry of marine life inhabiting this tranquil aquatic world.



Prompt: During a tranquil evening, a grey seal with inquisitive eyes explores a **vibrantly illustrated book** left on the **coarse sands of a deserted beach**. The book's pages flutter in the gentle sea breeze, **revealing bursts of purples, greens, and reds**. Nearby, **scattered seashells** and **washed-up seaweed** provide a natural backdrop to this unusual scene.

Prompt: A small, **teardrop-shaped candle** with a **pale blue hue** graces the surface of a large, **cubic storage box**. The box itself features a textured, glossy white finish and sits squarely in the corner of a room. To the side of the box, there's a stack of **neatly folded towels in varying shades of beige and cream**.

Figure 6. **Qualitative results of text-to-image generation.** We showcase samples generated by FlashU across diverse prompts. The model successfully handles complex descriptions, rendering high-quality textures and correct object placements, demonstrating that our acceleration framework preserves the generative performance of the backbone model.

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1

[4] 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 *ICML*, 2024. 1

[5] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. 1

[6] 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. 1

[7] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 1

[8] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. 1, 2

[9] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. 1

[10] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 1

[11] 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*, abs/2402.17245, 2024. 1, 2

[12] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024. 1

[13] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024. 1

[14] 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. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 1, 2

[15] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[16] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024. 1, 2

[17] 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. 1, 2

[18] 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. 1

[19] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. ILLUME: illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024. 1

[20] 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

[21] Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable multi-modal generators. *arXiv preprint arXiv:2412.04332*, 2024. 1, 2

[22] 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. 1, 2

[23] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 1, 2

[24] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. MUSE-VL: modeling unified VLM through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024. 1

[25] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pages 9556–9567. IEEE, 2024. 1

[26] Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025. 1