Cross-Modal Generative Semantic Communications for Mobile AIGC: Joint Semantic Encoding and Prompt Engineering

Yinqiu Liu, Hongyang Du, *Graduate Student Member, IEEE*, Dusit Niyato, *Fellow, IEEE*, Jiawen Kang, *Senior Member, IEEE*, Zehui Xiong, *Senior Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, Ping Zhang, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Employing massive Mobile AI-Generated Content (AIGC) Service Providers (MASPs) with powerful models, high-quality AIGC services become accessible for resource-constrained end users. However, this advancement, referred to as mobile AIGC, also introduces a significant challenge: users should download large AIGC outputs from the MASPs, leading to substantial bandwidth consumption and potential transmission failures. In this paper, we apply cross-modal Generative Semantic Communications (G-SemCom) in mobile AIGC to overcome wireless bandwidth constraints. Specifically, we utilize cross-modal attention maps to indicate the correlation between user prompts and each part of AIGC outputs. In this way, the MASP can analyze the prompt context and filter the most semantically important content efficiently. Only semantic information is transmitted, with which users can recover the entire AIGC output with high quality while saving mobile bandwidth. Since the transmitted information not only preserves the semantics but also prompts the recovery, we formulate a joint semantic encoding and prompt engineering problem to optimize the bandwidth allocation among users. Particularly, we present a human-perceptual metric named Joint Perceptual Similarity and Quality (JPSQ), which is fused by two learning-based measurements regarding semantic similarity and aesthetic quality, respectively. Furthermore, we develop the Attention-aware Deep Diffusion (ADD) algorithm, which learns attention maps and leverages the diffusion process to enhance the environment exploration ability of traditional deep reinforcement learning (DRL). Extensive experiments demonstrate that our proposal can reduce the bandwidth consumption of mobile users by 49.4% on average, with almost no perceptual difference in AIGC output quality. Moreover, the ADD algorithm shows superior performance over baseline DRL methods, with 1.74× higher overall reward.**

***Index Terms*—Cross-Modal attention, diffusion, generative semantic communications, mobile AIGC.**

Received 20 December 2023; revised 18 June 2024; accepted 20 August 2024. Date of publication 26 August 2024; date of current version 5 November 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62102099 and Grant U22A2054, in part by Guangzhou Basic Research Program under Grant 2023A04J1699, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A151514 0137. This research was also supported in part by the National Research Foundation, Singapore, in part by Infocomm Media Development Authority under its Future Communications Research and Development Programme, in part by Defence Science Organisation (DSO) National Laboratories under the AI Singapore Programme under Grant FCP-NTU-RG-2022-010 and under Grant FCP-ASTAR-TG-2022-003, in part by the Singapore Ministry of Education (MOE) Tier 1 under Grant RG87/22, in part by the NTU Centre for Computational Technologies in Finance (NTU-CCTF). The research was also supported in part by SUTD SRG-ISTD-2021-165, in part by SUTD-ZJU IDEA under Grant SUTD-ZJU (VP) 202102, in part by the Ministry of Education, Singapore, under its SMU-SUTD Joint under Grant 22-SIS-SMU-048, and in part by SUTD Kickstarter Initiative under Grant SKI 20210204. The work of Shiwen Mao was supported in part by NSF under Grant CNS-2148382. Recommended for acceptance by J. Ren. *(Corresponding author: Jiawen Kang.)*

Yinqiu Liu and Dusit Niyato are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yinqiu001@e.ntu.edu.sg; dniyato@ntu.edu.sg).

Hongyang Du is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Pok Fu Lam, Hong Kong (e-mail: duhy@eee.hku.hk).

Jiawen Kang is with the School of Automation, Guangdong University of Technology, Guangdong 510006, China (e-mail: kavinkang@gdut.edu.cn).

Zehui Xiong is with the Pillar of Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372 (e-mail: zehuixiong@sutd.edu.sg).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

Ping Zhang is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: pzhang@bupt.edu.cn).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TMC.2024.3449645

I. Introduction

As THE latest paradigm for content creation, AI-Generated Content (AIGC) [1], [2] has attracted great attention from both academia and industry. Recently, we have witnessed the phenomenal success of AIGC in various fields, such as Stable Diffusion and DALL-E· 3in text-to-image generation, ChatGPT in Q & A, and MusicLM in music composition [3]. Nonetheless, the strong power of AIGC models relies on extremely large neural networks with billions of parameters. For instance, DALL-E· 2 and GPT-3 contain 3.5 and 175 billion parameters, respectively [3]. Moreover, considering the difficulty of generating high-dimensional content, such as images and videos, each round of generative inference costs considerable power. Such resource-intensive features severely hinder the further application of AIGC, especially in mobile/edge scenarios with resource constraints.

To overcome resource limitations and provide ubiquitous high-quality AIGC services, researchers sought help from mobile-edge computing and presented the concept of *Mobile AIGC* [1]. Specifically, massive end users can offload their AIGC tasks to Mobile AIGC Service Providers (MASPs), e.g., base stations. With abundant computing resources to operate

AIGC models, the MASP can provide paid AIGC inference services according to users' task description/input, so-called prompts. In this way, users can receive high-quality AIGC outputs while circumventing hefty computation costs on their mobile devices. Recently, a series of breakthroughs regarding optimizing AIGC models and managing mobile AIGC networks have been proposed. For instance, Qualcomm published the world's first on-device Stable Diffusion [4]. Chen et al. [5] performed GPU-aware optimization on large diffusion models, accomplishing fast text-to-image AIGC on mobile-edge servers and devices. From the network perspective, Du et al. [6] and Wen et al. [7] scheduled the task allocation between users and the MASPs and designed the incentive mechanism for mobile AIGC, respectively.

Despite the achievements that have been made, the existing works ignore the bandwidth consumption of mobile devices. We observe that mobile AIGC just reduces users' computation overhead at the expense of increasing bandwidth consumption since users should download large AIGC outputs from the MASP after each round of inference. Hence, two challenges exist in the current paradigm.

- *Modality Transfer during AIGC Inference:* AIGC inference generally involves generating high-dimension information from low-dimension prompts, e.g., generating images (hundreds of KBs) from texts (hundreds of bytes). Such modality transfers might cause failed transmission if large AIGC outputs block the downlink channel. Incomplete or damaged AIGC outputs are less useful to users and downstream applications.
- *Contradiction between Generation Quality and Bandwidth Consumption:* The higher the quality of AIGC outputs, typically, the larger their sizes, and the more bandwidth is required for transmission. Therefore, if encountering transmission failure, users need to adjust and/or reduce their requirements for the quality of AIGC outputs and ask the MASP for regeneration. Such a contradiction prevents users from receiving high-quality AIGC outputs. Moreover, regeneration consumes additional bandwidth.

In this paper, we adopt Semantic Communications (SemCom) [8] in mobile AIGC to overcome the bandwidth constraints. Instead of transmitting every bit, SemCom circumvents the channel capacity limitation by only transmitting critical semantic information, enabling users to accomplish specific applications while saving wireless bandwidth [8]. In SemCom-aided mobile AIGC, a MASP can extract semantic features of the AIGC outputs, thereby compressing the content to be transmitted. Then, the users can apply a lightweight decoder to recover the source AIGC outputs with high fidelity. Note that several studies have explored the potential of SemCom in mobile AIGC [9], [10]. However, they do not implement the systematic SemCom-aided mobile AIGC and perform intensive experiments to illustrate how much bandwidth can be saved by SemCom without affecting the AIGC output quality on the user side. In contrast, we present a novel SemCom framework containing three designs oriented to mobile AIGC: i) To extract AIGC outputs' semantics and perform output recovery efficiently, we introduce a semantic extraction module in the MASP's AIGC model and equip users with generative decoders, forming the Generative SemCom (G-SemCom). ii) Noticing the modality transfers during AIGC inferences, our semantic information takes the form of a series of cross-modal attention maps, which associate each prompt word to certain parts of the AIGC output by attention scores. Hence, we can perform fine-grained semantic analysis of the AIGC outputs, filtering the content with the most important semantic meaning for users. iii) Traditional SemCom only optimizes the semantic similarity between the source and recovered information. However, in mobile AIGC, users require the recovered AIGC outputs to be high-quality. To this end, we present the joint optimization, which performs prompt engineering to ensure output quality when allocating wireless bandwidth for output transmission.

The main contributions of this paper can be summarized as follows:

- *G-SemCom Framework for Mobile AIGC: To the best of our knowledge, we are the first to present the cross-modal G-SemCom framework for mobile AIGC.* Supported by G-SemCom, each MASP only needs to transmit compressed semantic information of the AIGC output. On the user side, a generative decoder is deployed for recovery. In this way, the users can acquire high-quality AIGC outputs while saving considerable computation and bandwidth resources.
- *Attention-Aware Semantic Extraction:* Noticing the cross-modality feature of mobile AIGC, we propose an attention-aware method to extract semantic features of the source information. Specifically, we visualize the activation of the cross-attention layers in diffusion-based AIGC models, forming a series of cross-modal attention maps. By scoring the correlation between the user prompt and each part of the generated AIGC output, efficient semantic encoding can be performed from the user's perspective, thereby ensuring the semantic correctness of the recovered AIGC output.
- *Joint Semantic Encoding and Prompt Engineering:* We formulate a joint optimization problem to optimize the bandwidth allocation. Particularly, since the information sent by MASP not only preserves semantic features but also serves as the prompt for guiding the recovery of AIGC outputs, we consider the joint semantic encoding and prompt engineering with the goal of simultaneously maximizing the semantic similarity and output quality while saving wireless bandwidth. To do so, we define a novel human-perceptual metric called Joint Perceptual Similarity and Quality (JPSQ) to indicate the efficiency of G-SemCom in mobile AIGC. Moreover, we develop the Attention-aware Deep Diffusion (ADD) algorithm to solve the optimization, which utilizes diffusion steps to achieve strong exploration ability.
- *Experimental Results:* Extensive experiments prove the validity of our proposals. Specifically, the bandwidth consumption of mobile users can be reduced by 49.4% on average, while the perceptual output quality score [11] only drops by 0.0299. Moreover, the ADD algorithm significantly outperforms baseline Deep Reinforcement Learning (DRL) algorithms regarding converge speed and efficiency for bandwidth allocation.

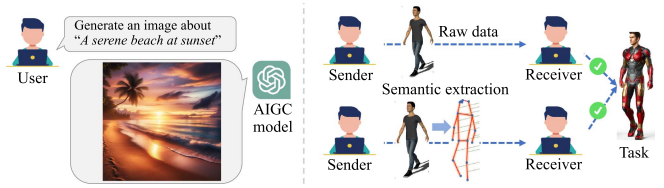

Fig. 1. An example of AIGC (left). The comparison between traditional communication and SemCom (right).

The remainder of this paper is organized as follows. Section II presents some preliminary and reviews the related works. Our motivation and system model are shown in Section III. In Section IV, we elaborate on our G-SemCom framework, especially the attention-aware semantic extraction, for mobile AIGC. Then, Section V describes the joint optimization of semantic encoding and prompt engineering via ADD. The experimental results and analysis are illustrated in Section VI. Finally, Section VII concludes the paper.

II. Preliminary and Related Work

A. Preliminary

1) AIGC: AIGC aims to assist or replace manual content generation by automatically generating content according to user-inputted prompts [1]. The core of AIGC is Generative AI (GAI) models, which are trained to produce data that mimic the distribution described by the input. Representative GAI models include diffusion, transformer, Generative Adversarial Networks (GAN), etc [1]. These models are integral to applications like creating photorealistic images or composing music. For example, users can request ChatGPT to generate an image of "a serene beach at sunset." As shown in Fig. 1(left), the generated image visually represents this scene.

2) Semantic Communication: Traditional data transmission methods primarily focus on transmitting every bit with minimal loss. In contrast, SemCom is task-oriented, prioritizing the meaning and utility of the data over its exact form [12]. This shift allows for the elimination of redundant data, thereby reducing transmission overhead. Hence, the process of SemCom typically involves semantic extraction on the sender side, where the vital semantics of source information that contributes to task accomplishment is distilled and encoded. Afterward, on the receiver side, such semantics can be decoded to efficiently accomplish the designated task. For instance, consider a scenario where the receiver needs to render 3D avatars. As shown in Fig. 1(right), instead of sending the entire user photo, the sender can train a semantic extractor, which only transmits the skeleton of the user.

B. Mobile AIGC

Mobile Resource Constraints: Given the constraints of mobile resources, a series of lightweight AIGC models have been presented. For instance, Chen et al. [5] conducted GPU-aware optimization on large diffusion models, realizing the on-device text-to-image generation in 12 seconds. SnapFusion [13] utilized step distillation and further reduced the inference time to 2 seconds on mobile devices. On Feb. 2023, Qualcomm developed the world's first on-device Stable Diffusion [4]. Likewise, Google and Apple also presented MediaPipe [14] and Core-ML Stable Diffusion [15], respectively. From the system perspective, the architecture and management of mobile AIGC are also evolving rapidly. Xu et al. [1] systemically introduced the potential of mobile-edge networks for accommodating AIGC services. Du et al. [6] discussed the task scheduling of mobile-edge AIGC, improving the system capacity by assigning each AIGC task to the most appropriate MASP. Wen et al. [7] designed the incentive mechanism for rewarding MASPs, ensuring the participation and economic sustainability of mobile AIGC. Despite reducing computing consumption, mobile AIGC users need to frequently download large AIGC outputs from MASPs, which costs huge wireless bandwidth. To this end, we present an end-to-end SemCom framework for mobile AIGC.

Modality Transfer during AIGC Inferences: AIGC inferences refer to the process of generating high-dimensional content from user-friendly low-dimensional prompts [1]. To semantically align the generated content with prompts, cross-modal attention mechanisms are pivotal and widely adopted by Stable Diffusion, CLIP, etc. [16]. As presented in Transformer [16], attention mechanisms enable neural networks to focus dynamically on relevant parts of the input during the training process, ensuring that the output adheres closely to the input's described attributes and context. In AIGC, cross-modal attention extends this concept to bridge data in different modalities, aligning features across these modalities and ensuring the accurate translation of the prompt's semantics into the generated content [17]. Hence, this paper exploits semantics from cross-model attention maps.

C. Generative Semantic Communications (G-SemCom)

Generative models have shown great potential to be incorporated into SemCom, which we coin as G-SemCom [18]. On the sender side, generative models can help extract human-interpretable semantic features. For instance, Wang et al. [8] generated semantic triples (formed by *object A-relationship-object B*) to compress source textual message. For image-oriented G-SemCom, Liu et al. [19] evaluated various formats for representing visual semantics, e.g., skeleton and depth maps. Compared with parameterized semantic features used by traditional SemCom, such human-interpretable semantics are easy to analyze, making human-in-the-loop SemCom optimization possible. On the receiver side, efficient semantic decoding and information recovery can be realized by generative decoders. For instance, He et al. [20] and Grassucci et al. [18] adopted GANs and diffusion models to reconstruct images that are semantically equivalent to the source images. Finally, by training on huge datasets, large generative models can serve as powerful shared knowledge bases between senders and receivers [21]. For example, Jiang et al. [21] developed a training-free knowledge base by Meta Segment Anything, realizing zero-shot semantic extraction and supporting fast information recovery. Motivated by such progress, this paper leverages G-SemCom to help relieve the heavy transmission burden of mobile AIGC.

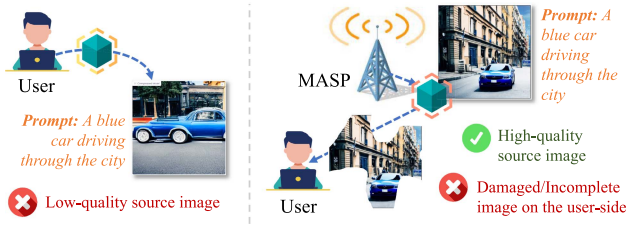

Fig. 2. Local AIGC (left) and Traditional Mobile AIGC (right).

D. Prompt Engineering

Prompt engineering refers to the process of crafting/finding the most appropriate prompt for the given downstream task, aiming to maximize the generation quality [22]. According to the specific prompt form, various prompt engineering methods have been proposed. Taking textual prompts, e.g., the instructions that we input to ChatGPT, as an example, the authors in [23] and [24] presented the prompt paraphrasing and searching, respectively. Despite adopting different strategies, the common principle is finding the best textual template for reformulating raw prompts, facilitating pretrained generative models to associate downstream tasks with the learned knowledge. Apart from determining the generation quality, in mobile AIGC, prompt engineering also directly affects the network-level performance. Liu et al. [3] stated that if users keep using low-quality prompts, frequent re-generation will cause considerable service fees, extra service latency, and bandwidth consumption. To this end, we jointly perform semantic encoding and prompt engineering on the sender side, thereby optimizing the input fed to the receiver's generative decoder while saving bandwidth.

III. Motivation and System Model

A. Motivation

Without loss of generality, we consider the text-to-image AIGC scenario in this paper. However, the proposed framework and algorithms are applicable to other forms of AIGC, which will be discussed in Section VI-E. Suppose that users adopt "A blue car driving through the city." as the prompt for image generation. The existing AIGC paradigms include:

- *Local AIGC:* Due to constrained computing resources, if generating images locally, the users can only afford to utilize compressed AIGC models [25], resulting in poor generation quality [see Fig. 2(left)].
- *Traditional Mobile AIGC:* Leveraging mobile AIGC, the users can call MASPs to generate high-quality images using powerful AIGC models [26]. Nonetheless, repeated image downloads from the MASPs consume considerable communication resources. Moreover, given the limited wireless bandwidth, large output images may not be fully transmitted. Damaged or incomplete images are useless to users [see Fig. 2(right)].

To this end, we develop G-SemCom for mobile AIGC. Our goal is to enable mobile users to acquire high-quality images under computing and communication resource constraints.

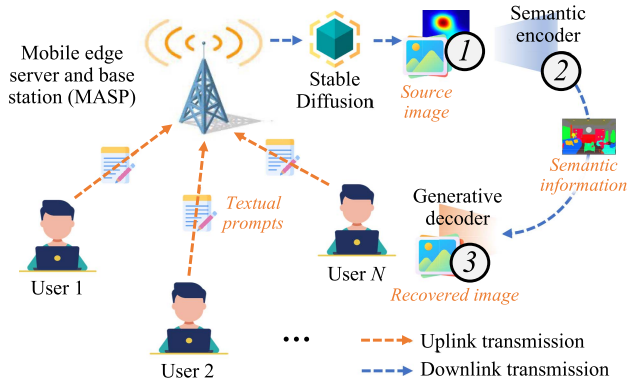

Fig. 3. The system model. Step 1: *Cross-modal attention map generation*, Step 2: *Attention-aware semantic extraction*, and Step 3: *Generative decoding*.

B. System Model

As shown in Fig. 3, we consider the system with one MASP and N users, denoted by $\mathcal{U} = \{U_1, U_2, \ldots, U_N\}$. However, the model can be extended straightforwardly for multiple MASPs. To acquire high-quality images, the users first send their prompts to the MASP. Serving by mobile edge servers and base stations [27], the MASP operates Stable Diffusion [26] [1], the start-of-the-art text-to-image model, and provides generative inference services for the users. In this way, high-quality source images can be generated. Then, G-SemCom is applied to overcome the bandwidth constraints. Specifically, the MASP generates a series of cross-modal attention maps during the inference, which associate each prompt word with certain source image pixels (Step 1). After attention-aware semantic extraction (Step 2), only the most semantically important pixels serve as semantic information and are transmitted over a wireless channel. Then, the users employ a generative decoder, taking semantic information as the prompt to recover the source image (Step 3). Particularly, to strike an optimal balance between the limited bandwidth and the human-perceptual G-SemCom experience, we formulate a joint semantic encoding and prompt engineering problem. This aims to optimize the bandwidth allocation among multiple mobile AIGC users. Next, we illustrate the transmission model. Afterward, Sections IV and V discuss the G-SemCom design and the joint optimization problem, respectively.

C. Transmission Model

We utilize the orthogonal frequency division multiple access (OFDMA) technique [8] to model the wireless transmission between the MASP and users. Specifically, each user is allocated one downlink orthogonal resource block (RB). Suppose that the i^{th} RB is assigned for transmitting semantic information $\boldsymbol{S}_i$ to user U_i, the corresponding downlink channel capacity is defined

[1]This paper selects Stable Diffusion as an example due to its well-proven generation quality and easy accessibility. Our framework can adapt to various mainstream AIGC models that incorporate attention mechanisms, such as Sora [28] and StoryDiffusion [29].

TABLE I
THE MAIN MATHEMATICAL NOTATIONS

Notation	Description
$A_z^{\mathbb{R}^+}[x,y]$	Cross-model attention map
$A_z^{0,1}[x,y]$	Binary attention map
β_t	Noise added in forward diffusion at step t
$\boldsymbol{C}$, $\boldsymbol{C}^*$	Boolean dependency matrix (original and compressed)
ϕ_i	Channel gain
γ	Discount factor of ADD
$\boldsymbol{D}$, $\boldsymbol{D}^*$	Dependency level matrix (original and compressed)
$F_t^{(i)\mathrm{d}}$, $F_t^{(i)\mathrm{u}}$	Cross-modal attention scores for down/upstream blocks
$\boldsymbol{I}$	Identity matrix
η	Learning rate of ADD
L_i	Transmission latency of user U_i
N_0	Noise power spectral density
O	Bandwidth consumption of transmitting each token in $\mathcal{S}_{A_i^{0,1}}$
P	Transmission power of MASP
$\mathbf{p}$	Prompt provide by user for image generation
Q_{th}	User threshold for aesthetic quality
T	Number of denoising steps in source image generation and ADD
ξ	Threshold for constructing binary attention maps
$\mathbf{x}$	Latent vector used in the diffusion process
$\mathbf{w}$	Word embeddings
W	Bandwidth of each RB
$\omega_0, \omega_1, \omega_2$	Weighting factors in optimization problems
$\boldsymbol{S}_i$	Semantic information for user U_i
$\mathcal{S}_{A_i^{0,1}}$	Image segments
$\mathbf{s}$	Semantic importance of each prompt word

as [8]

$$c_i = W \log_2 \left(1 + \frac{P\phi_i}{I_n + WN_0}\right), i \in \{1, 2, \ldots, N\} \quad (1)$$

where W is the bandwidth of each RB; P means the transmission power of the MASP, I_q means the interference caused by the base stations that are located in other service areas and use the i^{th} RB, and N_0 is the noise power spectral density. $\phi_i = \gamma_i\, d_i^{-2}$ represents the channel gain between the MASP and user U_i with γ_i being the Rayleigh fading parameter and d_i being their physical distance. Here, we consider that the transmission latency between the MASP and user U_i is limited to L_i. Hence, given the data rate c_i, the maximum size of semantic information can be determined. Note that the important mathematical notations used in this paper are summarized in Table I.

IV. CROSS-MODAL G-SEMCOM FOR MOBILE AIGC

In this section, we illustrate the design of our G-SemCom framework for mobile AIGC. First, we introduce the process of source image generation. Then, we demonstrate the generation of cross-modal attention maps. Finally, we develop the G-SemCom encoder and decoder.

A. Source Image Generation

According to the textual prompt, the MASP can generate a source image using Stable Diffusion, depicting the objects and scenes described by the user. As shown in Figs. 4(a) and (b), to realize such text-to-image generations, Stable Diffusion adopts a modular architecture with three components, namely a deep visual-language model called CLIP [30], a variational autoencoder (VAE) [3], and a UNet-based noise predictor [17]. To train Stable Diffusion, a large dataset containing massive caption-image pairs is first prepared. During each training iteration, the fetched image and its caption are encoded by VAE and CLIP into a latent vector $\mathbf{x}_0$ and word embeddings $\mathbf{w} := [w_1, \ldots, w_{l_W}]$, respectively. Afterward, a Markov process called forward diffusion is performed. Specifically, $\mathbf{x}_0$ is gradually perturbed by adding noise for T times, until it becomes a pure Gaussian noise $\mathbf{x}_T$, i.e.,

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \quad (2)$$

where each denoising step satisfies

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}\right), \quad (3)$$

where $\boldsymbol{I}$ represents the identity matrix. Note that $\{\beta_t\}_{t=1}^T$ follows a pre-defined schedule so that $p(\mathbf{x}_T)$ is approximately zero-mean isotropic [16]. The forward diffusion aims to train the noise predictor, which utilizes the UNet to learn the amount of noise that should be added in each step. For generating new images, Stable Diffusion first randomly generates a latent vector $\mathbf{x}_T$. Then, it performs the reverse diffusion process to subtract noise from $\mathbf{x}_T$. According to [31], such a denoising process can be expressed as

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathbf{w}), \beta_t \boldsymbol{I}), \quad (4)$$

$$\mu_\theta(\mathbf{x}_t, t, \mathbf{w}) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right), \quad (4a)$$

$$\alpha_t := 1 - \beta_t,\ \bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i, \quad (4b)$$

where $\epsilon_\theta(\mathbf{x}_t, t; \mathbf{w})$ means the noise predicted by UNet with parameters θ. Iteratively processing (4), the latent representation of the required image, i.e., $\mathbf{x}_0$, can be generated. Finally, $\mathbf{x}_0$ is decoded by VAE and becomes a high-quality and user-perceivable source image.

B. Cross-Modal Attention Map

From (4), we can observe that text embeddings $\mathbf{w}$ condition the image generation, which explains why the generated images are semantically equivalent to user prompts. As shown in Fig. 4(b), in Stable Diffusion, the text embeddings and latent image vector are bridged by UNet's spatial transformer blocks in the form of cross-modal attention. To be specific, UNet is basically composed of K downsampling convolutional blocks and the corresponding upsampling blocks [see Fig. 4(b)]. Suppose that given a latent image vector $\mathbf{x}_t \in \mathbb{R}^{\omega \times h}$ ($t \in \{1, 2, \ldots, T\}$), first, the downsampling blocks output a series of vectors $\{\mathbf{v}_{i,t}^{\mathrm{d}}\}_{i=1}^K$, where $\mathbf{v}_{i,t}^{\mathrm{d}} \in \mathbb{R}^{\lceil \frac{\omega}{c_i} \rceil \times \lceil \frac{h}{c_i} \rceil}$ for some $c > 1$. Then, the upsampling blocks iteratively upscale $\mathbf{v}_{K,t}^{\mathrm{d}}$ to $\{\mathbf{v}_{i,t}^{\mathrm{u}}\}_{i=K-1}^0$,

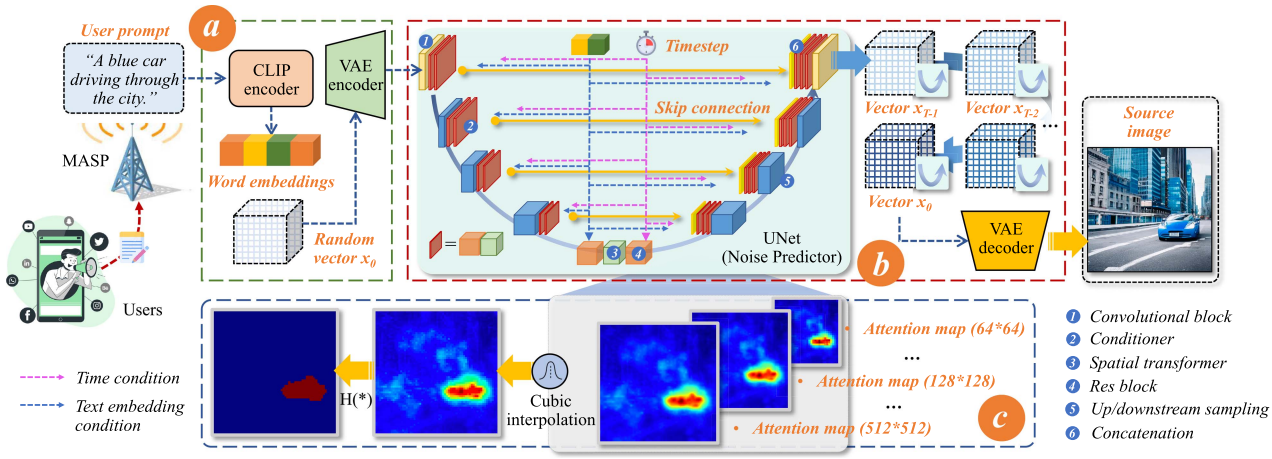

Fig. 4. The illustration of generating source images and cross-modal attention maps. (a): The CLIP and VAE modules. (b): The UNet architecture and diffusion process. (c): The attention map of word [car].

where $\mathbf{v}_{i,t}^{\mathrm{u}} \in \mathbb{R}^{\lceil \frac{\omega}{c_i} \rceil \times \lceil \frac{h}{c_i} \rceil}$. To support conditioned content generation, diffusion-based AIGC models like Stable Diffusion attach two more blocks to each downsampling/upsampling block, namely a resblock and a spatial transformer [17]. They are responsible for providing t and $\mathbf{w}$ conditions in (4), respectively. In this paper, we focus on the latter since the cross-modal attention reflects the modality transfer happening during the AIGC inference. For each downsampling/upsampling block, the cross-modal attention corresponding to it can be denoted by

$$\mathbf{v}_{i,t}^{\mathrm{d}} := F_t^{(i)\mathrm{d}}\left(\hat{\mathbf{v}}_{i,t}^{\mathrm{d}}, \mathbf{w}\right)\left(\boldsymbol{W}_v^{(i)}\mathbf{w}\right), \tag{5}$$

$$F_t^{(i)\mathrm{d}}\left(\hat{\mathbf{v}}_{i,t}^{\mathrm{d}}, \mathbf{w}\right) := \mathrm{softmax}\left(\frac{\left(\boldsymbol{W}_q^{(i)}\hat{\mathbf{v}}_{i,t}^{\mathrm{d}}\right)\left(\boldsymbol{W}_k^{(i)}\mathbf{w}\right)^T}{\sqrt{d}}\right), \tag{6}$$

where $F_t^{(i)\mathrm{d}}$ denotes the normalized downsampling attention score array. The attention of each word $w_z, z \in \{1, 2, \ldots, l_W\}$ on the 2D intermediate coordinate of the l^{th} head ($l \in \{1, 2, \ldots, l_H\}$) belonging to the i^{th} downsampling block can be measured with a score within [0, 1]. $\boldsymbol{W}_k$, $\boldsymbol{W}_q$, and $\boldsymbol{W}_v$ are projection matrices with l_H attention heads; d is a scaling factor. Note that for simplicity, we do not show the equations of upsampling attention score array $F_t^{(i)\mathrm{u}}$, which are similar to (5) and (6). As shown in Fig. 4(c), the intermediate coordinate, in the form of $[x, y]$, is locally mapped to a surrounding affected square area in the source image. In this way, we can quantify the correlation between the given prompt word and each image pixel according to the attention score value. However, the downsampling/upsampling blocks of UNet vary in size, resulting in a series of attention maps with different scales. Based on [16], as shown in Fig. 4(c), we upscale all $F_t^{(i)\mathrm{d}}$ and $F_t^{(i)\mathrm{u}}$ to the original image size, i.e., $\omega \times h$, using bicubic interpolation. Then, the attention scores are summed up over the heads, layers, and diffusion steps, forming the cross-modal attention map as follows

$$A_z^{\mathbb{R}^+}[x, y] := \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{\ell=1}^{l_H}\left(F_{t,(z,\ell)}^{(i)\mathrm{d}}[x, y] + F_{t,(z,\ell)}^{(i)\mathrm{u}}[x, y]\right), \tag{7}$$

where z and l represent the indexes of the word embedding and downsampling/upsampling block, respectively. $\mathbb{R}^+$ indicates that any $A_z^{\mathbb{R}^+}[x, y]$ belongs to positive real number. Finally, we generate the binary cross-modal attention maps by

$$A_z^{\{0,1\}}[x, y] := \mathrm{H}\left(A_z^{\mathbb{R}^+}[x, y] \geq \xi \max_{x,y} A_z^{\mathbb{R}^+}[x, y]\right), \tag{8}$$

where $\xi \max_{x,y} A_z^{\mathbb{R}^+}[x, y]$ is the pre-defined threshold. $\mathrm{H}(\cdot)$ represents the Heaviside step function, which outputs $\boldsymbol{1}$ when the value of $A_z^{\{0,1\}}[x, y]$ exceeds the threshold, and $\boldsymbol{0}$ otherwise. Compared with fine-grained attention scores, i.e., $A_z^{\mathbb{R}^+}[x, y]$, $A_z^{\{0,1\}}[x, y]$ facilitates the set operations on multiple attention maps, which are discussed below.

C. Attention-Aware Semantic Extraction

With the cross-modal attention maps, in this part, we extract semantic features from the source image. The entire procedure is shown in Algorithm 1.

1) Textual Prompt Deconstruction: First, the MASP analyzes the textual prompts provided by users, denoted by $\mathbf{p} := [p_1, p_2, \ldots, p_M]$, trying to understand the semantic meaning of users' requirements. To do so, we utilize *Spacy* [16] to perform the *Part-of-Speech* tagging, i.e., classifying the words according to their linguistic functions. As shown in Table II, we consider seven part-of-speech types that are semantically important [16] and use $\boldsymbol{X}$ to include all other types (e.g., determiner, interjection, and conjunction) with less semantic meanings. Take "A blue car driving through the city." as an example. Words [car] and [city] belong to $\boldsymbol{NN}$; [blue] belongs to $\boldsymbol{ADJ}$; [driving] belongs to $\boldsymbol{VERB}$; [through] belongs to $\boldsymbol{ADP}$; [A], [the], and [.] belong to $\boldsymbol{X}$. Afterward, we perform dependency parsing [32], aiming to analyze the grammatical structure of $\mathbf{p}$ and find out related words as well as their correlation. As shown in Fig. 5(a), each dependency item takes the form of an arrow, from head to the word that modifies it, called dependent. Similarly, this step can be realized by *Spacy*. In this way, the Boolean dependency matrix $\boldsymbol{C} \in \mathbb{R}^{M \times M}$ can

TABLE II
THE PART-OF-SPEECH OF TEXTUAL PROMPTS

Type	Definition	Examples
NN	noun	`dog, people, city ...`
PROPN	proper noun	`iPhone, IEEE, Alice ...`
NUM	numeral	`one, two, 100, 10th ...`
ADJ	adjective	`red, beautiful, good ...`
VERB	actions	`run, drive, think ...`
ADV	adverb	`quickly, rapidly, here ...`
ADP	adposition	`on, at, through ...`
X	other words/symbols	`(a, the), (and, but), (Wow), (she, he), (".", "$") ...`

Note that *X* includes all the words/symbols that do not belong to any of the above types.

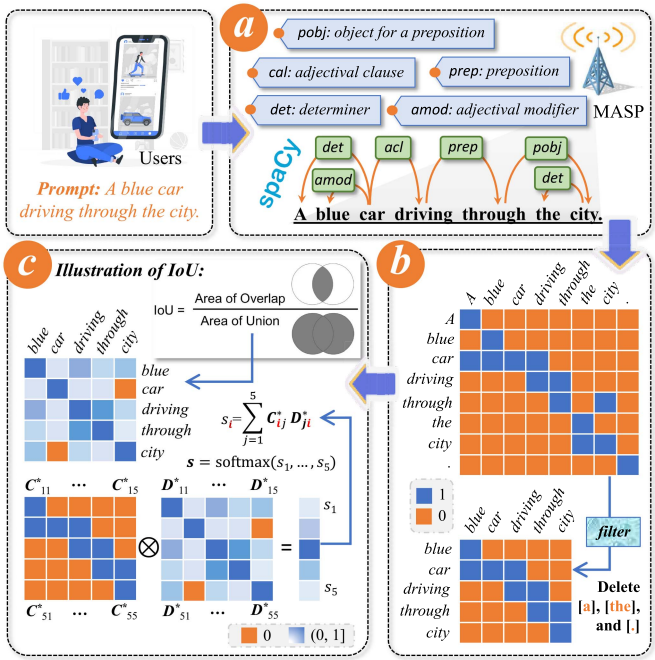

Fig. 5. Attention-aware semantic extraction. (a): Dependency parsing. (b): Boolean dependency matrices $\boldsymbol{C}$ and $\boldsymbol{C}^*$. (c): Dependency level matrix $\boldsymbol{D}^*$ and $\otimes$ operation.

be constructed as follows

$$\boldsymbol{C} = \begin{bmatrix} R^{\{0,1\}}_{(1\leftarrow 1)} & \cdots & R^{\{0,1\}}_{(1\leftarrow i)} & \cdots & R^{\{0,1\}}_{(1\leftarrow M)} \\ R^{\{0,1\}}_{(i\leftarrow 1)} & \ddots & R^{\{0,1\}}_{(i\leftarrow i)} & \ddots & R^{\{0,1\}}_{(i\leftarrow M)} \\ R^{\{0,1\}}_{(M\leftarrow 1)} & \cdots & R^{\{0,1\}}_{(M\leftarrow i)} & \cdots & R^{\{0,1\}}_{(M\leftarrow M)} \end{bmatrix}, \tag{9}$$

where $R^{\{0,1\}}_{(i\leftarrow j)}$ $(i, j \in \{1, 2, \ldots, M\})$ takes the value $\boldsymbol{1}$ if the dependency exists between p_i and p_j, i.e., the head and the dependent, respectively, and $\boldsymbol{0}$ otherwise.

Due to weak semantic meaning, the words belonging to $\boldsymbol{X}$ can be filtered out to reduce the computation complexity. Accordingly, $\boldsymbol{C}$ can be compressed to $\boldsymbol{C}^* \in \mathbb{R}^{(M-\zeta)\times(M-\zeta)}$, where ζ represents the number of $\boldsymbol{X}$-type words in $\mathbf{p}$. The original and compressed dependency matrices of "A blue car driving through the city." are shown in Fig. 5(b).

$\boldsymbol{C}^*$ can reflect the importance of each word based on the number of dependencies that it involves. Nonetheless, according to the types of the head and the dependent, as well as the function that the dependent acts on the head, there exist more than 18 kinds of dependencies [32]. Some dependencies, such as (`amod: [blue] ← [car]`), are strong, while the others, such as (`det: [A] ← [car]`), are weak. To this end, we leverage the *Mean Intersection over Union* (mIoU) to calculate the fine-grained semantic importance of each word. Suppose that the pixels of the entire source image construct the Universe $\mathcal{S}$, and the pixels included by the binary attention maps of words p_i and p_j are sets $\mathcal{S}_{A_i^{\{0,1\}}}$ and $\mathcal{S}_{A_j^{\{0,1\}}}$, respectively. mIoU can be derived as

$$\text{mIoU}_{(i\leftarrow j)} = \frac{|\mathcal{S}_{A_i^{\{0,1\}}} \cap \mathcal{S}_{A_j^{\{0,1\}}}|}{|\mathcal{S}_{A_i^{\{0,1\}}} \cup \mathcal{S}_{A_j^{\{0,1\}}}|}. \tag{10}$$

As shown in Fig. 5(c), $\text{mIoU}_{(i\leftarrow j)}$ can measure the similarity of the areas covered by the binary attention maps of words p_i and p_j and is within [0, 1]. Consequently, the higher the mIoU value, the stronger the dependency exists in two words. Using mIoU, we can acquire the following dependency level matrix, denoted by $\boldsymbol{D}^* \in \mathbb{R}^{(M-\zeta)\times(M-\zeta)}$.

$$\boldsymbol{D}^* = \begin{bmatrix} L^{\{0,\mathbb{R}^+\}}_{(1\leftarrow 1)} & \cdots & L^{\{0,\mathbb{R}^+\}}_{(1\leftarrow i)} & \cdots & L^{\{0,\mathbb{R}^+\}}_{(1\leftarrow \sigma)} \\ L^{\{0,\mathbb{R}^+\}}_{(i\leftarrow 1)} & \ddots & L^{\{0,\mathbb{R}^+\}}_{(i\leftarrow i)} & \ddots & L^{\{0,\mathbb{R}^+\}}_{(i\leftarrow \sigma)} \\ L^{\{0,\mathbb{R}^+\}}_{(\sigma\leftarrow 1)} & \cdots & L^{\{0,\mathbb{R}^+\}}_{(\sigma\leftarrow i)} & \cdots & L^{\{0,\mathbb{R}^+\}}_{(\sigma\leftarrow \sigma)} \end{bmatrix}, \tag{11}$$

where $L^{\{0,\mathbb{R}^+\}}_{(i\leftarrow j)}$ $(i, j \in \{1, 2, \ldots, (M-\zeta)\})$ takes the value of $\text{mIoU}_{(i\leftarrow j)}$; σ equals $M - \zeta$. Finally, the semantic importance of each word, denoted as $\mathbf{s} := \{s_1, s_2, \ldots, s_{M-\zeta}\}$, in which the $\mathbf{X}$-type words have been filtered out, can be derived as

$$\mathbf{s} = \text{softmax}(\boldsymbol{C}^* \otimes \boldsymbol{D}^*), \tag{12}$$

where $\otimes$ represents the matrix multiplication operation and is shown in Fig. 5(c). Note that $\mathbf{s}$ is normalized by softmax, ensuring that every s_i $(i \in \{1, 2, \ldots, (M-\zeta)\})$ is within [0, 1] and $\sum_{i=1}^{M-\zeta} s_i = 1$. In our example, the importance of `[blue]`, `[car]`, `[driving]`, `[through]`, and `[city]` are 0.16, 0.20, 0.31, 0.17, and 0.16, respectively. Hence, two major objects and their relationship, i.e., `[car]`, `[city]`, and `[driving]`, convey the major semantic meaning of the entire source image. In contrast, the preposition, i.e., `[through]`, is weak in terms of semantic importance.

2) Visual Prompt Segmentation: Up till now, we can evaluate the semantic importance of each word and link it to certain areas of the source image. However, as shown in Fig. 6(a), the attention distribution of some words (especially *ADV*- and *VERB*-type ones) in the source image is scattered, containing a lot of outlier noise. Such noise not only wastes bandwidth resources but also increases the difficulty of image recovery. To this end, we intend to perform clustering on the cross-modal attention maps according to the attention density and remove the noise. Therefore, leveraging *Density-Based Spatial Clustering*

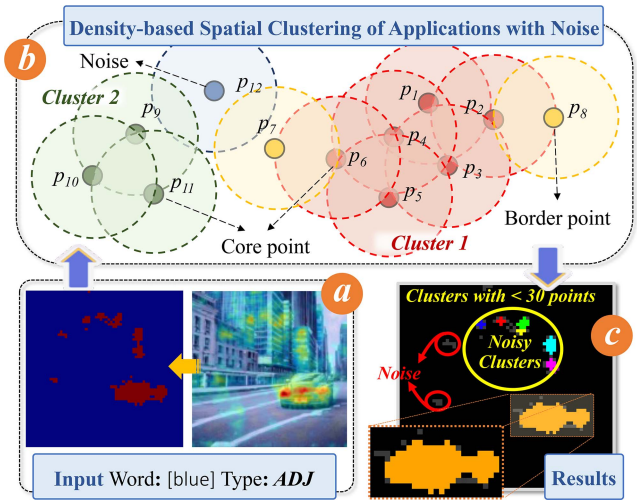

Fig. 6. Illustration of attention clustering. (a): The raw and binary attention maps. (b): The illustration of the DBSCAN algorithm. (c): The clustering results. Note that noise and noise clusters refer to scatters and clusters with less than 30 points, respectively. Due to the limited size, they cannot convey enough semantics. Hence, they will be filtered out.

of Applications with Noise (DBSCAN) [33], we perform the following three-step attention-aware visual prompt segmentation.

Attention Point Clustering: Given a cross-modal attention map, i.e., $A_z^{\{0,1\}}$ $(z \in \{1, 2, \ldots, (M-\zeta)\})$, we first cluster dense attention points and filter the noise, using a sophisticated clustering algorithm called DBSCAN[2]. DBSCAN conducts clustering by first detecting all the core points that have at least Ω neighbors, i.e.,

$$N_\varepsilon(p) \geq \Omega, N_\varepsilon(p) = |\{q \in \mathcal{S}_{A_z^{\{0,1\}}} | \mathrm{d}(p,q) \leq \varepsilon\}|, \quad (13)$$

where ε and Ω are user-defined and represent the distance threshold and the required number of neighbors within ε, respectively. d is the function for distance measurement. As shown in Fig. 6(b), DBSCAN starts from a random core point, e.g., p_1, and iteratively groups all neighboring core points (i.e., $p_2 \ldots, p_6$) into the same cluster. The border points that are close to any of the aforementioned core points are also included (i.e., p_7 and p_8). Afterward, another core point that has not been clustered, e.g., p_8, can be selected, and the above process is repeated. The algorithm will stop when all the core points are clustered. Accordingly, the remaining points are viewed as noise. More details are shown in Algorithm 1. Fig. 6(c) illustrates the clustering results of the attention map of the word `[blue]`.

Source Image Segmentation: Step 1 will be performed for all the acquired attention maps. Afterward, we can acquire a series of clean image segments, denoted by $\mathcal{S}^*_{A_z^{\{0,1\}}}$ $(\forall z \in \{1, 2, \ldots, (M-\zeta)\})$.

Semantic Information Packing: Finally, the MASP packs semantic information, which guides the users to recover semantically similar and high-quality images. Therefore, the pixels owning stronger semantic meanings should be prioritized. Next, we reorder $\mathbf{p}$ according to the semantic importance of each word, i.e., $\mathbf{s}$. Then, the semantic information matrix $\boldsymbol{S}$ can be

[2]DBSCAN is applied due to its wide adoption and well-proven performance in clustering points following complex distributions. The cluster algorithm is designed as a pluggable module in our framework. Other algorithms, such as k-means and OPTICS [34], can also be applied.

Algorithm 1: The Operations on the MASP-Side.

Require: g_0, $\mathbf{p} = [p_1, p_2, \ldots, p_M]$, $A_z^{\{0,1\}}[x, y], z \in \{1, 2, \ldots, M\}$ *## source image, prompt, and binary attention maps*
Ensure: $\boldsymbol{I}$ *## semantic information*

procedure Textual Prompt Extraction$\mathbf{p}$
  Call *spacy* to perform text-to-speech tagging and dependency parsing
  Initialize $\boldsymbol{C} = 0^{M \times M}$
  for all $p_i \in \mathbf{p}$ **do** *## row iteration*
    for all $p_j \in \mathbf{p}$ **do** *## column iteration*
      if $i = j$ **then**
        $\boldsymbol{C}_{ij} = 1$ *## each world is correlated to itself*
      else
        if p_i and p_j has dependency with p_i as the head and p_j as the dependent **then**
          $\boldsymbol{C}_{ij} = 1$ *## mark the dependency*
        end if
      end if
    end for
  end for
  for all $p_i \in \mathbf{p}$ **do** *## filter non-important words, acquiring* $\boldsymbol{C}^*$
    if p_i belongs to $\boldsymbol{X}$-type **then**
      Delete the i^{th} column and row of $\boldsymbol{C}$
    end if
  end for
  Initialize $\boldsymbol{D}^* = \boldsymbol{C}^*$ *## the compressed dependency matrix*
  Initialize $\boldsymbol{D}^* = 0^{(M-\zeta)\times(M-\zeta)}$
  for all $\boldsymbol{D}^*_{ij} \in \boldsymbol{D}^*$ **do**
    $\boldsymbol{D}^*_{ij} = \mathrm{IoU}_{(i \leftarrow j)}$
  end for
  $\mathbf{s} = \mathrm{softmax}(\boldsymbol{C}^* \otimes \boldsymbol{D}^*)$
end procedure
procedure Visual Prompt Segmentation$A_z^{\{0,1\}}[x, y]$
  for all $z \in \{1, 2, \ldots, N\}$ **do**
    Call DBSCAN to cluster the attention points of $A_z^{\{0,1\}}[x, y]$
    Filter noise points, as well as the clusters with less than 30 points
  end for
end procedure
procedure Semantic Information Packing$A_z^{\{0,1\}}[x, y]$, g_0, $\mathbf{p}$
  for all $A_z^{\{0,1\}}[x, y], z \in \{1, 2, \ldots, N\}$ **do**
    Sort the attention maps according to $\mathbf{s}$
  end for
  Construct $\boldsymbol{S}$ according to (14)
  Send as many tokens in $\boldsymbol{S}$ as possible
end procedure

generated.

$$\boldsymbol{S} = \begin{bmatrix} \mathbf{s}_1 = \mathcal{S}^*_{A_1^{\{0,1\}}}, \\ \mathbf{s}_2 = \mathcal{S}^*_{A_2^{\{0,1\}}} \setminus \mathbf{s}_1, \\ \mathbf{s}_3 = \mathcal{S}^*_{A_3^{\{0,1\}}} \setminus (\mathbf{s}_1 \cup \mathbf{s}_2), \\ \dots \\ \mathbf{s}_{M-\zeta} = \mathcal{S}^*_{A_{M-\zeta}^{\{0,1\}}} \setminus (\mathbf{s}_1 \cup \dots \cup \mathbf{s}_{M-\zeta-1}) \end{bmatrix}. \tag{14}$$

The MASP will send $[\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{M-\zeta}]$ in sequence until all available bandwidth is used.

3) Discussion: By generating cross-modal attention maps, analyzing the semantic meaning of user prompts, performing attention-aware segmentation of the source image, and only transmitting the semantically important pixels, the size of data that users should download can be efficiently reduced. Furthermore, the proposed G-SemCom framework will not incur considerable workloads for the MASP. First, the cross-modal attention maps can be regarded as by-products during the source image generation, causing no additional computation costs. The computation complexity of DBSCAN is $\mathcal{O}(n \log n)$, where n represents the number of attention points [35]. The remaining operations, such as matrix filtering and multiplications, have the complexity of $\mathcal{O}(M-\zeta)$ to $\mathcal{O}((M-\zeta)^2)$. Despite exponential complexity, the practical computation overhead can be ignored since the length of textual prompts is typically 10-100 words [36]. Hence, we can conclude that the proposed mechanisms will not bring a considerable burden to the MASP.

D. Generative Semantic Decoder

After the MASP finishes textual prompt deconstruction and visual prompt segmentation, it can send well-packed semantic information to users. Users then recover the source image by inpainting the pixels that are not transmitted. Note that untransmitted parts only have weak semantic importance. Take "A blue car driving through the city" as an example. The illustration of the road and sky will not affect the semantic correctness of the recovered images since they are not mentioned in the prompt. Moreover, from the image composition perspective, the road and sky are only used as a background to connect semantically strong objects (i.e., cars and cities), slightly influencing the image's aesthetic quality. Hence, the users can adopt various lightweight open-source image inpainting models based on diffusion or generative adversarial networks. We can consider that the image recovery model (e.g., [37] or [38]) is well-trained and shared among all users as the generative semantic decoders.

V. Joint Semantic Encoding and Prompt Engineering

To effectively reflect the human-perceptual experience of the G-SemCom-aided mobile AIGC services, this section first presents a novel metric named JPSQ. Then, we formulate the joint optimization problem for bandwidth allocation and present the ADD algorithm to solve it.

A. JPSQ Definition

Traditionally, to evaluate the effectiveness of SemCom, users can adopt pixel- or structure-level metrics, such as *Mean Square Error* and *Structural Similarity Index Metric*, to measure the similarity between the source and recovered images [39]. However, these metrics can only capture the difference in terms of luminance, contrast, and structure while failing to consider the image semantics. Then, task-oriented metrics for SemCom have been presented, emphasizing whether the proposed SemCom framework can accomplish specific communication tasks [40]. For instance, the authors in [41] utilized the classification accuracy for the observed objects to evaluate the effectiveness of the UAV-based SemCom. Following this principle, we design a novel task-oriented metric for G-SemCom in mobile AIGC called JPSQ. Particularly, in G-SemCom-aided mobile AIGC, the semantic information sent by the MASP undertakes two tasks. First, it guarantees that users can recover images that maintain the same semantics as source images. Meanwhile, recall that the users aim to acquire high-quality AIGC images. Hence, the semantic information also serves as the prompts fed to the generative decoder, facilitating it to recover images with high aesthetic quality. Motivated by this, we jointly consider the semantic similarity and image quality when designing JPSQ. Furthermore, considering that AI-generated images are consumed by human users, we utilize learning-based metrics trained on large-scale human feedback datasets rather than mathematical methods to capture human perceptual similarity and quality. Last but not least, we adopt the Weber-Fechner Law [42] to fuse these two aspects and construct JPSQ.

1) Perceptual Semantic Similarity: To evaluate the perceptual semantic similarity from the user perspective, we utilize the state-of-the-art learning-based metric called *DreamSim* [43]. As shown in Fig. 7(a), the difference between each pair of images (g_0, g_1) is measured by the cosine distance, i.e.,

$$D(g_0, g_1; f_\theta) = 1 - \text{cosine}(f_\theta(g_0), f_\theta(g_1)), \tag{15}$$

where f_θ represents the learnable network that extracts important perceptual semantic features from input images. Such a network is assembled by multiple pretrained models, such as DNIO [44] and OpenCLIP [45] and fine-tuned by Low-Rank Adaptation (LoRA) mechanisms, which align the backbone models with the similarity evaluation task. The smaller the DreamSim value, the more similar the two images are.

2) Perceptual Aesthetic Quality: To measure the aesthetic quality of the recovered images, we adopt a learning-based image assessment framework called NIMA [11]. As shown in Fig 7(b), NIMA converts the image quality measurement to a classification problem, with ten possible classes representing the quality score from 1 to 10. Such classification is realized by a pluggable classifier network, supporting VGG16, Inceptio-v2, and MobileNet [11]. Accordingly, the classification output is defined as $\mathbf{c}(g) = [c_1, c_2, \dots, c_{10}]$, where c_i indicates the probability that the given image g achieves score i. Note that $\sum_{i=1}^{10} c_i = 1$ can be guaranteed since a softmax operation is employed. Finally, the aesthetic quality of image g can be

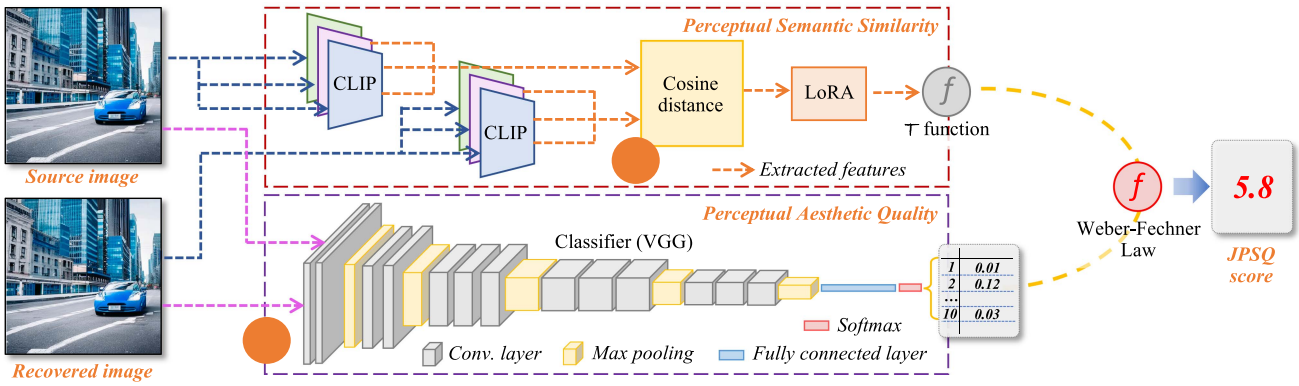

Fig. 7. The definition of JPSQ metric for G-SemCom-aided Mobile AIGC, which consists of perceptual semantic similarity and aesthetic quality. (a): The framework of DreamSim metric. (b): The framework of NIMA metric. Note that these two metrics are flexible. Hence, the CLIP and VGG 16 can be replaced by other models for different application scenarios.

Algorithm 2: The Procedure of ADD Algorithm.

Require: $\mathbf{s}$, N_b, T, η, γ ## *AIGC environment, batch size, diffusion step number, discount factor, and learning rate*
Ensure: $\mathbf{b}$ ## *bandwidth allocation scheme*
procedure ADD Trainings, N_b, T, η, γ
  Initialize networks: policy generation network ϵ_θ, Q-networks Q_{v_1}, Q_{v_2}, $Q^*_{v_1}$, and $Q^*_{v_2}$.
  while not converged **do**
    Initialize random noise $\mathbf{b}_T$; generate bandwidth allocation scheme $\mathbf{b}_0$ by denoising process shown in (20).
    Add exploration noise to $\mathbf{b}_0$.
    Execute resource allocation and calculate utility u by (19).
    Store the record $(\mathbf{s}, \mathbf{b}_0, u)$ in the replay buffer
    Randomly select N_b records
    Update the policy generation network by (22)
    Update the Q-networks by (23)
  end while
end procedure
procedure ADD Inferences, N_b, T, η, γ
  Observe the environment $\mathbf{s}$
  Generate bandwidth allocation scheme $\mathbf{b}_0$
  Return $\mathbf{b}_0$
end procedure

expressed as:

$$Q(g) \sim \mathcal{N}(\mu, \sigma),$$
$$\mu = \sum_{i=1}^{10} i \times c_i, \quad \sigma = \sqrt{\left(\sum_{i=1}^{10} (i-\mu)^2 \times c_i\right)}. \quad (16)$$

In this paper, we utilize μ acquired by NIMA to reflect the aesthetic quality of images.

3) Metric Fusion: Finally, we fuse the perceptual semantic similarity and aesthetic quality by Weber-Fechner Law [42]. Denoting source and recovered images as g_0 and g_1, respectively, JPSQ can be calculated as

$$\mathcal{J}(g_0, g_1) = \mathcal{T}(D(g_0, g_1; f_\theta)) \ln\left(\frac{\omega_0 Q(g_1)}{Q_{\text{th}}}\right), \quad (17)$$

where ω_0 serves as a weighting factor and Q_{th} indicates the minimal image quality required by users. $\mathcal{T}$ is defined as

$$\mathcal{T}(t) = \frac{t_{\max} - t}{t_{\max} - t_{\min}}, \quad (18)$$

where $t_{\min}$ and $t_{\max}$ represent the lower and upper bounds of the DreamSim score, respectively, and $t_{\min}$ is 0. In this paper, we acquire $t_{\max}$ for our case by generating 1000 AIGC images, measuring their DreamSim scores with a pure Gaussian noise, and calculating the average. Function $\mathcal{T}(\cdot)$ plays two roles. First, the denominator inverts the differences reflected by the DreamSim score into similarities. In addition, the effect of the magnitudes can be eliminated.

B. Problem Formulation

In this part, we formulate the joint semantic encoding and prompt engineering problem based on JPSQ. Recall that in our OFDMA-based transmission model, each user can be assigned an RB to receive data from the MASP. However, the overall bandwidth resources of the MASP are limited. Therefore, we intend to optimize the bandwidth allocation among users, acquiring the best trade-off between the overall G-SemCom performance and consumed bandwidth. The optimization problem can be formulated as follows:

$$\max_{\boldsymbol{b_i}} \sum_{i=1}^{N} \left[\left(\omega_1 \mathcal{J}(g_0^i, g_1^{b_i}) \cdot \mathrm{H}(Q(g_1^{b_i}) \geq Q_{\text{th}}) - \omega_2 b_i\right)\right] \quad (19)$$

$$\text{s.t. } 0 \leq b_i \leq \min\{L_i c_i, O|\boldsymbol{S}_i|\}, \forall i \in \{1, 2, \ldots, N\} \quad (19a)$$

where b_i means the bandwidth resources allocated to user U_i. $g_1^{b_i}$ represents the recovered image using the bandwidth of b_i, and g_0^i is the corresponding source image. Note that a step function $\mathrm{H}(\cdot)$ is applied since the images whose quality is lower than the user threshold are unacceptable in AIGC. Additionally, (19b) constrains the range of b_i. Specifically, L_i means the latency threshold between user U_i and the MASP. Hence, $L_i c_i$ represents

the maximum bandwidth that can be transmitted within the required latency. O indicates the bandwidth consumption for transmitting each item in $\boldsymbol{S}_i$. Hence, $O|\boldsymbol{S}_i|$ means the required bandwidth for transmitting the entire semantic information. The upper bound of b_i is the smaller value between these two terms.

C. Components of Attention-Aware Deep Diffusion

Traditionally, the joint optimization problem can be solved by DRL-based methods, such as Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC) algorithms. However, since the state of our problem, i.e., attention maps, is high-dimensional, the existing methods may lack enough exploration ability and yield only sub-optimal solutions. Hence, to realize efficient bandwidth allocation in complex environments, we introduce a deep diffusion module into traditional DRL for policy optimization, forming the ADD algorithm. Next, we demonstrate the components of ADD.

Agent: In the proposed G-SemCom-aided mobile AIGC, the agent represents the MASP, which allocates available bandwidth among multiple users for transmitting semantic information.

State: The state of the mobile AIGC environment takes the form of $\mathbf{s} := [\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_N]$, i.e., the attention-based semantic information of the source images generated for users U_1 to U_N. Note that to reduce the complexity of ADD for representing and learning the states, the original *512* $\times$ *512* attention maps are resized to *16* $\times$ *16*. The state space is a discrete space, using Boolean values to indicate whether a certain pixel is associated with the user prompts.

Action: The action of ADD is a vector $\mathbf{b} := \{b_1, b_2, \ldots, b_N\}$, denoting the bandwidth allocating to each user. With $\mathbf{b}$, the MASP encodes the semantic information for each user U_i ($i \in [1, 2, \ldots, N]$), i.e., calculating the number of pixels that can be sent by b_i/O and sending the pixels following the mechanism stated in Section IV-C.

Policy: The policy refers to the probability of the agent taking action $\mathbf{b}$ at the state $\mathbf{s}$. Particularly, the ADD algorithm adopts a deep diffusion network parameterized by θ to learn the relationship between the input state $\mathbf{s}$ and the output action $\mathbf{b}$ that can optimize the reward defined below. Such a policy network can be expressed as $\pi_\theta(\mathbf{s}, \mathbf{b}) = P(\mathbf{b}|\mathbf{s})$.

Reward: Finally, given the environment state $\mathbf{s}$, The reward of taking action $\mathbf{b}$ can be defined as $R(\mathbf{b}|\mathbf{s}) = \sum_{i=1}^{N}[(\omega_1 \mathcal{J}(g_0^i, g_1^{b_i}) \cdot \mathrm{H}(Q(g_1^{b_i}) \geq Q_{\text{th}}) - \omega_2 b_i)]$. Note that if the constraint is not satisfied, we use a negative reward as the penalty term.

D. Attention-Aware Deep Diffusion for Optimization

The deep diffusion network is introduced to learn the optimized policy $\pi_\theta(\mathbf{s}, \mathbf{b})$ [2]. Following the diffusion principle, the final action $\mathbf{b}_0$ can be generated from random noise $\mathbf{b}_T$ after T steps of denoising, i.e.,

$$\begin{aligned}\pi_\theta(\mathbf{s}, \mathbf{b}) &= p_\theta(\mathbf{b}_{0:T}|\mathbf{s}) \\ &= \mathcal{N}(\mathbf{b}_T; 0, \boldsymbol{I}) \prod_{t=1}^{T} p_\theta(\mathbf{b}_{t-1}|\mathbf{b}_t, \mathbf{s}).\end{aligned} \tag{20}$$

Recall that the definition of $p_\theta(\mathbf{b}_{t-1}|\mathbf{b}_t, \mathbf{s})$ has been shown in (4). Based on (4), the probability of each denoising step can be derived as [31]

$$\mathbf{b}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{b}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\mathbf{x}_t, t\right)\right) + \sigma_t \mathbf{z}, \tag{21}$$

where $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I})$, and according to [31], $\sigma_t = \beta_t$ can achieve good performance. Then, we adopt a Double Deep Q-Network (DDQN) learning architecture [46] to organize the ADD training. Specifically, with the action $\mathbf{b}_0$ generated by policy $\pi_\theta(\mathbf{s}, \mathbf{b})$, ADD applies $\mathbf{b}_0$ in the mobile AIGC environment $\mathbf{s}$ and acquires $R(\mathbf{b}_0|\mathbf{s})$. The solution evaluation network Q_v can help train parameter θ in ϵ_θ and optimize the policy $\pi_\theta(\mathbf{s}, \mathbf{b})$. To do so, Q_v calculates the Q-value of $P(\mathbf{b}_0|\mathbf{s})$, and the optimal θ is the one that can lead to the highest expected Q-value. In this case, the optimal policy generation network can be obtained by

$$\arg\min_{\epsilon_\theta} \mathcal{L}_\epsilon(\theta) = -\mathbb{E}_{\boldsymbol{b}_0 \sim \epsilon_\theta}\left[Q_v\left(\mathbf{s}, \mathbf{b}_0\right)\right]. \tag{22}$$

The Q_v should be trained to predict the best Q-values, which is achieved by minimizing the Bellman operator [46]. In the proposed ADD, there are two Q-networks to be trained, namely Q_{v_1} and Q_{v_2}, with the corresponding target networks $Q^*_{v_1}$ and $Q^*_{v_2}$, respectively. Note that the target networks are used to compute the target for the Q-value updates. The weights of $Q^*_{v_1}$ and $Q^*_{v_2}$ are kept fixed for a number of steps and then periodically updated to match the weights of Q_{v_1} and Q_{v_2}, respectively. By decoupling the targets from the parameters, the learning process can be stabilized. Based on (22), the joint optimization of the two Q-networks can be expressed as minimizing the following expectation

$$\mathbb{E}_{\mathbf{b}_0 \sim \pi^*_\theta}\left[\left\| \begin{array}{l} \left(R\left(\mathbf{b}_0|\mathbf{s}\right) + \gamma \min\limits_{i=1,2} Q_{v'_i}\left(\mathbf{s}, \mathbf{b}_0\right)\right) \\ -Q_{v_j}\left(\mathbf{s}, \mathbf{b}_0\right) \end{array} \right\|^2\right], \tag{23}$$

where γ is the discount factor; $j \in \{1, 2\}$ equals the value of i that leads to the minimum $Q_{v'_i}(\mathbf{s}, \mathbf{b}_0)$. The policy can be optimized with the optimal Q-networks, and the optimal bandwidth allocation scheme $\mathbf{b}$ in any given AIGC state $\mathbf{s}$ can be generated. The detailed training and inference procedures of ADD are shown in Algorithm 2.

E. Complexity Analysis

Here, we analyze the complexity of the proposed ADD algorithm. Suppose that the sizes of the diffusion-based policy network and Q-network are S_p and S_q, respectively. The architectural complexity is $\mathcal{O}(S_p + 2S_q)$. Since generating each bandwidth allocation scheme requests T times diffusion denoising, the policy generation complexity is $\mathcal{O}(TS_p)$. Hence, the overall complexity can be derived as $\mathcal{O}((T+1)S_p + 2S_q)$. Accordingly, supposing that δ epochs are performed, and the batch size is S_b, the computational complexity for training is $\mathcal{O}(\delta S_b((T+1)S_p + 2S_q))$. Finally, the corresponding inference-stage complexity is $\mathcal{O}(S_p)$.

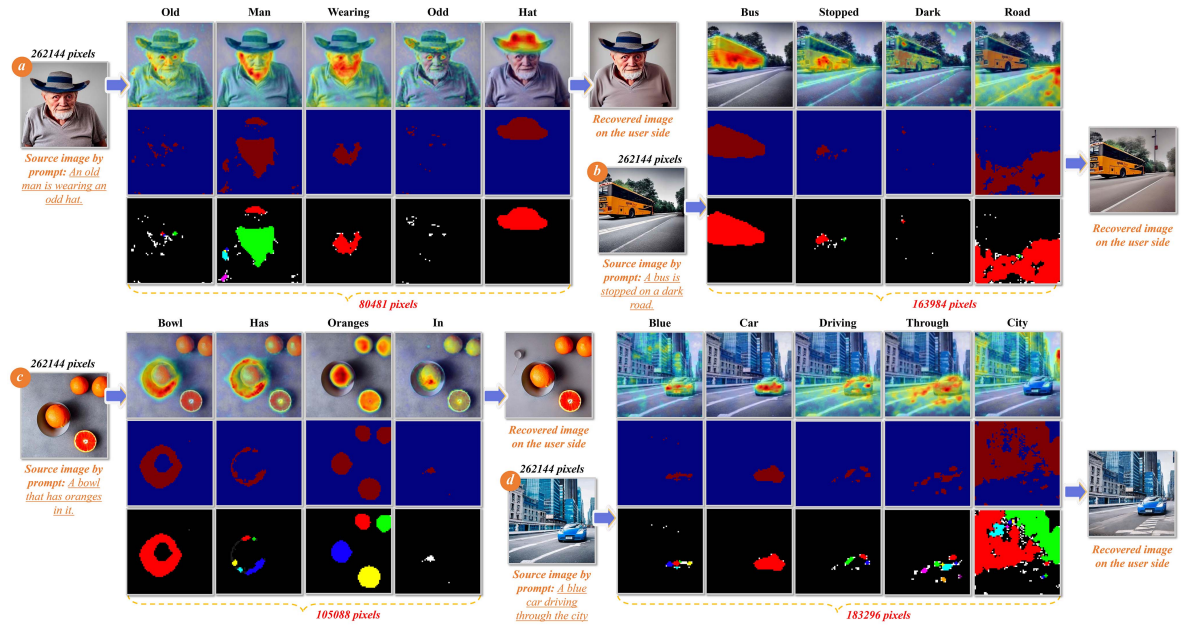

Fig. 8. The case study to illustrate the effectiveness of G-SemCom-aided mobile AIGC. (a): The case of prompt: *An old man is wearing an odd hat*. (b): The case of prompt: *A bus is stopped on a dark road*. (c): The case of prompt: *A bowl that has orange in it*. (d): The case of prompt: *A blue car driving through the city*. For each case, the three rows are cross-modal attention maps, binary attention maps, and clustering results, respectively.

VI. Performance Evaluation

In this section, we implement the proposed G-SemCom framework and build the experimental mobile AIGC system. Then, we conduct extensive experiments that aim to answer two questions: 1) whether the proposed G-SemCom framework for mobile AIGC can effectively reduce the bandwidth consumption of users while ensuring them acquire high-quality AI-generated images and 2) whether the ADD algorithm can efficiently allocate bandwidth resources among users, thus maximizing the overall reward defined by JPSQ. The analysis of the experimental results is also described.

Implementation: To generate high-quality source images, we equip MASPs with Stable Diffusion v2 [26], the state-of-the-art text-to-image AIGC model. The number of diffusion steps is set to 25. On the user side, we deploy a diffusion-based image inpainting model [37] as the generative decoders. Since the users only need to recover the image background with less semantic importance, the diffusion step number is set to 5 to reduce resource consumption. The prompts that the users send to the MASP are selected from the image captions in the *COCO 2017* dataset [47]. We adopt the implementation of the DreamSim metric in [43], which ensembles CLIP, OpenCLIP, and DINO to construct the backbone model. For NIMA, we utilize MobileNet to implement the image quality classifier and load the pretrained model weights from [48]. All the steps of Algorithm 1 are packed into a pipeline written in Python, based on *diffusers*, *daam*, *spacy*, and *sklearn* libraries. Finally, we leverage *PyTorch* to implement the proposed ADD algorithm, combining the basic DDQN architecture in [49] and our deep diffusion module for policy generation.

Testbed. The experiments are conducted on a server with an NVIDIA RTX A5000 GPU with 24GB of memory and an AMD Ryzen Threadripper PRO 3975WX 32-Core CPU with 263GB of RAM. The operating system is Ubuntu 20.04 LTS with *PyTorch* 2.0.1. We utilize this server to simulate one MASP and multiple uniformed distributed mobile users. The OFDMA transmission model between the MASP and users is implemented based on [8].

TABLE III
The Summary of Experimental Settings [8]

Symbol	Description	Value
γ	Discount factor of ADD	0.95
L_i	Latency threshold of U_i	5s
O	Bandwidth consumption for transmitting each pixel	1
Q_{th}	Threshold of aesthetic quality	4.9827
η	Learning rate of ADD	10^{-4}
T	Diffusion step number of ADD	5 & 6
ξ	Threshold of attention value	0.9, 0.8, 0.5
ω_0	Weighting factor in Eq. (17)	1.25
ω_1	Weighting factor in Eq. (19)	500
ω_2	Weighting factor in Eq. (19)	0.05

Experimental Settings. The important environmental and hyperparameter configurations in terms of the proposed G-SemCom pipeline and ADD are shown in Table III.

A. Effectiveness of G-SemCom Framework

In this part, we evaluate the effectiveness of the proposed G-SemCom framework for mobile AIGC.

1) Case Study: First, Fig. 8 illustrates four cases where the users request different images from the MASP. We can observe that the cross-modal attention maps can effectively associate any given word to certain pixels of the generated images. However,

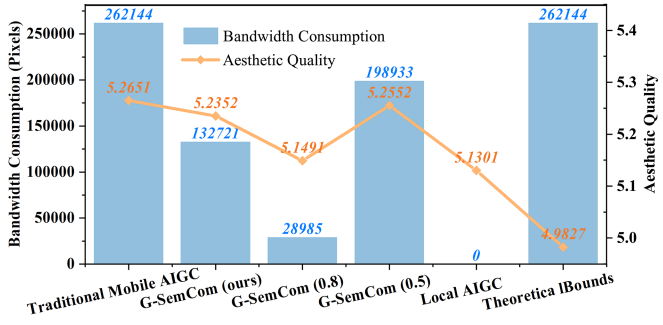

Fig. 9. The performance of different AIGC paradigms in terms of bandwidth consumption and image quality. Note that *Bounds* include the theoretical upper and lower bounds of the bandwidth consumption and NIMA score, respectively.

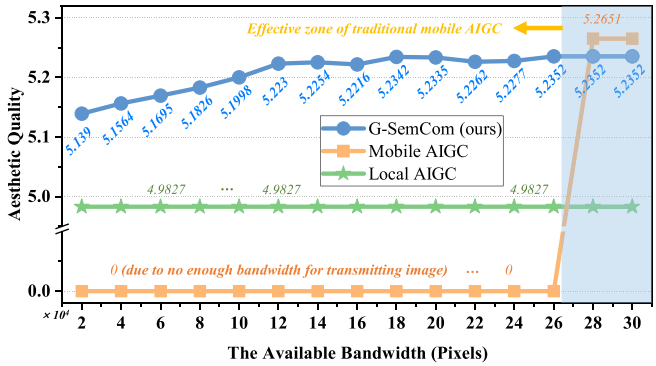

Fig. 10. The service robustness of different AIGC paradigms to channel error. Note that the blue bar indicates the effective zone of mobile AIGC since it can only provide the users with useful images within this range.

the attention distribution of $\boldsymbol{ADJ}$- and $\boldsymbol{VERB}$-typed words, e.g., `[old]` in case *a* and `[stopped]` in case *b*, are scattered due to less clear semantic meaning. Take `[old]` as an example. Its attention covers the entire image, while only those pixels highlighting the character's facial features are meaningful. To this end, we utilize (14) to filter out attention points without clear semantic meaning, forming the binary attention maps. Note that we assign ξ in (14) with different values according to the general semantic importance of different part-of-speech types. Specifically, for $\boldsymbol{PROPN}$-typed, $\boldsymbol{NN}$-typed, and other words, ξ equals *0.9*, *0.8*, and *0.5*, respectively. We can observe that the meaningless pixels are effectively removed. In contrast, the pixels with strong semantic meaning, such as the pixels associated with words `[hat]` and `[bus]`, are fully maintained. Atop binary attention maps, the proposed attention clustering algorithm based on DBSCAN can further remove the noise and noisy clusters, which are too small and will affect image recovery. Finally, the recovered images can hold the high quality of source images, with almost no perceptual quality difference. Meanwhile, the semantic information takes only 80481, 163984, 105088, and 183296pixels compared with the 512×512 source image with 262144pixels, achieving 69.3%, 27.4%, 60.0%, and 30.1% reduction, respectively. More in-depth experiments on bandwidth reduction are shown below.

2) G-SemCom Performance: Fig. 9 illustrates the average performance of the traditional and our proposed G-SemCom-aided mobile AIGC for generating 1000 images. Note that we take the number of pixels as the bandwidth unit, which can circumvent the errors caused by different standards for packing images, e.g., . *jpg* and . *png*. From Fig. 9, we can observe that compared with traditional mobile AIGC, the proposed G-SemCom [with $\xi = \{0.9, 0.8, 0.5\}$] can reduce the bandwidth consumption of the users by 49.4% on average, while the average image quality, measured by NIMA score, drops only 0.0299. Moreover, if ξ is relaxed to 0.8 or tightened to 0.5, the bandwidth and quality will further decrease and increase, respectively, exhibiting the outstanding flexibility of our proposal. Recall that the generative decoder on the user side is a lightweight image inpainting model, which inpaints the masked image with 5 diffusion steps. We then provide fully masked images to the decoder, thereby exploring the quality of the images that the users can generate locally using the same computation resources with G-SemCom. As shown in Fig. 9, the image quality of local AIGC is only 5.1301 due to less powerful models and fewer diffusion steps. Given the lower bound of the NIMA score is 4.9827, the decrement from 5.2651 to 5.1301 means that the image quality drops by 47.8%. Note that such a lower bound is acquired by generating 1000 pure Gaussian noise images, meaning their NIMA scores, and taking the minimum value.

3) Service Robustness: Besides saving bandwidth, another significant advantage of our G-SemCom framework is enhancing the robustness of mobile AIGC services to channel errors. Note that channel error means that the connection between the user and MASP is interrupted due to unexpected circumstances, resulting in only a part of the semantic information being transmitted. Traditionally, users need to download the entire image from the MASP. Hence, the channel error will cause transmission failure, in which the users can only receive broken images. Assisted by G-SemCom, regardless of how many bits have been transmitted, users can recover full images using the generative decoders, which is extremely important if the application has strict requirements for latency and cannot tolerate re-transmission. As shown in Fig. 10, even though only 20000pixels are transmitted, the average NIMA score reaches 5.1390, which exceeds that of the local AIGC (i.e., 5.1301). With the increasing number of transmitted pixels, the quality of the recovered images grows gradually. In contrast, the image quality of traditional mobile AIGC remains 0 until all the image information can be transmitted. Apart from enhancing service robustness, the proposed ADD can further determine the number of pixels to be transmitted to achieve the best JPSQ bandwidth balance in the given state. The corresponding experiments are discussed in Section IV-B.

4) Resource Consumption on the User Side: Here, we explore the resource consumption of users to operate the generative decoder. To this end, we first evaluate the relationship between diffusion steps adopted by generative decoders and the image quality. As shown in Fig. 11, the increasing diffusion step number fails to improve the image quality linearly because the pixels yet to be recovered only have weak semantics. Take case *a* in Fig. 8 as an example. Compared with the man's face, which

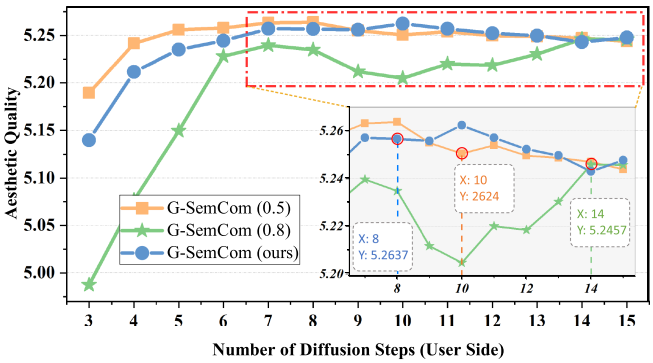

Fig. 11. The aesthetic quality with the number of diffusion steps on the user side.

TABLE IV
THE COMPUTATION AND MEMORY CONSUMPTION OF THE MASP AND USERS

Stakeholder	GPU time	CPU time	GPU memory
MASP	2.955s	3.181s	7846 MB
User	1.078s	1.170s	3919 MB

accommodates rich semantics (e.g., facial features, expression, and beard), human perception is much less sensitive to the cloth and background. Therefore, increasing resources for rendering such parts cannot improve the human-perceptual quality of the recovered image. For the recommended ξ scheme, i.e., {0.9, 0.8, 0.5}, setting the diffusion step number as 5-7 can already lead to satisfying image quality. Table IV illustrates the resource overhead when the diffusion step number equals 5. We can observe that compared with generating contents, decoding only consumes 63.4%, 63.3%, and 51.1% of CPU time, GPU time, and GPU memory, respectively.

B. Ablation Study

In this part, we perform an in-depth ablation study, aiming to investigate the effectiveness of each proposed step performed by the MASP in Algorithm 1.

1) Binary Attention Map: This operation refers to filtering the less important attention points from the original attention maps, whose major purpose is reducing semantic information size. Fig. 12(a) illustrates the case *a* in Fig. 8 with and without filtering. Note that we adopt the recommended ξ scheme, i.e., {0.9, 0.8, 0.5} when constructing the binary attention maps. We can observe that without filtering, the attention maps contain all the pixels, while the majority of them only have weak semantic meaning. In contrast, by setting the threshold ξ and forming the binary attention map, the size of the semantic information can be reduced by 69.3%.

2) Dependency Parsing: Dependency parsing facilitates the MASP in evaluating the importance of each word in the user prompts. Without dependency parsing, the MASP can only randomly select pixels when packing the semantic information. As shown in Fig. 12(b), the resulting semantic information is blurred, with a NIMA score of 4.9102. In contrast, we first understand which pairs of words are correlated by dependency

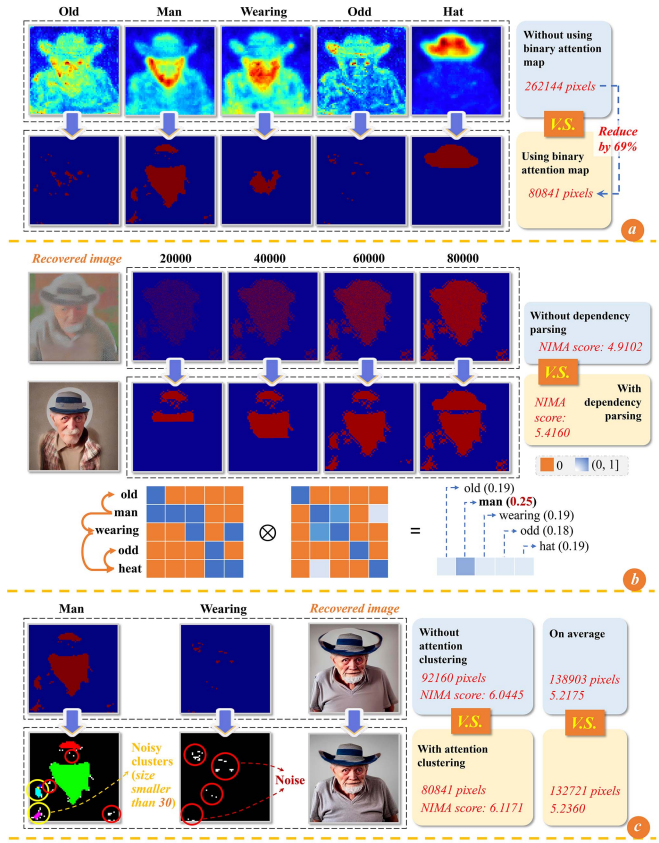

Fig. 12. The ablation study. (a): The inspection on binary attention map. (b): The inspection on dependency parsing. (c): The inspection on attention clustering.

parsing. For instance, Fig. 12(b) shows the dependencies existing in our case. Then, the $\boldsymbol{C}^*$ and $\boldsymbol{D}^*$ matrices can be established. Finally, the semantic importance can be calculated, which is $\mathbf{s} = [0.19, 0.25, 0.19, 0.18, 0.19]$. The packing of semantic information can then follow the order of $\mathbf{s}$. We can observe that the pixels associated with `[man]` are prioritized. In this case, even though only 20000pixels are transmitted, the core semantic information is well preserved in the recovered image, resulting in a much higher NIMA score.

3) Attention Clustering: Finally, attention clustering means adopting DBSCAN to remove noise and noisy clusters, which do not carry enough image semantics due to small sizes. Moreover, they might affect image recovery since the generative decoder can hardly generate. Fig. 12(c) illustrates the clustering results of `[man]` and `[wearing]`. The pixels marked by red and yellow circles are noise and noisy clusters, respectively. We can observe the noise associated with the word `[wearing]` affects the recovery of the man's hat, decreasing the image quality from 6.1171 to 6.0445. Using the 1000 images of Fig. 8, by performing attention clustering, the average bandwidth consumption reduces by 6182pixels while the quality increases by 0.0185.

C. Efficiency of ADD

Then, we study the efficiency of the proposed ADD algorithm in optimizing the resource allocation scheme.

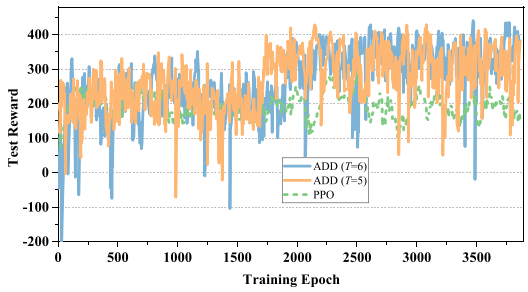

Fig. 13. The training curves of the proposed ADD (with $T = 5$ and 6) and baseline PPO.

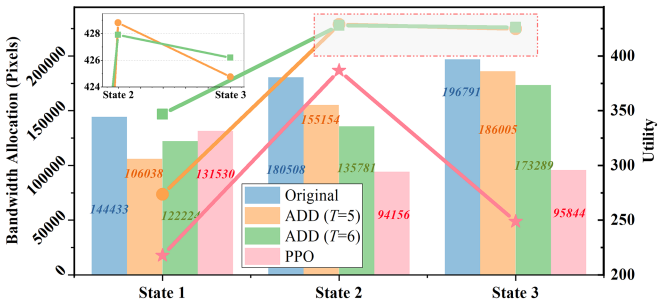

Fig. 14. The bandwidth allocation schemes and resulting reward in different states. Note that *state* represents the semantic information $\boldsymbol{S}$ of a mobile user.

1) ADD Training: First, Fig. 13 illustrates the training curves of the ADD algorithm with 5 and 6 steps of diffusion. Similar to Section VI-A, the training dataset is constructed by images generated from captions in *COCO 2017* dataset [47]. Additionally, to demonstrate the superiority of our proposal, we adopt a standard DRL algorithm as the baseline, called PPO [50]. As shown in Fig. 13, the ADD (T=5) and PPO take a similar time to converge. The ADD (T =6) converges slower while achieving almost 90% higher rewards than PPO. This can be explained by the enhanced ability of ADD to explore the environment since an exploration noise is added to the generated bandwidth allocation scheme during each training iteration. Hence, the training process can avoid getting trapped in sub-optimal solutions. However, the number of diffusion steps is not always better. This is because ADD will lose its ability to explore the state effectively, as excessive denoising might lead to overfitting.

2) Optimization of JPSQ: Then, we investigate the efficiency of ADD in scheduling bandwidth. Fig. 14 shows the bandwidth allocation schemes for three images whose semantic information sizes are 144433, 180508, and 196791pixels, respectively. We can observe that the proposed ADD algorithm with $T = 5$ and 6 can achieve 32.2% and 40.9% higher utility [defined in (19)] than PPO on average. Such results demonstrate that our algorithm can better balance the human-perceptual AIGC service quality and bandwidth costs. Take State *3* as an example. The PPO algorithm only assigns 94156pixels for transmitting the semantic information. In this case, even though bandwidth consumption is low, the DreamSim score is 0.198, meaning the recovered image holds 80.2% similarity with the source image at the semantic level. In contrast, the ADD ($T = 6$) algorithm uses 173289pixels

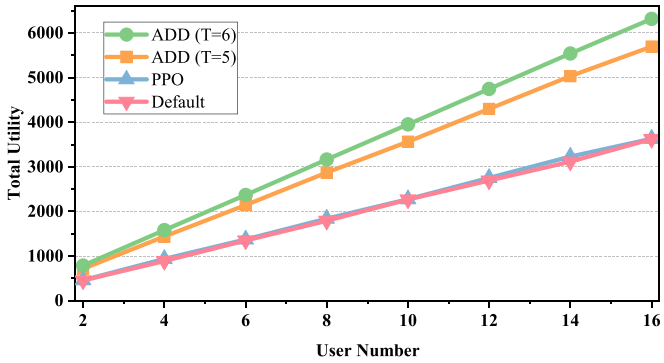

Fig. 15. The total utility with increasing user number. *Fixed allocation* means allocating bandwidth for the entire semantic information without optimization.

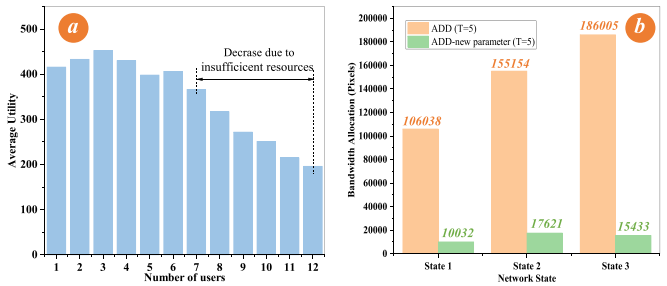

Fig. 16. (a): The average utility with varying number of users connected to one MASP. (b): The generated bandwidth allocation schemes with the updated weighting factors.

to achieve 96.7% similarity and a 5.56 NIMA score, resulting in 72.5% higher overall utility. Furthermore, we evaluate the total utility with the increasing number of users, as shown in Fig. 15. From Fig. 15, we can observe that the ADD algorithm with $T =$ 5 and 6 significantly outperforms the PPO and default schemes. Efficient bandwidth allocation is crucial for mobile AIGC to meet users' demand for high-quality AIGC outputs with limited bandwidth resources.

D. Scalability Analysis

Finally, we analyze the scalability of our proposed framework to support large-scale mobile AIGC networks with dense users. First, we conclude that our proposals can scale linearly with the increasing number of MASPs. This is because the G-SemCom pipeline and ADD algorithm are deployed separately on each MASP. Since all the MASPs serve mobile users independently, the network-wide utility can increase linearly with the network scale. Such scalability enables our framework to support large-scale mobile AIGC applications.

Nonetheless, if the number of users connected to one MASP keeps increasing and the required bandwidth resources exceed the MASP's capability, the average utility of connected users will decrease. To demonstrate this statement, we suppose a batch of users (the number increases from 1 to 12) submit service requests to one MASP simultaneously, and the tolerance for transmission latency is 2s. Different from Fig. 15, the bandwidth of MASP is constrained. As shown in Fig. 16(a), initially, the

average utility remains stable. Once the total required resources exceed the MASP's capability, the MASP should suspend some service requests, causing a drop in average utility. To alleviate such a performance drop, a mixture of expert architecture [51] is worth developing, which dynamically schedules MASPs for each user. Instead of working independently, these MASPs can collaborate to undertake network-wide workload, realizing the load-balancing.

E. Discussions and Future Work

1) Weighting Factors Setting: Here, we explore the impact of weighting factors ω_0, ω_1, and ω_2 on performance. To do so, we reduce the values of ω_0 and ω_1 to 1 and 200, respectively. Meanwhile, the value of ω_2 is increased to 0.2. Such adjustment represents the shift from seeking JPSQ-bandwidth balance to emphasizing saving bandwidth resources. As shown in Fig. 16(b), guided by the updated reward, the ADD algorithm tends to allocate a small bandwidth to each user. In conclusion, these weighting factors should be configured according to the specific application scenarios or the requirements of users.

2) Security and Privacy: User privacy and data security are critical concerns in mobile AIGC. First, AIGC outputs can be regarded as digital assets and can be traded in the market [52]. In this case, attackers can plagiarize AIGC works generated by MASP and sell them to make profits. In addition, attackers can obtain sensitive information by eavesdropping on the communication link between MASP and users. To defend against plagiarism, a digital watermark can be embedded into the AIGC output [52], which serves as a unique identifier that proves ownership and authenticity of the digital asset. Additionally, blockchains can help to maintain an immutable record of AIGC ownership [52]. To protect the communications between the MASP and users, advanced privacy protection techniques such as differential privacy and covert communication can be applied.

3) Adaptability: Finally, we discuss the adaptability of our proposals to other AIGC forms. It can be concluded that the proposed G-SemCom framework and ADD algorithm are applicable if two requirements are satisfied: *i) The AIGC model used by MASPs involves cross-modality attention, and ii) the users utilize textual prompts.* Since AIGC inferences are typically cross-modality, attention mechanisms are widely adopted by mainstream AIGC models. Similarly, most AIGC models, e.g., ChatGPT and Stable Diffusion, leverage textual prompts to guide generation since human users are used to describe their requirements in natural language.

Here, we showcase how to apply G-SemCom to text-to-video AIGC, where MASPs generate transition videos by StoryDiffusion [29]. First, since cross-attention is utilized by StoryDiffusion to control the generation of each frame [29], we can fetch the attention maps accordingly. Following the G-SemCom pipeline, we then analyze the prompt logic, filtering the words/segments with vital semantics. Note that given the complexity of prompts used for video generation, we can perform fine-grained textual analysis, e.g., using a knowledge graph. Afterward, each video frame can be regarded as an individual image. Only the parts associated with the most semantically important prompt words/segments are extracted and transmitted to save transmission resources. Finally, on the receiver side, the semantic information can be decoded by a lightweight AIGC model to recover the video. We consider the detailed mechanism reconfiguration for supporting other AIGC forms as future work.

VII. Conclusion

In this paper, we have presented a novel G-SemCom framework for mobile AIGC, where the MASP only sends compressed semantic information, and users adopt a lightweight generative decoder to recover high-quality images. Specifically, by cross-modal attention maps, the MASP can filter the pixels with the highest semantic importance for transmission. Moreover, considering the bandwidth limitation, we have defined a joint optimization problem to allocate bandwidth among users. Utilizing attention maps and the diffusion principle, we have designed the ADD algorithm to maximize the human perceptual JPSQ. Extensive experiments demonstrate that our G-SemCom framework can reduce bandwidth consumption by 49.4% while ensuring image quality on the user side. In addition, the ADD has significantly outperformed traditional DRL, striking great balances between bandwidth and JPSQ in mobile environments.

References

[1] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surv. Tut.*, vol. 26, no. 2, pp. 1127–1170, Second Quarter, 2024.
[2] H. Du et al., "Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization," *IEEE Commun. Surv. Tut.*, early access, May 10, 2024, doi: 10.1109/COMST.2024.3400011.
[3] Y. Liu et al., "Optimizing mobile-edge AI-generated everything (AIGX) services by prompt engineering: Fundamental, framework, and case study," *IEEE Netw.*, early access, Nov. 28, 2023, doi: 10.1109/MNET.2023.3335255.
[4] "The word's first on-device stable diffusion version by qualcomm," 2023. [Online]. Available: https://www.qualcomm.com/news/onq/2023/02/worlds-first-on-device-demonstration-of-stable-diffusion-on-android
[5] Y. -H. Chen et al., "Speed is all you need: On-device acceleration of large diffusion models via GPU-aware optimizations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 4650–4654.
[6] H. Du et al., "Exploring collaborative distributed diffusion-based AI-Generated Content (AIGC) in wireless networks," *IEEE Netw.*, vol. 38, no. 3, pp. 178–186, May 2024.
[7] J. Wen et al., "Freshness-aware incentive mechanism for mobile ai-generated content (AIGC) networks," in *Proc. IEEE/CIC Int. Conf. Commun. China*, 2023, pp. 1–6.
[8] Y. Wang et al., "Performance optimization for semantic communications: An attention-based reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2598–2613, Sep. 2022.
[9] R. Cheng, Y. Sun, D. Niyato, L. Zhang, L. Zhang, and M. A. Imran, "A wireless ai-generated content (AIGC) provisioning framework empowered by semantic communication," 2023, *arXiv:2310.17705*.
[10] Y. Lin et al., "A unified framework for integrating semantic communication and AI-generated content in metaverse," *IEEE Netw.*, vol. 38, no. 4, pp. 174–181, Jul. 2023.
[11] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.
[12] X. Mu and Y. Liu, "Exploiting semantic communication for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, Aug. 2023.
[13] Y. Li et al., "Snapfusion: Text-to-image diffusion model on mobile devices within two seconds," 2023, *arXiv:2306.00980*.
[14] "Google on-device stable diffusion," 2023. [Online]. Available: https://developers.google.com/mediapipe
[15] "Apple on-device stable diffusion," 2023. [Online]. Available: https://github.com/apple/ml-stable-diffusion

[16] R. Tang et al., "What the DAAM: Interpreting stable diffusion using cross attention," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 5644–5659.
[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[18] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," 2023, *arXiv:2306.04321*.
[19] G. Liu et al., "Semantic communications for artificial intelligence generated content (AIGC) toward effective content creation," 2023, *arXiv:2308.04942*.
[20] Q. He, H. Yuan, D. Feng, B. Che, Z. Chen, and X.-G. Xia, "Robust semantic transmission of images with generative adversarial networks," in *Proc. IEEE Glob. Commun. Conf.*, 2022, pp. 3953–3958.
[21] F. Jiang et al., "Large AI model-based semantic communications," 2023, *arXiv:2307.03492*.
[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, 2023.
[23] A. Haviv, J. Berant, and A. Globerson, "BERTese: Learning to Speak to BERT," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 3618–3623.
[24] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing *NLP," EMNLP-IJCNLP*, 2019, pp. 2153–2162.
[25] B.-K. Kim, H.-K. Song, T. Castells, and S. Choi, "On architectural compression of text-to-image diffusion models," *ArXiv:2305.15798*, 2023.
[26] Stable diffusion model. 2023. [Online]. Available: https://stability.ai/blog/stable-diffusion-public-release
[27] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Cache-aided NOMA mobile edge computing: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6899–6915, Oct. 2020.
[28] "The Sora model," 2024. [Online]. Available: https://openai.com/index/sora/
[29] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou, "StoryDiffusion: Consistent self-attention for long-range image and video generation," 2024, *arXiv:2405.01434*.
[30] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[32] "The dependency parsing for texts," 2023. [Online]. Available: http://nlpprogress.com/english/dependency_parsing.html
[33] D. Deng, "DBSCAN clustering algorithm based on density," in *Proc. Int. Forum Elect. Eng. Automat.*, 2020, pp. 949–953.
[34] "The OPTICS clustering algorithm," 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html
[35] M. Ester, H. -P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.
[36] "The length of textual prompts," 2024. [Online]. Available: https://www.aipromptsearch.net/en/stable-diffusion-prompts
[37] "The image inpainting model based on diffusion," 2023. [Online]. Available: https://huggingface.co/runwayml/stable-diffusion-inpainting
[38] D. Cha and D. Kim, "DAM-GAN: Image inpainting using dynamic attention map based on fake texture detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4883–4887.
[39] T. Zhu et al., "How to evaluate semantic communications for images with vitscore metric?," 2023, *arXiv:2309.04891*.
[40] F. Liu, W. Tong, Y. Yang, Z. Sun, and C. Guo, "Task-oriented image semantic communication based on rate-distortion theory," 2022, *arXiv:2201.10929*.
[41] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, Aug. 2022.
[42] "The introduction to Weber-Fechner law," 2023. [Online]. Available: https://www.raggeduniversity.co.uk/wp-content/uploads/2018/03/Weber-Fechner-Law.pdf
[43] S. Fu et al., "DreamSim: Learning new dimensions of human visual similarity using synthetic data," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 50742–50768.
[44] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9630–9640.
[45] M. Cherti et al., "Reproducible scaling laws for contrastive language-image learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2818–2829.
[46] Y. Fan, J. Ge, S. Zhang, J. Wu, and B. Luo, "Decentralized scheduling for concurrent tasks in mobile edge computing via deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 23, no. 4, pp. 2765–2779, Apr. 2024.
[47] "COCO thedataset," 2023. [Online]. Available: https://cocodataset.org/#home
[48] "The weights for NIMA based on mobilenet," 2023. [Online]. Available: https://github.com/idealo/image-quality-assessment
[49] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.
[50] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv: 1707.06347*.
[51] G. Cheng, L. Dong, W. Cai, and C. Sun, "Multi-task reinforcement learning with attention-based mixture of experts," *IEEE Robot. Automat. Lett.*, vol. 8, no. 6, pp. 3812–3819, Jun. 2023.
[52] C. Chen, Y. Li, Z. Wu, M. Xu, R. Wang, and Z. Zheng, "Towards reliable utilization of AIGC: Blockchain-empowered ownership verification mechanism," *IEEE Open J. Comput. Soc.*, vol. 4, pp. 326–337, 2023.

Yinqiu Liu received BEng degree from the Nanjing University of Posts and Telecommunications, China, in 2020, and the MSc degree from the University of California, Los Angeles, in 2022. He is currently working toward the PhD degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His current research interests include wireless communications, mobile AIGC, and generative AI.

Hongyang Du (Graduate Student Member, IEEE) received the BEng degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, in 2021, and the PhD degree from Nanyang Technological University, Singapore, in 2024. He is an assistant professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong. He serves as the editor-in-chief assistant of *IEEE Communications Surveys & Tutorials* (2022-2024), and the guest editor for *IEEE Vehicular Technology Magazine*. He is the recipient of the IEEE Daniel E. Noble Fellowship Award from the IEEE Vehicular Technology Society, in 2022, the IEEE Signal Processing Society Scholarship from the IEEE Signal Processing Society, in 2023, the Singapore Data Science Consortium (SDSC) Dissertation Research Fellowship, in 2023, and NTU Graduate College's Research Excellence Award, in 2024. His research interests include edge intelligence, generative AI, semantic communications, and network management.

Dusit Niyato (Fellow, IEEE) received the BEng degree from the King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand, and the PhD degree in electrical and computer engineering from the University of Manitoba, Canada. He is a professor with the College of Computing and Data Science, at Nanyang Technological University, Singapore. His research interests are in the areas of mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.

Jiawen Kang (Senior Member, IEEE) received the PhD degree from the Guangdong University of Technology, China, in 2018. He has been a postdoc with Nanyang Technological University, Singapore from 2018 to 2021. He currently is a full professor with the Guangdong University of Technology, China. His research interests mainly focus on blockchain, security, and privacy protection in wireless communications and networking.

Zehui Xiong (Senior Member, IEEE) received the PhD degree from Nanyang Technological University (NTU), Singapore. He is currently an assistant professor with the Singapore University of Technology and Design, Singapore. His research interests include wireless communications, Internet of Things, blockchain, edge intelligence, and Metaverse. Recognized as a highly cited researcher, he has published more than 200 research papers in leading journals and flagship conferences. He has been honored with Forbes Asia 30u30, IEEE Early Career Researcher Award for Excellence in Scalable Computing, IEEE Technical Committee on Blockchain and Distributed Ledger Technologies Early Career Award, IEEE Internet Technical Committee Early Achievement Award, IEEE TCSVC Rising Star Award, IEEE TCI Rising Star Award, IEEE TCCLD Rising Star Award, IEEE Best Land Transportation Paper Award, IEEE CSIM Technical Committee Best Journal Paper Award, IEEE SPCC Technical Committee Best Paper Award, and IEEE VTS Singapore Best Paper Award. He is now serving as the associate director of Future Communications R&D Programme.

Shiwen Mao (Fellow, IEEE) received the PhD degree in electrical and computer engineering from Polytechnic University, Brooklyn, NY, USA, in 2004. He is a professor and Earle C. Williams Eminent Scholar, and director of the Wireless Engineering Research and Education Center, Auburn University, Auburn, AL. His research interest includes wireless networks, multimedia communications, and smart grid. He received the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award, in 2019 and NSF CAREER Award, in 2010. He is a co-recipient of the 2021 Best Paper Award of Elsevier/KeAi Digital Communications and Networks Journal, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the 2018 Best Journal Paper Award and the 2017 Best Conference Paper Award from IEEE ComSoc MMTC, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a co-recipient of the Best Paper Awards from IEEE ICC 2022 and 2013, IEEE GLOBECOM 2019, 2016, and 2015, and IEEE WCNC 2015, and the Best Demo Awards from IEEE INFOCOM 2022 and IEEE SECON 2017.

Ping Zhang (Fellow, IEEE) received the PhD degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1990. He is currently a professor with the Schoolof Information and Communication Engineering, BUPT, where he serves as the director of the State Key Laboratory of Networking and Switching Technology and also the director of the Department of Broadband Communication, Peng Cheng Laboratory, Shenzhen, China. He is an academician of the Chinese Academy of Engineering. He served as a chief scientist of National Basic Research Program (973 Program) and an expert in Information Technology Division of National High-Tech RD Program (863 Program). He is a member of IMT-2020 (5G) Experts Panel, Experts Panel for China's 6G Development, and Consultant Committee on International Cooperation of National Natural Science Foundation of China.

Xuemin Shen (Fellow, IEEE) received the PhD degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is a University professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, AI for networks, and vehicular networks. He is a registered professional engineer of Ontario, Canada, an Engineering Institute of Canada fellow, a Canadian Academy of Engineering fellow, a Royal Society of Canada fellow, a Chinese Academy of Engineering Foreign member, and a distinguished lecturer of the IEEE Vehicular Technology Society and Communications Society. He received "West Lake Friendship Award" from Zhejiang Province, in 2023, President's Excellence in Research from the University of Waterloo, in 2022, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT), in 2021, the R.A. Fessenden Award, in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario), in 2019, James Evans Avant Garde Award, in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award, in 2015 and Education Award, in 2017 from the IEEE Communications Society (ComSoc), and Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He has also received the Excellent Graduate Supervision Award, in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA), in 2003 from the Province of Ontario, Canada.