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

Zewei Xin, Qinya Li, Chaoyue Niu, Fan Wu, Guihai Chen
Shanghai Jiao Tong University

Introduction

- The impressive capabilities of text-to-image (T2I) models often come at the cost of larger model sizes and higher computational expenses.

Text-to-Image Model	#Param	Pricing ($/M)
Stable Diffusion 1.6 [28]	0.86 B	9 K
Stable Diffusion XL [24]	2.6 B	9 K
Stable Diffusion 3 Medium [9]	2 B	35 K
Stable Diffusion 3 [9]	8 B	65 K
Stable Diffusion 3.5 [1]	8 B	65 K

Introduction

- However, not all user requests require the most powerful model.
- In some cases, smaller models can produce results that are comparable, or even superior.

SD2.1 SD3

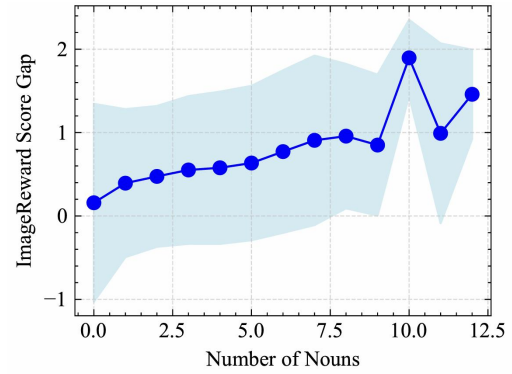
*There is a **dog** lying on the **ground**.*

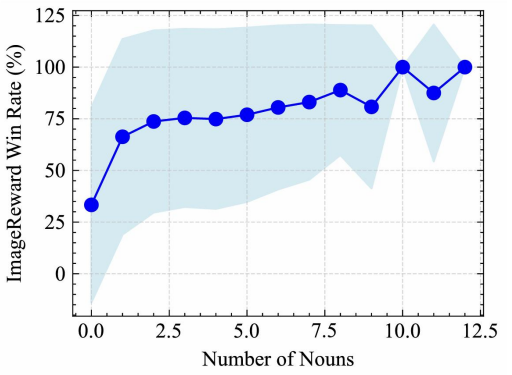
SD2.1 SD3

***Apples** are lined up in a wooden **box**.*

*A **cat** sitting motionless on a **television** stand behind a **mug** in the shape of a **skeleton head**.*

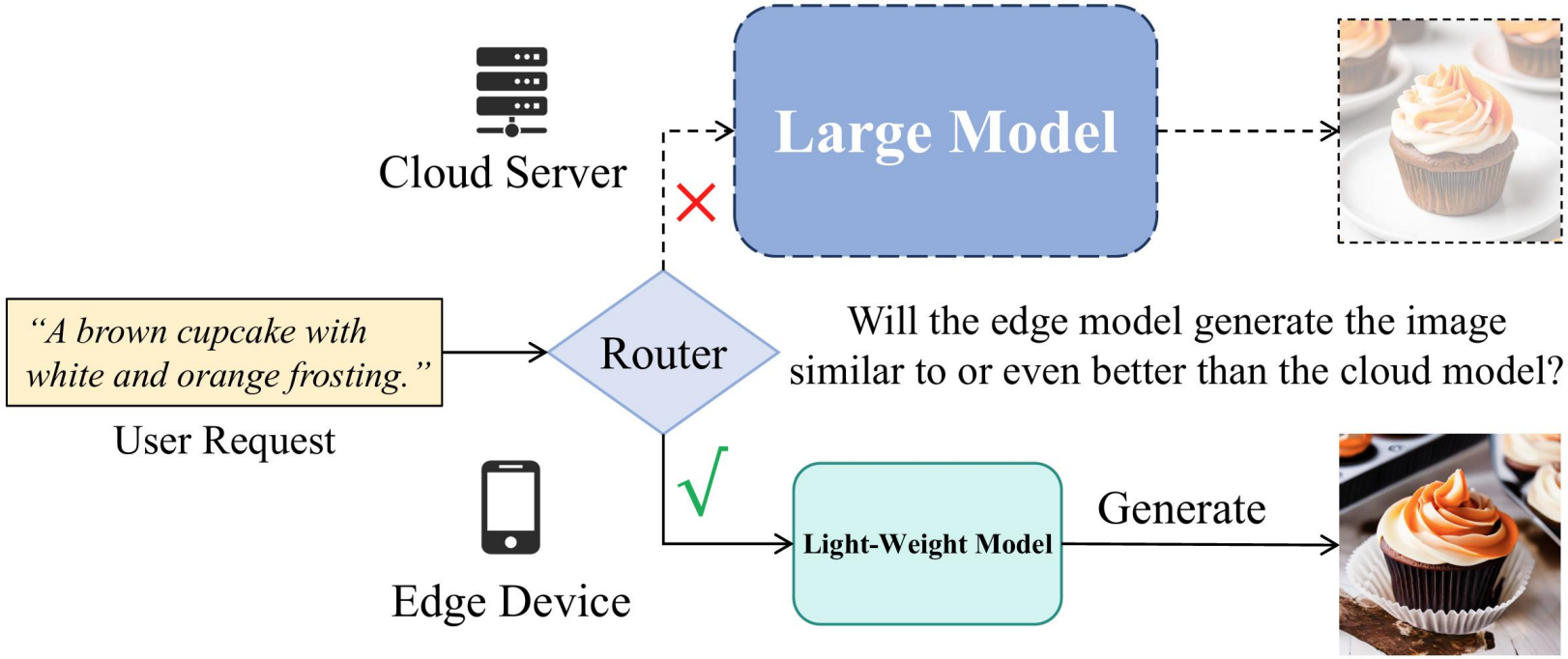
*A **sandwich** cut in half on a white **plate** with a side of **salad** and a **fork** and **knife** wrapped in a **napkin**.*





Introduction

- Therefore, a router that intelligently directs user requests between lightweight edge models and large-scale cloud models is essential.



Problem Formulation

- Our routing objective is to maximize generation quality while minimizing generation overhead.
- This overhead can be approximated by the proportion of requests routed to the cloud.

$$
\begin{aligned}
&\max\ R(\mathcal{X})\mathcal{Q}(\mathcal{I}_c) + (1 - R(\mathcal{X}))\mathcal{Q}(\mathcal{I}_e) \\
&\quad \text{s.t. } \mathbb{P}\{R(\mathcal{X}) = 1\} \cdot F_c \leq \tau_{fee} \\
&\qquad \mathbb{P}\{R(\mathcal{X}) = 1\} \cdot D_{\mathcal{M}_c} + \mathbb{P}\{R(\mathcal{X}) = 0\} \cdot D_{\mathcal{M}_e} \\
&\qquad + D_R \leq \tau_{time},
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
&\max\ R(\mathcal{X})\mathcal{Q}(\mathcal{I}_c) + (1 - R(\mathcal{X}))\mathcal{Q}(\mathcal{I}_e) \\
&\quad \text{s.t. } \mathbb{P}\{R(\mathcal{X}) = 1\} \leq \rho_r.
\end{aligned}
$$

Routing Objective

- We use multi-dimensional quality metrics to address the subjectivity of image quality.
- We rebalance quality trade-offs based on their relative distances.

$$q(I, m) = \sigma(CLIP(I, m^+) - CLIP(I, m^-)),$$

$$\mathcal{Q}(I) = [q(I, m_i)|i = 1, 2, \ldots, N],$$

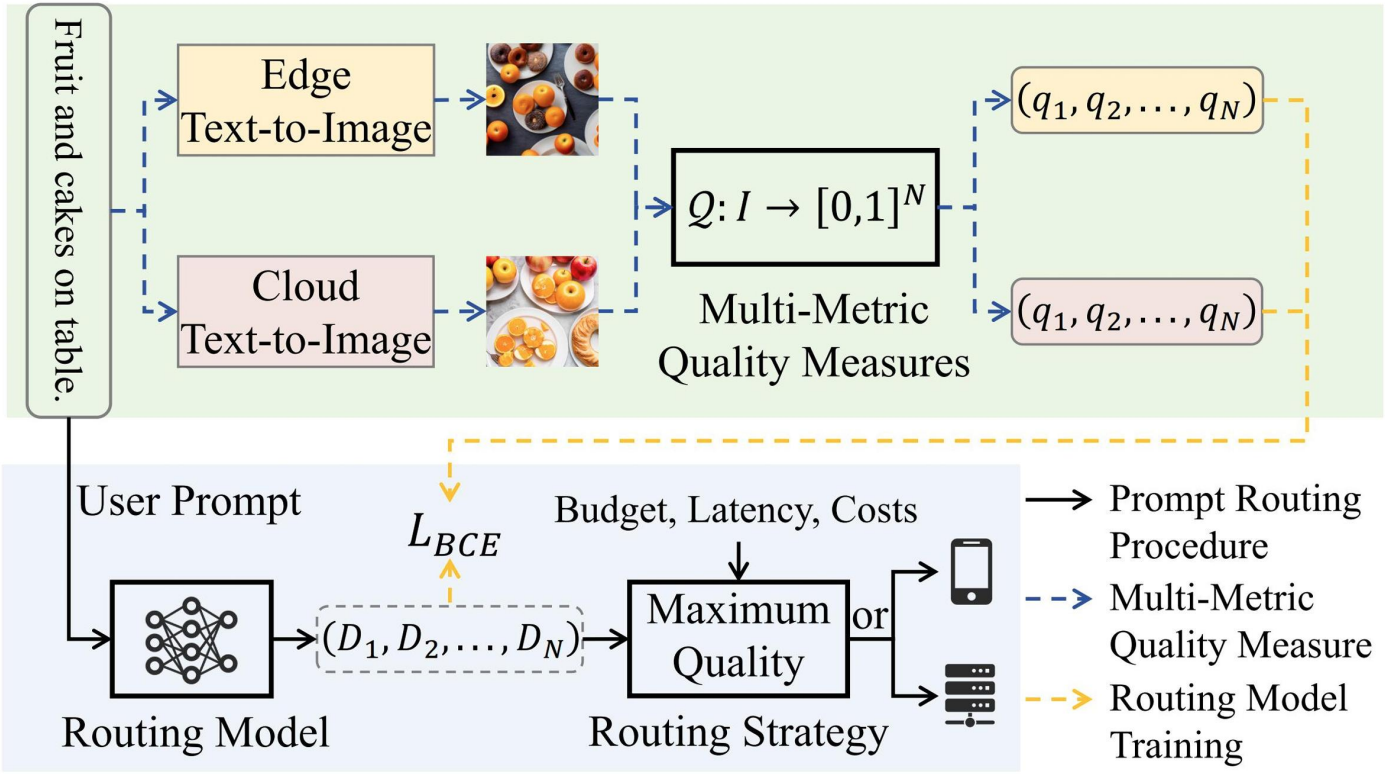
$$D_i(I_e, I_c) = \sigma\left(\frac{q(I_e, m_i) - q(I_c, m_i)}{\Gamma|\mu_i(\mathcal{I}_e) - \mu_i(\mathcal{I}_c)|}\right),$$

$$PRS(I_e, I_c) = \sum_{i=1}^{N} w_i D_i(I_e, I_c),$$

$$\max_{\alpha \leq 1/2} \mathbb{P}\{PRS(I_e, I_c) < \alpha \mid I_e, I_c \in \mathcal{I}_e, \mathcal{I}_c\} \leq \rho_r.$$

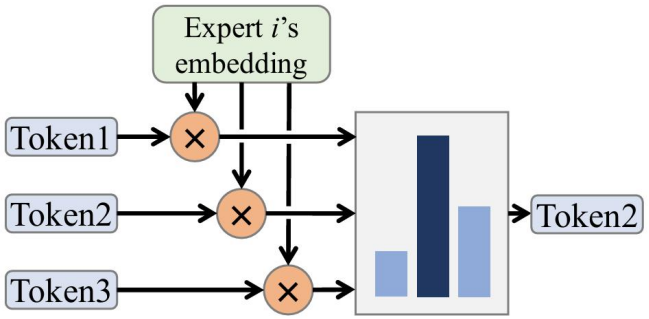
Overview of RouteT2I

- Overview of RouteT2I.

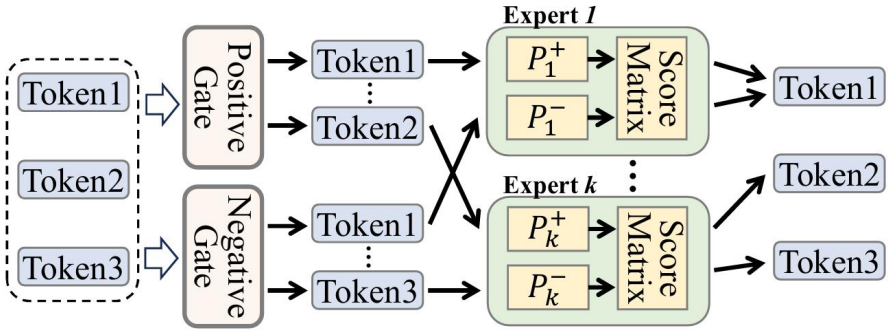


Routing Model

- Architecture of a dual-gate MoE that uses the token selection gate as the gate network.
- The routing model replaces the standard FFN in Transformer with a dual-gate MoE.



(a) Token selection gate.



(b) Dual-gate MoE.

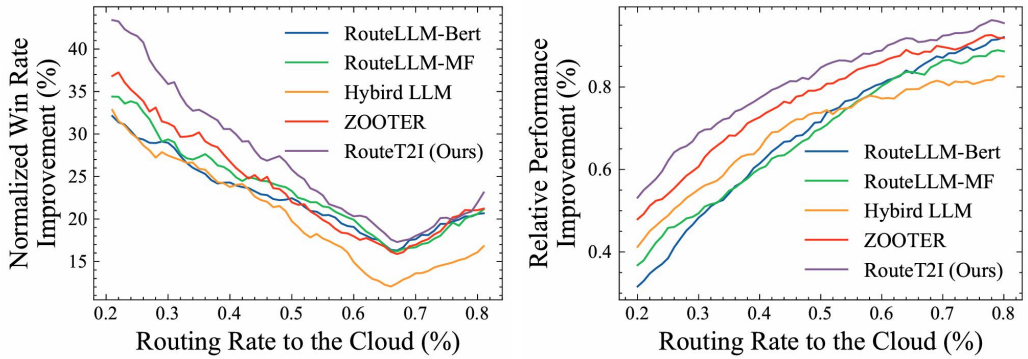
Experiment

- Visualization of our RouteT2I selection results

Experiment

- Comparisons with State-of-the-Art Methods

Router	Image Quality Metrics										$\Delta P(\%)$
	Definition	Detail	Clarity	Sharpness	Harmony	Realism	Color	Consistency	Layout	Integrity	
Edge Model	0.6251	0.6685	0.6076	0.6537	0.5949	0.5575	0.4680	0.5088	0.4860	0.4690	-
Cloud Model	0.6337	0.6847	0.6346	0.6703	0.5930	0.5868	0.5134	0.5199	0.5345	0.4972	-
Random	0.6294	0.6766	0.6211	0.6620	0.5939	0.5721	0.4907	0.5144	0.5102	0.4831	40.00
RouteLLM-BERT [23]	0.6347	0.6792	0.6305	0.6651	0.5960	0.5788	0.4982	0.5160	0.5167	**0.4866**	71.51
RouteLLM-MF [23]	**0.6364**	**0.6814**	0.6299	0.6660	0.5952	0.5776	0.4970	0.5164	0.5149	0.4850	69.90
Hybrid LLM [8]	0.6327	0.6784	0.6306	0.6677	0.5964	0.5787	0.5008	0.5161	**0.5191**	0.4864	73.49
ZOOTER [21]	0.6350	0.6796	0.6315	0.6672	0.5966	0.5788	0.5004	0.5166	0.5179	0.4854	77.95
RouteT2I (Ours)	0.6350	0.6786	**0.6318**	**0.6679**	**0.5975**	**0.5804**	**0.5010**	**0.5167**	0.5189	0.4865	**83.97**



Experiment

- Cost saving under a given performance target

Router \ ΔP	40%	50%	60%	70%	80%
RouteLLM-BERT [23]	56.15	51.39	46.92	42.70	40.21
RouteLLM-MF [23]	48.86	49.90	48.20	44.89	41.50
Hybrid LLM [8]	62.06	58.85	53.63	49.92	33.38
ZOOTER [21]	69.28	65.76	60.81	57.35	49.64
`RouteT2I` (Ours)	**71.81**	**70.24**	**66.61**	**60.01**	**53.53**

Thanks for listening!