Adaptive Routing of Text-to-Image Generation Requests Between Large Cloud Model and Light-Weight Edge Model

Supplementary Material

Text Encoder	#Parameter	Max Length (tokens)	Vocabulary Size
CLIP-ViT/L	123.65 M	77	49408
OpenCLIP-ViT/H	354.0 M	77	49408
OpenCLIP-ViT/G	694.7 M	77	49408
Flan-T5-XL	3 B	512	32128
T5-XXL	11 B	512	32128

Table 1. Text encoders of text-to-image models.

A. Details of Text-to-Image Model Choices on Edge and Cloud

To evaluate the performance of routing methods, we choose a series of text-to-image (T2I) models with varying sizes, computational costs, and performances on edge and cloud. This section provides a brief overview of the T2I models used in our experiments. As shown in Tab. 2, we utilize diffusion-based models such as Stable Diffusion 1.5 (SD1.5), Stable Diffusion 2.1 (SD2.1), Stable Diffusion XL (SDXL), and Stable Diffusion XL-Refiner (XL-Refiner), alongside diffusion Transformer models Stable Diffusion 3 (SD3) and PixArt-α, and an autoregressive model Infinity. Notably, Stable Diffusion XL-Refiner is a refinement model applied after image generation by SDXL to enhance image quality. We test the performance of our routing method under different edge-cloud model pairs to verify its effectiveness.

These T2I models also differ in their use of text encoders, which play a critical role in conditioning the models on textual input. As detailed in Tab. 2, the models vary in both the type and number of text encoders employed. For instance, SD1.5 and SD2.1 rely on lightweight encoders such as CLIP or OpenCLIP, while SD3 incorporates three text encoders: CLIP, OpenCLIP, and T5. As shown in Tab. 1, these text encoders vary significantly in parameters and maximum token length limits. It is worth noting that the T5 encoder, used in SD3, PixArt-α, and Infinity, is a large-scale text-to-text transfer transformer model whose size may even exceed that of the T2I model itself. This means that when comparing the relative size of T2I models, the size of the text encoder cannot be ignored and needs to be taken into account.

B. Comparison of Input and Output Spaces for Text-to-Image Models

To begin, we define the sizes of text and image spaces. For simplicity, we measure the size of a space by the number of possible samples it can contain.

For the text space, models typically have a fixed vocabulary size and a maximum input length. Assuming a vocabulary size of $|\mathcal{V}|$ and a maximum input token length of L, the size of the text space can be approximated as:

$$S_{\text{text}} = |\mathcal{V}|^L \tag{1}$$

The size of the image space is determined by the image resolution and the color depth per pixel. Each pixel's color depth defines the number of possible states for that pixel. For an RGB image with a resolution of $W \times H$ and a color depth of 24 bits (8 bits each for red, green, and blue), the size of the image space can be approximated as:

$$S_{\text{image}} = (2^{\text{color\_depth}})^{W \times H} \tag{2}$$

To facilitate comparison between large numerical values of space sizes, we define the scale ratio R as the logarithmic of the ratio between output space size S_{out} and input space size S_{in}:

$$R = \log(S_{\text{out}}/S_{\text{in}}) \tag{3}$$

If $R > 0$, the output space is larger than the input space; if $R < 0$, the input space is smaller than the output space; and if $R \approx 0$, the two spaces are of comparable size.

For large language models (LLMs), both the input and output spaces are text-based, with similar vocabulary sizes and token limits. Therefore, the scale ratio of output-to-input space sizes can be approximated as:

$$R_{\text{LLM}} = \log(S_{\text{text\_out}}/S_{\text{text\_in}}) \approx 0 \tag{4}$$

This indicates that the input and output spaces are roughly equal in scale.

For text-to-image models, the input space is text-based, while the output space is image-based. Assuming the text encoder uses CLIP with a vocabulary size of $|\mathcal{V}| = 49408$ and a maximum input length of $L = 77$ as shown in Tab. 1, and the output is a 512×512 RGB image with a color depth of 24 bits, the scale ratio of output-to-input space sizes can be approximated as:

$$\begin{aligned} R_{\text{T2I}} &= \log(S_{\text{image}}/S_{\text{in}}) && (5)\\ &= HW \cdot color\_depth \cdot \log 2 - L \cdot \log(|\mathcal{V}|) && (6)\\ &= 512 \times 512 \times 24 \log 2 - 77 \times \log 49408 && (7)\\ &\approx 4360072 && (8) \end{aligned}$$

Text-to-Image Model	Text Encoder	Type	#Param
Stable Diffusion 1.5	CLIP-ViT/L	Diffusion	0.86 B
Stable Diffusion 2.1	OpenCLIP-ViT/H	Diffusion	0.86 B
Stable Diffusion XL	OpenCLIP-ViT/G and CLIP-ViT/L	Diffusion	2.6 B
Stable Diffusion XL-Refiner	OpenCLIP-ViT/G and CLIP-ViT/L	Diffusion	-
Stable Diffusion 3	OpenCLIP-ViT/G, CLIP-ViT/L and T5-XXL	Diffusion Transformer	8 B
PixArt-α	T5-XXL	Diffusion Transformer	0.6 B
Infinity	Flan-T5-XL	AutoRegressive	2 B

Table 2. Details of text-to-image models used in routing.

Metrics	Positive/Negative Text Pairs
Definition	(“High definition photo”, “Low definition photo”)
Detail	(“Detailed photo”, “Lacking Detail photo”)
Clarity	(“Clear photo”, “Blurred photo”)
Sharpness	(“Sharp”, “Hazy”)
Harmony	(“Visually harmonious”, “Visually chaotic”)
Realism	(“Realism”, “Distortion”)
Color	(“Color accurate”, “Color distorted”)
Consistency	(“Color consistency”, “Color conflict”)
Layout	(“Reasonable composition”, “Chaotic composition”)
Integrity	(“Object completion”, “Object twisting”)

Table 3. Multi-metric image generation quality and their positive/negative text description pairs.

Similarly, for T5 text encoder with a vocabulary size of $|\mathcal{V}| = 32128$ and a maximum input length of $L = 512$, the scale ratio is:

$$R_{\text{T2I}} = \log(S_{\text{image}}/S_{\text{in}}) \tag{9}$$
$$= 512 \times 512 \times 24 \log 2 - 512 \times \log 32128 \tag{10}$$
$$\approx 4355591 \tag{11}$$

This indicates that, for T2I models, the output image space is at least $e^{4\times10^6}$ times larger than the input text space. Considering that T2I models can generate higher-resolution images, such as 768×768 or 1024×1024, this ratio becomes even larger.

Thus, for LLMs, the input and output spaces are roughly comparable in size. In contrast, for T2I models, the output image space is significantly larger than the input text space. This means that predictive routing for T2I models must infer quality changes in a vastly larger and more complex output image space based on a constrained input text space. This significant disparity poses a greater challenge for predictive routing of T2I models. To address this, we propose a routing optimization strategy based on multi-dimensional image quality metrics to reduce noise and inaccuracies in predictions, while designing the routing model to capture the complex mapping between input and output spaces.

C. Image Quality Metrics

To comprehensively evaluate the quality of generated images, we introduce a set of metrics that include both general criteria applicable to real photos and unique indicators specific to generated images. As shown in Tab. 3, our evaluation, which targets a realistic object generation task, adopts 10 commonly used metrics suggested by HEIM [2], T2I-Scorer [6], and VisionPrefer [8]. Metrics such as clarity and sharpness are considerations for real photos but also apply to generated images, and metrics like object integrity and realism are unique to evaluating generated images. Each of these metrics is accompanied by a pair of positive and negative descriptive texts that characterize the nature of the attribute being evaluated. This multi-dimensional quality metric allows us to more comprehensively consider subtle differences between images generated by different T2I models, enabling more accurate routing decision. Furthermore, our approach is not limited to the aforementioned metrics. Tab. 3 merely illustrates an example. Considering that different generation scenarios have varying quality requirements, the multi-metric quality formulation and the weights for different metrics in Eq. (3) and Eq. (4) of the main text also allow tailoring to specific application needs. Image quality in our routing objective can encompass any number of dimensions and utilize arbitrary evaluation methods without requiring modifications to our routing framework.

D. Human Preference

We evaluate alignment with human judgments using the SOTA Human Preference Score (HPSv2) [7] and Multi-dimensional Human Preference (MPS) [9], whose scoring models were trained on human preference datasets. Although our routing is not specifically optimized for these metrics, it still achieves 19.50% and 7.57% improvements over random routing, as shown in Tab. 4, on ΔP, which quantifies how much of the quality gap between the edge and cloud models is recovered by routing. These results show that our routing generalizes well and aligns closely with human preferences.

Router	HPSv2 [7]	$\Delta P(\%)$	MPS [9]	$\Delta P(\%)$
Edge	0.2725	-	0.3193	-
Cloud	0.2925	-	0.6806	-
Random	0.2825	50.00	0.5000	50.00
RouteLLM-BERT [4]	0.2843	59.00	0.5058	51.62
RouteLLM-MF [4]	0.2852	63.50	0.5138	53.83
HybridLLM [1]	0.2842	58.50	0.5058	51.62
ZOOTER [3]	0.2846	60.50	0.5137	53.81
`RouteT2I` (Ours)	**0.2864**	**69.50**	**0.5273**	**57.57**

Table 4. The alignment of images generated by text-to-image models with human judgments, with the router selecting between edge and cloud models for each prompt at a 50% routing rate.

E. Training Cost and Inference Latency

Our routing model, comprising 58.17 million parameters, can be trained in 7 minutes on a single NVIDIA RTX 4090D GPU. During training, it efficiently utilizes only 2.7GB of system RAM and 2.5GB of GPU VRAM, demonstrating strong computational efficiency. For edge deployment, the model shows promising performance on embedded platforms. On the NVIDIA Jetson TX2, inference takes an average of 64.5ms per image, using 4.0GB RAM and 1.6GB swap space. On the less powerful NVIDIA Jetson Nano, the inference time increases to 131.3ms per image, using 1.1GB RAM and 2.7GB swap. These results highlight the efficiency of our routing model and its suitability for resource-constrained edge devices, showing that on-device routing inference incurs only a small overhead while achieving notable performance improvements.

F. Results on Other Datasets

We also conduct experiments on a subset of the public LAION2B-en-aesthetic dataset [5], consisting of 20k samples for training and 10k for validation. Prompts from both the COCO and the LAION datasets are collected from authentic human inputs. Compared to COCO prompts' concise and objective styles, LAION's are more colloquial and stylistic. Tab. 5 shows our method still performs well on the LAION dataset, demonstrating its generalization.

G. Visual Results

In Fig. 1, we demonstrate the visual results of our `RouteT2I` at a 50% routing rate, where the router selects the most suitable model between the edge or cloud text-to-image model for each input prompt. The results show that when the edge model performs comparably or even better than the cloud model, the router tends to choose the edge model for generation. Conversely, when the cloud model is significantly superior, it is selected instead. This approach allows us to maintain high generation quality while reducing reliance on cloud services.

H. Results of More Text-to-Image Model Pairs

To validate the performance of our routing approach across various edge-cloud T2I model pairs, we conduct experiments not only with SD3 and SD2.1 pairs as shown in Tab. 2 (main text) but also with other model pairs. These T2I models, used on the edge and the cloud, differ in architecture, text encoders, parameter counts, and performance, serving to verify the robustness of our routing method under diverse conditions.

We present the multi-metric image quality performances and their corresponding relative performance improvements for various T2I models on edge and cloud at a routing rate of 50% in Tab. 6 to Tab. 22. Our proposed `RouteT2I` demonstrates outstanding performance across all combinations. For instance, when SD2.1 is used as the edge model and Pixart-α is used in the cloud in Tab. 16, the performance gap between them is relatively small. In this scenario, `RouteT2I` achieves its maximum relative performance improvement, enhancing performance by 235.22% compared to the improvement of cloud models over edge models. Conversely, when SD1.5 is used as the edge model and SD3 is used in the cloud in Tab. 8, the performance gap between them is much larger. In this case, `RouteT2I` still achieves significant gain, reaching 89.27% of the improvement brought by the cloud model. Notably, these performance enhancements are achieved at a routing rate of 50%, meaning that we can reduce cloud model calls by half while still approaching or even surpassing the generation quality of more powerful cloud T2I models. This demonstrates the efficiency of our T2I routing method, which remains effective across diverse combinations of edge and cloud T2I models. By minimizing reliance on costly cloud services, `RouteT2I` not only reduces operational expenses but also ensures high-quality image generation.

Additionally, the T2I models used as cloud and edge models may have different text encoders. For example, SD1.5 uses CLIP, PixArt-α uses T5, and SD3 uses multiple encoders simultaneously. These encoders map text into different vector spaces, posing challenges for prompt-based routing. However, our method remains effective. For instance, when routing between SDXL and SD1.5 in Tab. 14, which use different numbers of text encoders, our method achieves a quality improvement equivalent to 41.79% of the cloud model's gain over the edge model, surpassing other methods by at least 6%. For routing between models with different encoder types, such as PixArt-α and SD2.1 in Tab. 16, our method reaches an improvement of 235.22% of the cloud model's gain, outperforming others by at least 20%. Notably, for the pair of SD3 and SD1.5 in Tab. 8, where SD3's text encoder far exceed that of SD1.5 in both the number of encoders and parameters, we still achieve an 89.2% performance gain. This demonstrates that our routing model, equipped with dual-gate token selection MoE,

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model	0.6922	0.6496	0.6019	0.6477	0.6284	0.5974	0.4922	0.5308	0.4835	0.4611	-
Cloud Model	0.7057	0.6716	0.6326	0.6657	0.6323	0.6454	0.5334	0.5481	0.5315	0.4791	-
Random	0.6990	0.6606	0.6173	0.6567	0.6304	0.6214	0.5128	0.5395	0.5075	0.4701	50.00
RouteLLM-BERT [4]	**0.7073**	0.6674	0.6226	0.6591	0.6331	0.6257	0.5150	0.5424	0.5130	0.4707	73.89
RouteLLM-MF [4]	0.7067	**0.6680**	0.6240	0.6619	0.6326	0.6276	0.5159	0.5422	0.5131	0.4709	75.21
Hybrid LLM [1]	**0.7073**	0.6670	0.6235	0.6613	0.6331	0.6273	0.5169	0.5428	0.5138	0.4712	76.80
ZOOTER [3]	0.7053	0.6657	0.6222	0.6619	0.6337	0.6285	0.5167	0.5427	0.5131	**0.4719**	76.45
`RouteT2I` (Ours)	0.7063	0.6677	**0.6241**	**0.6624**	**0.6344**	**0.6291**	**0.5175**	**0.5437**	**0.5145**	0.4710	**81.38**

Table 5. The multi-dimensional quality of images generated by text-to-image models on the LAION dataset, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better.

Figure 1. Visual results of our `RouteT2I` at a routing rate of 50%, using SD2.1 as the edge model and SD3 as the cloud model. The selected side is highlighted in red. This routing decision is made before generation begins, and the unselected side does not actually perform any generation tasks.

focuses on the semantic meaning of tokens rather than the vector spaces associated with specific encoders. Our approach is consistently efficient across T2I models using diverse text encoders.

We also experiment with the impact of different T2I model architectures on routing. Specifically, we select three mainstream generation architectures: diffusion, diffusion Transformer, and autoregressive. We evaluate routing across different combinations of model types, such as diffusion models, diffusion and diffusion Transformer models, diffusion and autoregressive models, and diffusion Transformer and autoregressive models in Tab. 6 to Tab. 22. Our `RouteT2I` demonstrates consistent effectiveness across all architecture types, achieving significant performance gains in every scenario.

References

[1] Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024. 3, 4

[2] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023. 2

[3] Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. *arXiv preprint arXiv:2311.08692*, 2023. 3, 4

[4] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024. 3, 4

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information*

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: XL-Refiner	0.6083	0.6584	0.5835	0.6013	0.6086	0.5773	0.4701	0.5195	0.4824	0.4850	-
Cloud Model: SD3	0.6337	0.6847	0.6346	0.6703	0.5930	0.5868	0.5134	0.5199	0.5345	0.4972	-
Random	0.6210	0.6715	0.6091	0.6358	0.6008	0.5821	0.4917	0.5197	0.5084	0.4911	40.00
RouteLLM-BERT	0.6236	0.6672	**0.6210**	**0.6431**	0.6057	0.5956	0.5024	0.5236	0.5158	0.4963	152.48
RouteLLM-MF	0.6231	0.6662	0.6201	0.6429	0.6043	0.5956	0.5022	0.5231	0.5145	0.4955	140.40
HybridLLM	0.6239	**0.6694**	**0.6210**	0.6421	0.6044	0.5936	0.5036	0.5237	**0.5178**	0.4951	152.13
ZOOTER	**0.6239**	0.6687	0.6186	0.6413	0.6054	0.5945	0.5026	0.5230	0.5168	0.4956	138.95
`RouteT2I` (Ours)	0.6231	0.6676	0.6200	0.6410	**0.6063**	**0.5967**	**0.5039**	**0.5250**	0.5164	**0.4964**	**184.92**

Table 6. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. XL-Refiner is used as the edge model, while SD3 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SDXL	0.5900	0.6564	0.5645	0.5832	0.6042	0.5660	0.4622	0.5282	0.4866	0.4892	-
Cloud Model: SD3	0.6337	0.6847	0.6346	0.6703	0.5930	0.5868	0.5134	0.5199	0.5345	0.4972	-
Random	0.6119	0.6705	0.5996	0.6267	0.5986	0.5764	0.4878	0.5241	0.5105	0.4932	30.00
RouteLLM-BERT	0.6161	0.6667	0.6105	**0.6335**	0.6020	0.5893	0.4983	0.5277	0.5171	0.4983	55.28
RouteLLM-MF	0.6170	0.6659	0.6103	0.6325	0.6019	0.5896	0.4984	0.5277	0.5172	0.4987	55.70
HybridLLM	0.6156	**0.6694**	0.6110	0.6331	0.6012	0.5880	0.5008	0.5257	**0.5206**	0.4978	53.08
ZOOTER	**0.6184**	0.6680	0.6107	0.6328	0.6011	0.5882	0.5006	0.5266	0.5198	0.4973	53.41
`RouteT2I` (Ours)	0.6155	0.6669	**0.6121**	0.6316	**0.6031**	**0.5909**	**0.5014**	**0.5286**	0.5200	**0.5002**	**61.83**

Table 7. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SDXL is used as the edge model, while SD3 serves as the cloud model.

processing systems, 35:25278–25294, 2022. 3

[6] Haoning Wu, Xiele Wu, Chunyi Li, Zicheng Zhang, Chaofeng Chen, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. T2i-scorer: Quantitative evaluation on text-to-image generation via fine-tuned large multi-modal models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3676–3685, 2024. 2

[7] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 3

[8] Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Multimodal large language models make text-to-image generative models align better. *Advances in Neural Information Processing Systems*, 37:81287–81323, 2024. 2

[9] Sixian Zhang, Bohan Wang, Junqiang Wu, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Learning multi-dimensional human preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2024. 2, 3

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SD1.5	0.6035	0.6615	0.5912	0.6454	0.5944	0.5444	0.4659	0.5149	0.4945	0.4730	-
Cloud Model: SD3	0.6337	0.6847	0.6346	0.6703	0.5930	0.5868	0.5134	0.5199	0.5345	0.4972	-
Random	0.6186	0.6731	0.6129	0.6579	0.5937	0.5656	0.4896	0.5174	0.5145	0.4851	40.00
RouteLLM-BERT	0.6227	**0.6782**	0.6197	0.6635	0.5968	0.5704	0.4947	0.5197	0.5202	0.4861	77.48
RouteLLM-MF	**0.6234**	0.6773	0.6195	0.6634	0.5961	0.5708	0.4947	0.5195	0.5199	0.4868	71.83
HybridLLM	0.6219	0.6766	**0.6214**	0.6641	0.5970	0.5723	0.4977	0.5199	0.5223	0.4872	80.95
ZOOTER	0.6211	0.6762	0.6213	0.6638	0.5966	0.5719	0.4975	0.5197	0.5223	0.4877	76.94
`RouteT2I` (Ours)	0.6211	0.6749	**0.6214**	**0.6643**	**0.5979**	**0.5727**	**0.4985**	**0.5210**	**0.5227**	**0.4880**	**89.27**

Table 8. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SD1.5 is used as the edge model, while SD3 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: XL-Refiner	0.6083	0.6584	0.5835	0.6013	0.6086	0.5773	0.4701	0.5195	0.4824	0.4850	-
Cloud Model: SD2.1	0.6251	0.6685	0.6076	0.6537	0.5949	0.5575	0.4680	0.5088	0.4860	0.4690	-
Random	0.6167	0.6634	0.5956	0.6275	0.6018	0.5674	0.4690	0.5141	0.4842	0.4770	0.00
RouteLLM-BERT	0.6219	0.6649	0.6068	0.6389	0.6065	0.5768	0.4798	0.5180	0.4913	0.4778	94.96
RouteLLM-MF	0.6212	0.6638	0.6072	**0.6391**	0.6057	0.5765	**0.4804**	0.5183	0.4913	0.4777	96.10
HybridLLM	0.6224	**0.6653**	0.6059	0.6383	0.6072	0.5773	0.4798	0.5184	0.4924	**0.4781**	99.58
ZOOTER	0.6220	**0.6653**	0.6052	0.6364	0.6066	0.5766	0.4790	0.5184	0.4910	0.4776	90.11
`RouteT2I` (Ours)	**0.6226**	0.6651	**0.6075**	0.6384	**0.6079**	**0.5778**	**0.4804**	**0.5185**	**0.4926**	0.4779	**104.21**

Table 9. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. XL-Refiner is used as the edge model, while SD2.1 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SDXL	0.5900	0.6564	0.5645	0.5832	0.6042	0.5660	0.4622	0.5282	0.4866	0.4892	-
Cloud Model: SD2.1	0.6251	0.6685	0.6076	0.6537	0.5949	0.5575	0.4680	0.5088	0.4860	0.4690	-
Random	0.6076	0.6624	0.5861	0.6184	0.5996	0.5617	0.4651	0.5185	0.4863	0.4791	0.00
RouteLLM-BERT	**0.6142**	0.6645	0.5964	0.6290	0.6025	0.5708	0.4751	0.5231	0.4926	0.4799	134.47
RouteLLM-MF	0.6141	0.6636	**0.5967**	**0.6299**	0.6019	0.5699	0.4756	0.5231	0.4928	0.4798	136.41
HybridLLM	0.6130	0.6642	0.5950	0.6269	0.6039	0.5708	0.4756	**0.5241**	0.4924	**0.4806**	133.10
ZOOTER	**0.6142**	0.6641	0.5949	0.6269	0.6041	0.5704	0.4736	0.5227	0.4921	0.4795	124.23
`RouteT2I` (Ours)	0.6140	**0.6646**	0.5958	0.6274	**0.6055**	**0.5715**	**0.4760**	**0.5241**	**0.4934**	0.4803	**152.44**

Table 10. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SDXL is used as the edge model, while SD2.1 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SD1.5	0.6035	0.6615	0.5912	0.6454	0.5944	0.5444	0.4659	0.5149	0.4945	0.4730	-
Cloud Model: SD2.1	0.6251	0.6685	0.6076	0.6537	0.5949	0.5575	0.4680	0.5088	0.4860	0.4690	-
Random	0.6143	0.6650	0.5994	0.6496	0.5947	0.5510	0.4669	0.5118	0.4903	0.4710	20.00
RouteLLM-BERT	0.6183	0.6663	0.6046	**0.6531**	0.5969	0.5570	0.4723	0.5122	0.4926	0.4720	112.83
RouteLLM-MF	0.6176	0.6657	**0.6054**	0.6528	0.5959	0.5571	0.4725	0.5125	0.4927	**0.4726**	94.83
HybridLLM	0.6167	0.6667	0.6050	0.6523	**0.5970**	0.5556	0.4712	0.5127	**0.4933**	0.4720	109.52
ZOOTER	0.6158	0.6662	0.6031	0.6525	0.5969	0.5555	0.4709	**0.5135**	0.4932	0.4715	103.70
`RouteT2I` (Ours)	**0.6186**	**0.6668**	0.6048	0.6521	**0.5970**	**0.5573**	**0.4726**	0.5126	**0.4933**	**0.4726**	**119.18**

Table 11. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SD1.5 is used as the edge model, while SD2.1 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SDXL	0.5900	0.6564	0.5645	0.5832	0.6042	0.5660	0.4622	0.5282	0.4866	0.4892	-
Cloud Model: XL-Refiner	0.6083	0.6584	0.5835	0.6013	0.6086	0.5773	0.4701	0.5195	0.4824	0.4850	-
Random	0.5992	0.6574	0.5740	0.5922	0.6064	0.5717	0.4661	0.5239	0.4845	0.4871	20.00
RouteLLM-BERT	0.6008	0.6567	0.5758	0.5948	0.6108	0.5736	0.4696	**0.5249**	0.4871	0.4874	44.04
RouteLLM-MF	0.6012	0.6574	0.5761	**0.5950**	0.6108	0.5742	0.4694	0.5239	0.4870	0.4874	46.82
HybridLLM	0.6005	0.6564	0.5759	0.5947	**0.6115**	0.5730	0.4690	0.5245	0.4873	0.4866	40.89
ZOOTER	0.6003	0.6574	0.5761	0.5948	0.6110	0.5741	0.4696	0.5245	0.4870	0.4875	48.22
`RouteT2I` (Ours)	**0.6014**	**0.6575**	**0.5763**	**0.5950**	0.6113	**0.5748**	**0.4700**	0.5239	**0.4874**	**0.4876**	**51.25**

Table 12. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SDXL is used as the edge model, while XL-Refiner serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SD1.5	0.6035	0.6615	0.5912	0.6454	0.5944	0.5444	0.4659	0.5149	0.4945	0.4730	-
Cloud Model: XL-Refiner	0.6083	0.6584	0.5835	0.6013	0.6086	0.5773	0.4701	0.5195	0.4824	0.4850	-
Random	0.6059	0.6600	0.5874	0.6234	0.6015	0.5609	0.4680	0.5172	0.4885	0.4790	10.00
RouteLLM-BERT	0.6090	0.6614	**0.5983**	0.6354	0.6079	0.5693	0.4763	0.5196	0.4945	0.4804	76.34
RouteLLM-MF	0.6078	0.6603	0.5977	**0.6360**	0.6075	0.5694	0.4761	0.5198	0.4944	0.4804	69.25
HybridLLM	0.6084	0.6606	0.5968	0.6359	0.6073	0.5680	0.4754	0.5191	0.4948	0.4802	66.31
ZOOTER	**0.6104**	**0.6615**	0.5964	0.6357	0.6071	0.5677	0.4760	0.5209	0.4951	0.4795	77.68
`RouteT2I` (Ours)	0.6091	**0.6615**	0.5969	0.6354	**0.6082**	**0.5697**	**0.4768**	**0.5212**	**0.4957**	**0.4808**	**80.83**

Table 13. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SD1.5 is used as the edge model, while XL-Refiner serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SD1.5	0.6035	0.6615	0.5912	0.6454	0.5944	0.5444	0.4659	0.5149	0.4945	0.4730	-
Cloud Model: SDXL	0.5900	0.6564	0.5645	0.5832	0.6042	0.5660	0.4622	0.5282	0.4866	0.4892	-
Random	0.5968	0.6590	0.5779	0.6143	0.5993	0.5552	0.4640	0.5216	0.4906	0.4811	-10.00
RouteLLM-BERT	**0.6026**	0.6609	**0.5885**	0.6253	0.6019	0.5627	0.4706	0.5232	0.4954	0.4826	35.75
RouteLLM-MF	0.5993	0.6602	0.5878	0.6250	0.6034	0.5628	0.4708	0.5241	0.4959	0.4832	35.36
HybridLLM	0.5989	**0.6610**	0.5862	0.6237	0.6040	0.5614	0.4700	0.5240	0.4954	0.4833	33.08
ZOOTER	0.5992	0.6603	0.5870	**0.6254**	0.6021	0.5608	0.4709	**0.5248**	0.4957	0.4830	33.41
`RouteT2I` (Ours)	0.6004	**0.6610**	0.5874	0.6242	**0.6042**	**0.5635**	**0.4716**	**0.5248**	**0.4960**	**0.4836**	**41.79**

Table 14. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SD1.5 is used as the edge model, while SDXL serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: PixArt-α	0.6660	0.7039	0.6069	0.6499	0.6114	0.6022	0.4562	0.4882	0.4581	0.4753	-
Cloud Model: SD3	0.6337	0.6847	0.6346	0.6703	0.5930	0.5868	0.5134	0.5199	0.5345	0.4972	-
Random	0.6499	0.6943	0.6208	0.6601	0.6022	0.5945	0.4848	0.5041	0.4963	0.4862	10.00
RouteLLM-BERT	0.6534	0.6917	0.6341	0.6641	0.6062	0.6104	0.4974	**0.5115**	0.5050	0.4911	36.93
RouteLLM-MF	0.6545	0.6918	0.6348	0.6654	0.6058	0.6121	0.4970	0.5113	0.5049	0.4927	39.71
HybridLLM	0.6527	**0.6932**	0.6339	0.6648	0.6075	0.6119	**0.4991**	0.5105	**0.5083**	0.4938	41.10
ZOOTER	**0.6552**	0.6930	0.6344	0.6642	0.6068	0.6124	0.4976	0.5103	0.5064	0.4921	41.10
`RouteT2I` (Ours)	0.6543	0.6924	**0.6353**	**0.6674**	**0.6096**	**0.6136**	0.4984	0.5104	0.5081	**0.4940**	**45.13**

Table 15. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. Pixart-α is used as the edge model, while SD3 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SD2.1	0.6251	0.6685	0.6076	0.6537	0.5949	0.5575	0.4680	0.5088	0.4860	0.4690	-
Cloud Model: PixArt-α	0.6660	0.7039	0.6069	0.6499	0.6114	0.6022	0.4562	0.4882	0.4581	0.4753	-
Random	0.6456	0.6862	0.6073	0.6518	0.6032	0.5798	0.4621	0.4985	0.4720	0.4721	0.00
RouteLLM-BERT	0.6524	0.6874	0.6183	0.6612	0.6088	0.5926	0.4738	0.5049	0.4812	0.4752	212.61
RouteLLM-MF	0.6525	0.6878	0.6175	0.6618	0.6087	0.5921	0.4741	0.5049	0.4813	0.4757	204.33
HybridLLM	**0.6536**	0.6880	0.6172	0.6609	0.6084	0.5926	0.4741	0.5058	0.4819	0.4757	197.92
ZOOTER	0.6520	**0.6881**	0.6174	**0.6621**	0.6096	0.5916	0.4732	**0.5064**	0.4822	0.4755	203.06
`RouteT2I` (Ours)	0.6525	0.6868	**0.6193**	**0.6621**	**0.6102**	**0.5937**	**0.4757**	0.5052	**0.4831**	**0.4763**	**235.22**

Table 16. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SD2.1 is used as the edge model, while Pixart-α serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SDXL	0.5900	0.6564	0.5645	0.5832	0.6042	0.5660	0.4622	0.5282	0.4866	0.4892	-
Cloud Model: PixArt-α	0.6660	0.7039	0.6069	0.6499	0.6114	0.6022	0.4562	0.4882	0.4581	0.4753	-
Random	0.6280	0.6801	0.5857	0.6165	0.6078	0.5841	0.4592	0.5082	0.4724	0.4822	10.00
RouteLLM-BERT	0.6354	0.6804	0.5940	0.6244	0.6039	0.5933	0.4657	0.5110	0.4729	0.4809	22.09
RouteLLM-MF	0.6354	0.6807	0.5977	**0.6268**	0.6076	0.5962	0.4685	0.5125	0.4792	0.4827	37.97
HybridLLM	0.6356	**0.6814**	0.5962	0.6239	0.6093	0.5956	0.4675	0.5120	0.4793	0.4828	37.81
ZOOTER	0.6350	0.6804	0.5942	0.6241	**0.6116**	0.5940	0.4668	0.5115	0.4787	0.4836	38.84
`RouteT2I` (Ours)	**0.6360**	0.6805	**0.5981**	0.6256	0.6101	**0.5973**	**0.4695**	**0.5130**	**0.4801**	**0.4837**	**44.34**

Table 17. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SDXL is used as the edge model, while Pixart-α serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: XL-Refiner	0.6083	0.6584	0.5835	0.6013	0.6086	0.5773	0.4701	0.5195	0.4824	0.4850	-
Cloud Model: PixArt-α	0.6660	0.7039	0.6069	0.6499	0.6114	0.6022	0.4562	0.4882	0.4581	0.4753	-
Random	0.6371	0.6811	0.5952	0.6256	0.6100	0.5898	0.4632	0.5039	0.4702	0.4801	10.00
RouteLLM-BERT	0.6454	0.6812	**0.6075**	0.6341	0.6149	0.6031	0.4738	**0.5099**	0.4793	0.4813	55.85
RouteLLM-MF	0.6453	0.6806	0.6062	0.6342	0.6131	0.6031	0.4738	0.5097	0.4793	0.4809	48.20
HybridLLM	**0.6458**	0.6805	0.6063	**0.6345**	0.6132	0.6034	0.4737	**0.5099**	0.4784	0.4814	48.98
ZOOTER	0.6449	**0.6814**	0.6054	**0.6345**	0.6138	0.6007	0.4719	0.5094	0.4781	0.4823	48.84
`RouteT2I` (Ours)	0.6434	0.6793	0.6058	0.6328	**0.6165**	**0.6040**	**0.4749**	0.5093	**0.4799**	**0.4830**	**62.41**

Table 18. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. XL-Refiner is used as the edge model, while Pixart-α serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: SD1.5	0.6035	0.6615	0.5912	0.6454	0.5944	0.5444	0.4659	0.5149	0.4945	0.4730	-
Cloud Model: PixArt-α	0.6660	0.7039	0.6069	0.6499	0.6114	0.6022	0.4562	0.4882	0.4581	0.4753	-
Random	0.6347	0.6827	0.5991	0.6477	0.6029	0.5733	0.4611	0.5016	0.4763	0.4741	20.00
RouteLLM-BERT	0.6382	**0.6844**	0.6094	0.6580	0.6082	0.5840	0.4701	**0.5100**	0.4830	0.4781	87.13
RouteLLM-MF	0.6384	0.6837	0.6093	0.6588	0.6106	0.5854	0.4710	0.5095	0.4849	0.4776	89.28
HybridLLM	0.6395	0.6837	0.6085	0.6573	0.6107	**0.5861**	0.4717	0.5091	0.4856	0.4780	88.25
ZOOTER	**0.6400**	0.6839	0.6086	0.6587	**0.6108**	0.5842	0.4711	0.5075	0.4860	0.4781	90.87
`RouteT2I` (Ours)	0.6393	0.6830	**0.6104**	**0.6589**	**0.6108**	0.5859	**0.4723**	0.5068	**0.4861**	**0.4785**	**95.00**

Table 19. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. SD1.5 is used as the edge model, while Pixart-α serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: Infinity	0.5329	0.6810	0.5577	0.5817	0.6222	0.5817	0.4587	0.4926	0.4766	0.5009	-
Cloud Model: SD3	0.6337	0.6847	0.6346	0.6703	0.5930	0.5868	0.5134	0.5199	0.5345	0.4972	-
Random	0.5833	0.6828	0.5961	0.6260	0.6076	0.5843	0.4860	0.5063	0.5055	0.4990	30.00
RouteLLM-BERT	0.5879	**0.6866**	0.6031	**0.6413**	0.6154	0.5947	0.4929	**0.5140**	0.5117	0.5017	78.93
RouteLLM-MF	0.5899	0.6849	0.6036	**0.6413**	0.6156	0.5962	0.4955	0.5135	0.5135	0.5026	80.57
HybridLLM	0.5932	0.6837	0.6093	0.6370	0.6196	0.6028	**0.5031**	0.5074	0.5210	0.5035	95.28
ZOOTER	0.5941	0.6853	0.6067	0.6365	0.6211	0.6030	0.5000	0.5087	0.5198	0.5025	97.54
`RouteT2I` (Ours)	**0.5947**	0.6854	**0.6095**	0.6384	**0.6218**	**0.6056**	0.5018	0.5095	**0.5212**	**0.5036**	**106.87**

Table 20. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. Infinity is used as the edge model, while SD3 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: Infinity	0.5329	0.6810	0.5577	0.5817	0.6222	0.5817	0.4587	0.4926	0.4766	0.5009	-
Cloud Model: SD2.1	0.6251	0.6685	0.6076	0.6537	0.5949	0.5575	0.4680	0.5088	0.4860	0.4690	-
Random	0.5790	0.6747	0.5827	0.6177	0.6086	0.5696	0.4633	0.5007	0.4813	0.4850	10.00
RouteLLM-BERT	0.5842	**0.6774**	0.5960	0.6313	0.6246	0.5872	0.4777	0.5043	0.4978	**0.4922**	67.89
RouteLLM-MF	0.5812	0.6758	0.5958	**0.6334**	0.6219	0.5839	0.4765	**0.5047**	0.4940	0.4921	59.12
HybridLLM	0.5888	0.6755	0.5946	0.6317	0.6238	0.5869	0.4774	0.5034	0.4956	0.4880	61.66
ZOOTER	0.5885	0.6764	0.5937	0.6314	0.6237	0.5861	0.4767	0.5045	0.4955	0.4884	61.68
`RouteT2I` (Ours)	**0.5889**	0.6760	**0.5980**	0.6332	**0.6279**	**0.5909**	**0.4803**	0.5040	**0.5001**	0.4899	**75.02**

Table 21. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. Infinity is used as the edge model, while SD2.1 serves as the cloud model.

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model: Infinity	0.5329	0.6810	0.5577	0.5817	0.6222	0.5817	0.4587	0.4926	0.4766	0.5009	-
Cloud Model: PixArt-α	0.6660	0.7039	0.6069	0.6499	0.6114	0.6022	0.4562	0.4882	0.4581	0.4753	-
Random	0.5995	0.6924	0.5823	0.6158	0.6168	0.5920	0.4574	0.4904	0.4673	0.4881	0.00
RouteLLM-BERT	0.6104	**0.6924**	0.5981	**0.6333**	0.6309	0.6185	0.4729	0.5002	0.4824	0.4926	127.82
RouteLLM-MF	0.6064	0.6909	0.5962	0.6329	0.6259	0.6162	0.4711	**0.5013**	0.4782	0.4928	113.45
HybridLLM	0.6088	0.6921	0.5979	0.6311	0.6304	0.6159	0.4735	0.4981	0.4827	0.4932	123.07
ZOOTER	**0.6105**	**0.6924**	0.5960	0.6312	0.6309	0.6169	0.4718	0.4985	0.4822	0.4918	117.25
`RouteT2I` (Ours)	0.6083	0.6910	**0.5983**	0.6294	**0.6327**	**0.6187**	**0.4743**	0.4978	**0.4853**	**0.4937**	**130.24**

Table 22. The multi-dimensional quality of images generated by text-to-image models, with the router selecting between edge and cloud models for each prompt at a 50% routing rate. The higher the metrics, the better. Infinity is used as the edge model, while Pixart-α serves as the cloud model.