IEEE ICC 2025

Optimal Energy-Delay Tradeoff for Mobile Edge Generation (MEG)

Xiaoxia Xu, Xidong Mu, Yuanwei Liu, and Arumugam Nallanathan

Queen Mary University of London (QMUL), U.K

Queen’s University Belfast, U. K.

The University of Hong Kong, Hong Kong

Background: What is LAM?

Large generative model (LAM) is reshaping the way to develop AI in all industries

Human-computer dialogue by ChatGPT

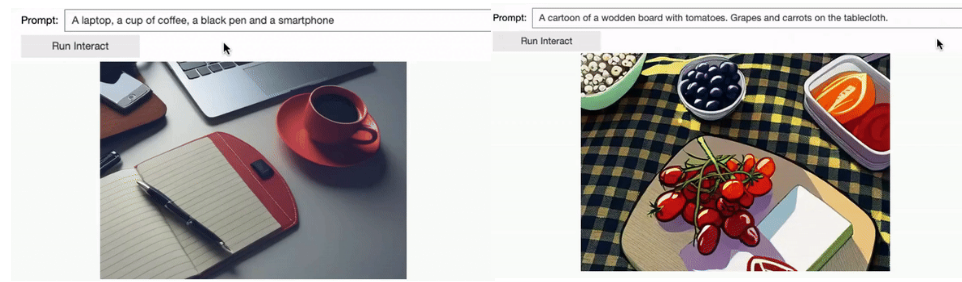


Image Generation & Editing by Nvidia

Key Concept & Advantage

- **LGM: large-scale** neural networks (NNs) with **human-like AI-generated contents (AIGC)** capabilities
- Built to **understand & create multimodal contents** across text, images, audios/music, videos, point clouds, etc. like human **experts**
- Innovatively support **smart human-computer dialogue/interaction**, infeasible in the past

Background: LAM vs. Conventional AI

Popular LAMs

Large language models (LLMs)

ChatGPT, deepseek, Meta Llama 3 …

Multi-modal visual generative model

Stable Diffusion, DALL·E, SORA …

LAM

- Generative AI (GAI) paradigm
- Model scale: billions/trillions of parameters
- Use case: AIGC/AI-generated everything (AIGX)
- Multimodal: text, image, video, audio …
- Inference latency: high (even for generating a single image/video)

Conventional AI Models

- Discriminative AI (DAI) paradigm
- Model scale: <100 millions parameters
- Use case: classification & regression
- Multimodal: lack of support
- Inference latency: low, millisecond-level/batch

Why Mobile Edge Generation (MEG)

Current AIGC: Centralized Generation

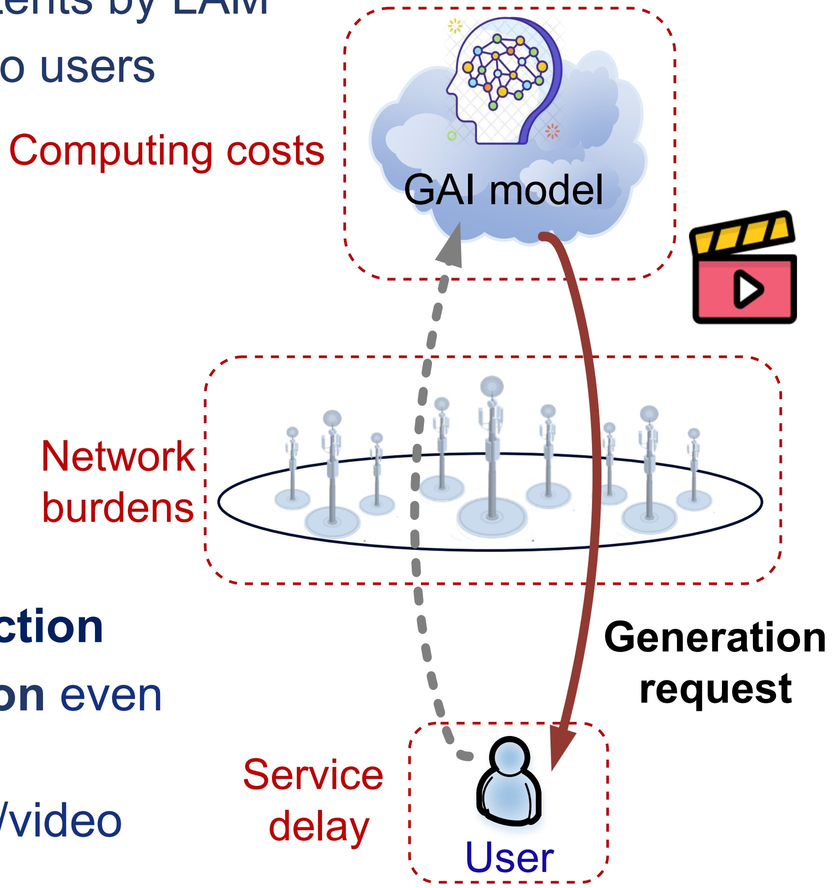
- **Upload:** Users upload request (e.g., prompt) & local data
- **Generate:** Cloud/edge server generates contents by LAM
- **Download:** Transmit large-volume contents to users

Limitations

- Large **communication** overheads for AIGC downloads
- High **delay** for massive user scenarios
- Huge **computing** and storage costs for LAM
- Privacy leakage risks

Cannot offer intensive human-machine interaction

- **Example: DeepSeek experiences congestion** even for text generation
- **Resource scarcity** is more serious for image/video generation



(Centralized Generation)

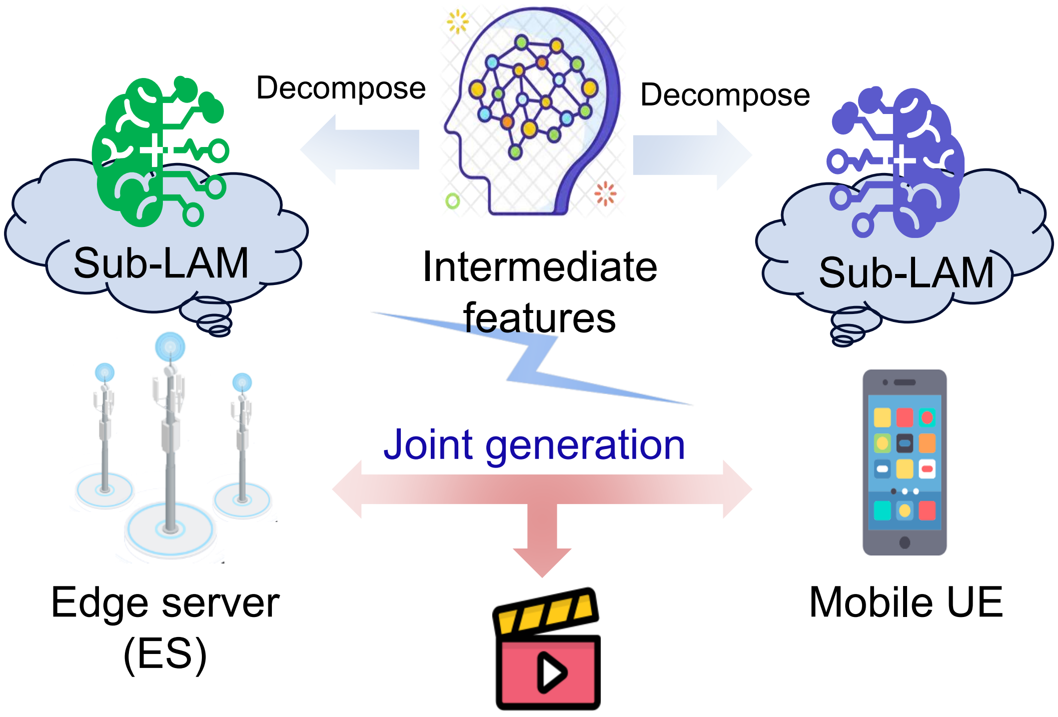
Motivation of Mobile Edge Generation (MEG)

Key idea of MEG

- Decompose LAM into distributed sub-models on different edge devices (ESs, UEs)
- 6G-connected ES-UE co-generation

Advantage

- Transmit low-dimension features, rather than large-volume generated contents
- Affordable computing complexity for mobile devices
- Enhance privacy and data security by feature coding
- Customized local LAM according to user preference



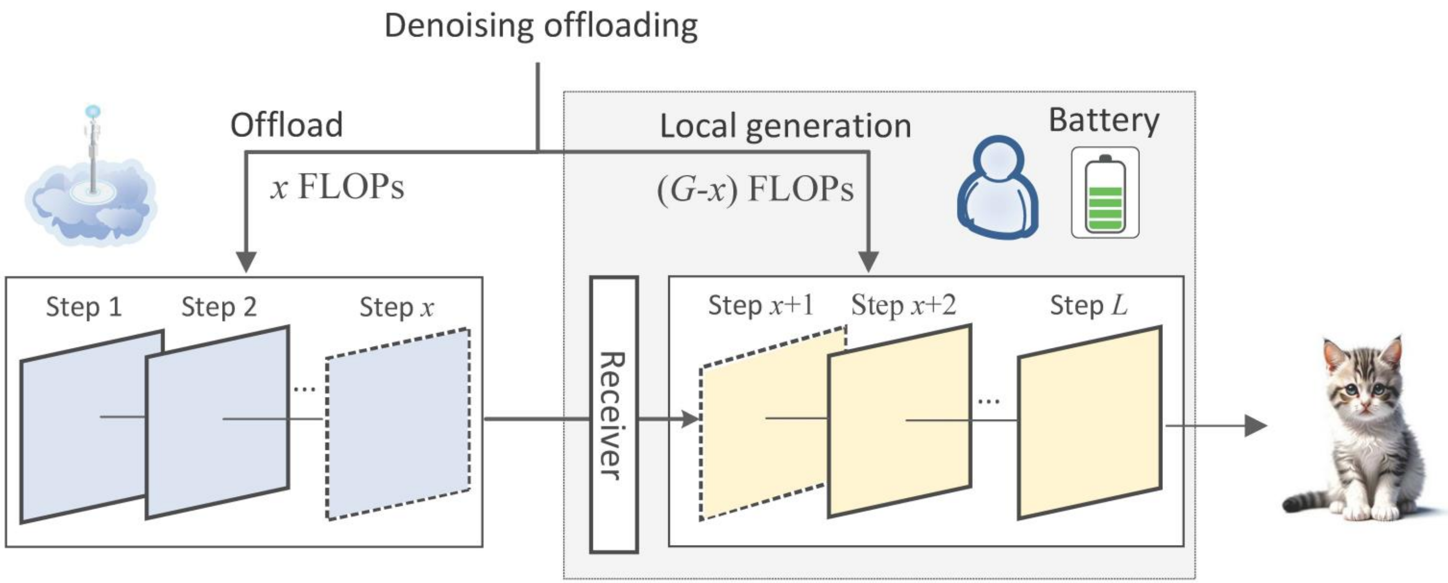
Key problem of MEG

- Low-latency collaborative generation mechanism?
- How to balance between latency and mobile energy consumption?

MEG: From Model Split to Generation Split

(Conventional) Model split based MEG

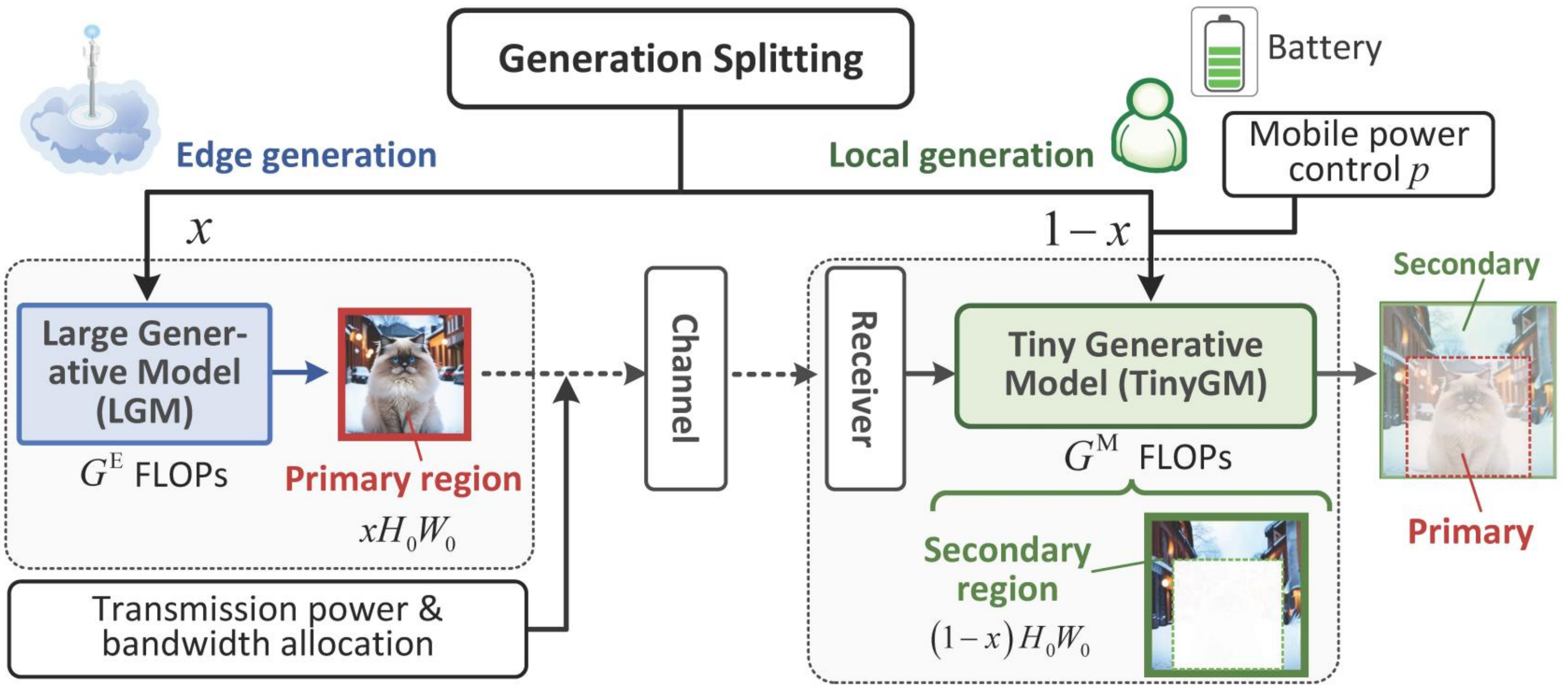
- Split LAM into two sub-models at specific neural layer
- **Pros**
  (i) Transmit intermediate features -> low transmission latency
  (ii) Achieve same quality with centralized generation
- **Cons**
  (i) Generation costs on edge/mobile devices & trans. cost inherently depend on predefined LGM structure (hidden layers) -> limits the scheduling flexibility
  (ii) Same complexity over the entire content -> ignore semantic difference



MEG: From Model Split to Generation Split

(Our proposed) Generation split based MEG

- Split generated content into primary region and secondary region
- **Exploit different model complexities and devices** over different regions
  - Edge: high-complexity LGM -> generate primary region & transmit to mobile
  - Mobile: TinyGM -> complete second region
- Mobile generation cost is **fully controllable** by **generation splitting ratio**



System Model

- **Latency model**

Mobile computing power

Mobile generation delay

$$\boxed{D^{tot}(x,p)} = \boxed{D^{\mathrm{E}}(x,p)} + \boxed{D^{\mathrm{M}}(x,p)} + \boxed{xZ^{\mathrm{bit}} / R(h)}$$

Generation splitting ratio Edge generation delay Transmission delay

$$\boxed{D^{\mathrm{M}}(x,p)} = (1-x)G^{\mathrm{M}} / \boxed{\mathrm{F}^{\mathrm{M}}(p)} \qquad \mathrm{F}^{M}(p) = v_0 \left(\frac{p}{\overline{P}}\right)^{\kappa} N_{core} N_{OPC}$$

Number of FLOPs for TinyGM FLOPS at mobile

Here, we ignore the fixed FLOPs to acquire coarse features of full region.
A more accurate model is considered in journal version

- **Mobile energy consumption**

$$E^{M}(x,p) = p(1-x)G^{M} / \mathrm{F}^{M}(x,p)$$

Joint Generation Splitting and Mobile Power Control

Multi-objective programming (MOP)

- Minimize both generation delay and mobile energy consumption

$$
\begin{aligned}
\text{(P1)}\quad & \min_{x,p} D^{tot}(x,p) = xU^{\mathrm{E}} + (1-x)U^{\mathrm{M}}p + xZ^{\mathrm{bit}}/R(h) \\
& \min_{x,p} E^{\mathrm{M}}(x,p) = (1-x)pG^{\mathrm{M}}/F^{\mathrm{M}}(p) \\
\text{s.t.}\quad & \underline{x} \le x \le \overline{x} \\
& 0 \le p \le P^{\mathrm{M}}
\end{aligned}
$$

unit generation cost at mobile

$$U^{M} = V_0 \cdot (p)^{-\kappa} = v_0 \left(\frac{p}{\overline{P}}\right)^{\kappa} N_{core} N_{OPC}$$

unit generation cost at the edge

$$U^{\mathrm{E}} = V_0 \cdot \left(P^{\mathrm{E}}\right)^{-\kappa}$$

we set $\overline{x} = 1$ by assuming a large computation capacity at the edge

- Transformation to single-objective programming (SOP) by ϵ-constraint method

$$
\begin{aligned}
\text{(P2)}\quad & \min_{x,p} D^{tot}(x,p) = xU^{\mathrm{E}} + (1-x)V_0 p^{-\kappa} + xZ^{\mathrm{bit}}/R(h) \\
\text{s.t.}\quad & E^{\mathrm{M}}(x,p) = (1-x)V_0 p^{1-\kappa} \\
& \underline{x} \le x \le \overline{x}, \quad 0 \le p \le P^{\mathrm{M}}
\end{aligned}
$$

Optimal Solution

Closed-form Optimal Solution

$$p^*(x,\ ) = \min\left\{ P^{\mathrm{M}}, \left(\frac{\ }{(1-x)V_0} \right)^{\frac{1}{1-\kappa}} \right\}$$

$$x^*(\ ) = \begin{cases} \max\left\{\underline{x}, x_0^*\right\}, & \text{if } U_{\min}^{\mathrm{M}} \quad (1-\kappa)\left(U^{\mathrm{E}} + \dfrac{Z^{\mathrm{bit}}}{R(h)}\right), \\ \max\left\{\underline{x}, x_1^*\right\}, & \text{if } (1-\kappa)\left(U^{\mathrm{E}} + \dfrac{Z^{\mathrm{bit}}}{R(h)}\right) < U_{\min}^{\mathrm{M}} \quad U^{\mathrm{E}} + \dfrac{Z^{\mathrm{bit}}}{R(h)}, \\ 1, & \text{if } U_{\min}^{\mathrm{M}} > U^{\mathrm{E}} + \dfrac{Z^{\mathrm{bit}}}{R(h)}, \end{cases} \qquad \left(U_{\min}^{\mathrm{M}} = V_0 \cdot \left(P^{\mathrm{M}}\right)^{-\kappa} \right)$$

Boundary of Pareto-optimal energy-delay (E-D) region

Compute optimal E-D objectives $\Rightarrow$ $$\mathcal{R}_{\mathrm{E-D}} = \bigcup_{\substack{\forall \epsilon \in \mathcal{E}, \\ \forall \underline{x} \leqslant x \leqslant \overline{x}}} \left\{ (E, D) : \ E \leqslant E^*(\epsilon), D \geqslant D^*(\epsilon) \right\}.$$

Optimal Solution

Lemma 2: Superiority over conventional fully edge-based generation

Given a fixed mobile power control, MEG always leads to equivalent or reduced latency compared to the fully edge-based generation, and the performance gain is given by

$$\Delta_{\mathrm{MEG}} = \frac{}{\eta^{\mathrm{M}} G^{\mathrm{M}}} \left[\frac{Z^{\mathrm{bit}}}{R(h)} + U^{\mathrm{E}} - U^{\mathrm{M}} \right]^{+}$$

- MEG strictly outperforms fully edge-based generation if channel state h and generative time costs U^{M} and U^{E} satisfy $U^{\mathrm{M}} \leq \frac{Z^{\mathrm{bit}}}{R(h)} + U^{\mathrm{E}}$. In this case, there is a tradeoff between mobile energy consumption and generation delay.
- Otherwise, if $U^{\mathrm{M}} > \frac{Z^{\mathrm{bit}}}{R(h)} + U^{\mathrm{E}}$, MEG reduces to the fully edge-based generation, and $x = 1$ is the unique minimizer of both E-D objective functions.

The performance gain of MEG over the fully edge-based generation increases as SNR or mobile generative time cost of TinyGM decreases. -> We can ensure the performance gain by deploying sufficiently lightweight TinyGM on mobile devices.

Simulation Result

Baseline

- Fully edge generation (FEG): a centralized SDXL/SD3 is deployed at the ES
- Model split (MS): SDXL is split into two parts. (i) VAE encoder + latent diffusion at the edge; (ii) VAE decoder at the mobile.

  [Note] MS generates the full-region latent feature at the edge, which is thus termed *FEG-Latent* in the paper.

Implementation of the proposed MEG framework

- Edge LGM: a 25-step standard SDXL/SD3 performs diffusion to generate primary region in latent space -> transmit primary-region latent features
- Mobile TinyGM: complete secondary region by a coarse-to-refine pipeline

  (i) Coarse module: generate coarse full-region by a downsampled & distilled 2-step SDXL

  (ii) Refinement: refine secondary region by lightweight Restormer

A demo of the above implementation is publicly available at https://github.com/xiaoxiaxusummer/MEGSplitting

Simulation Result

Prompt: A girl standing in the snowy street

Scheme	MEG, x=0.4	MEG, x=0.5	MEG, x=0.6	MEG, x=0.7	MEG, x=0.8	MEG, x=0.9	MS/FEG (baseline)
Latency (s)	2.0812	2.4722	2.8631	3.2540	3.6449	4.0358	4.1405/11.1120
FID score	0.7356	0.2516	0.2977	0.0638	0.0354	0.0129	Reference
SSIM	0.6366	0.6925	0.7624	0.8254	0.8882	0.9476	Reference

Samples generated by SDXL using different x

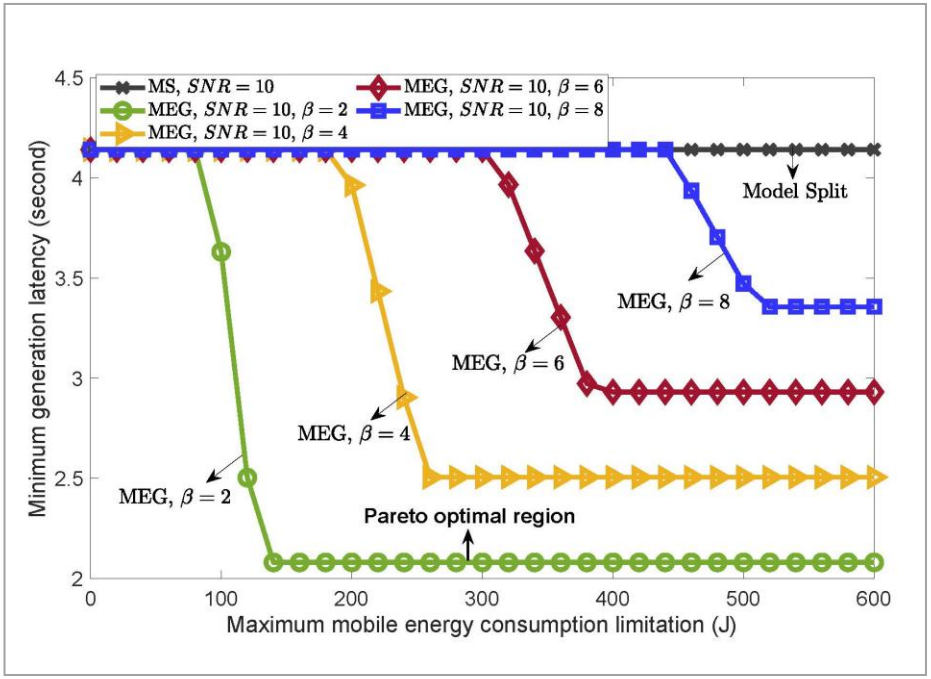
Scheme	MEG, x=0.4	MEG, x=0.5	MEG, x=0.6	MEG, x=0.7	MEG, x=0.8	MEG, x=0.9	MS/FEG (baseline)
Latency (s)	2.6948	3.2391	3.7834	4.3277	4.8720	5.4163	5.6744/12.6458
FID score	1.5303	1.0615	0.7984	0.7925	0.3833	0.0668	Reference
SSIM	0.6401	0.7100	0.7669	0.8243	0.8758	0.9381	Reference

Samples generated by SD3 using different x

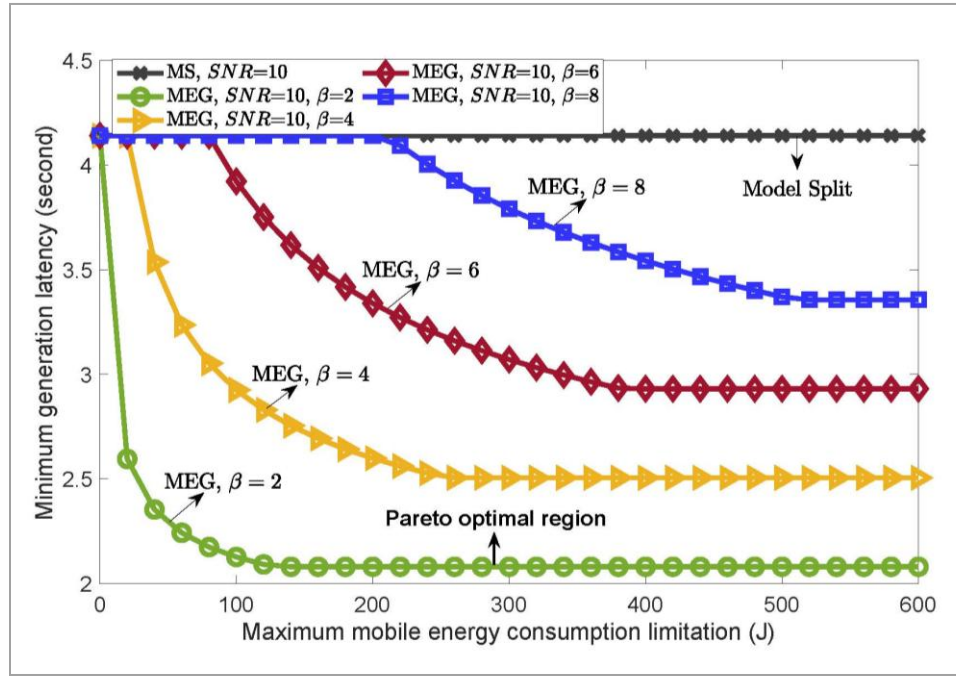
A demo to reproduce the generated samples is publicly available at https://github.com/xiaoxiaxusummer/MEGSplitting

Simulation Result

- Pareto E-D region boundary



Fixed mobile power



Optimal mobile power

- Generation delay ↓ as mobile energy consumption limitation ↑, demonstrating the E-D performance tradeoff in MEG.
- MEG reduces to MS when $\epsilon = 0$, achieves a lower generation latency than MS when $\epsilon > 0$. Performance gain ↑ as β ↓ (i.e., mobile generative time cost ↓)

Conclusion

- We proposed a novel MEG framework that enables flexible generation splitting between the edge and mobile devices -> reduce latency for edge-mobile co-generation of high-definition image
- An edge LDM and a mobile TinyGM are exploited to generate primary and secondary regions, enabling different model complexities for different contents.
- We formulated a multi-objective joint generation splitting and mobile power control problem, which simultaneously minimizes delay and mobile energy consumption. We derived Pareto-optimal solutions and E-D region boundary.
- Simulation results verified the superiority of the proposed MEG framework over centralized generation and model split schemes.

Future Direction

- Latency-distortion tradeoff based on the proposed framework
- Personalized primary region identification
- Generation splitting over noisy wireless channels
- Generation splitting in resource-constrained edge settings
- Security/privacy-aware generation splitting

Thank you

Email: x.xiaoxia@qmul.ac.uk