Agent-Driven Generative Semantic Communication With Cross-Modality and Prediction

Wanting Yang, Zehui Xiong, *Senior Member, IEEE*, Yanli Yuan, Wenchao Jiang, *Member, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, and Mérouane Debbah, *Fellow, IEEE*

Abstract— In the era of 6G, with compelling visions of intelligent transportation systems and digital twins, remote surveillance is poised to become a ubiquitous practice. Substantial data volume and frequent updates present challenges in wireless networks. To address these challenges, we propose a novel agent-driven generative semantic communication (A-GSC) framework based on reinforcement learning. In contrast to the existing research on semantic communication (SemCom), which mainly focuses on either semantic extraction or semantic sampling, we seamlessly integrate both by jointly considering the intrinsic attributes of source information and the contextual information regarding the task. Notably, the introduction of generative artificial intelligence (GAI) enables the independent design of semantic encoders and decoders. In this work, we develop an agent-assisted semantic encoder with cross-modality capability, which can track the semantic changes, channel condition, to perform adaptive semantic extraction and sampling. Accordingly, we design a semantic decoder with both predictive and generative capabilities, consisting of two tailored modules. Moreover, the effectiveness of the designed models has been verified using the UA-DETRAC dataset, demonstrating the performance gains of the overall A-GSC framework in both energy saving and reconstruction accuracy.

Index Terms— Semantic communication, video streaming, diffusion model, deep reinforcement learning, semantic sampling.

I. INTRODUCTION

AS HUMAN society advances towards remote management, coupled with emerging 6G applications like intelligent transportation system, digital twins and metaverses, remote surveillance is poised to become a ubiquitous practice [2]. The substantial data volume and frequent updates emphasize the challenges of redundancy and rigidity of information representation and transmission in conventional communication. Fortunately, 6G ushers in a new era of semantic communication (SemCom) [3], paving the way for more efficient and streamlined transmission.

In literature, two prominent and non-overlapping branches of semantic communication research for remote surveillance have emerged: deep learning (DL)-based semantic compression and age of information (AoI)-guided semantic sampling. DL-based SemCom concentrates on each individual data unit, performing task-agnostic semantic compression first and then decompression [4]. Meanwhile, AoI-guided SemCom prioritizes information freshness, filtering out outdated and irrelevant information, which some literature regards as a form of compression in the *temporal* domain [5]. To succinctly distinguish these two compression approaches operating in different domains, hereinafter, we refer to the latter as semantic sampling and the former as semantic extraction.

Both branches of research have demonstrated satisfactory performance and improvements in energy efficiency given constrained resources. Nonetheless, it is crucial to emphasize that these methods merely consider the intrinsic attributes of source information such as compressibility of high-dimensional information and information freshness, which implies that most SemCom research in remote surveillance primarily focuses on the *semantic* level.[1] In fact, contextual information regarding a specific task tends to play a more crucial role in reducing transmission redundancy [7]. Specifically, only goal-related semantic information is necessary for the destination. Taking traffic monitoring for instance, transmitting building details serves no purpose. Similarly, in scenarios like disaster monitoring, the value of intrinsic fresh information may be limited, while changing or unusual information holds greater benefits. Overall, integrating contextual information into the system design empowers SemCom to achieve a heightened level of *effectiveness* with higher compression efficiency.

---

[1]Shannon and Weaver suggest that SemCom encompasses two distinct levels [6]: semantic level, which focuses on the meaning of the transmitted symbols and effectiveness level, which is concerned with the final performance achieved by communication tasks.

To the best of our knowledge, there is a notable absence of research on effectiveness-level SemCom for surveillance scenarios.

In response to the identified research gap, we propose a novel modular agent-driven generative SemCom (A-GSC) framework, which envisions a holistic redesign of information processing, sampling, transmission, and reconstruction processes. By incorporating intrinsic and extrinsic information, we achieve a remarkably high level of compression efficiency by seamlessly cascading semantic extraction and semantic sampling. This integration produces a synergistic effect greater than the sum of its parts. An preliminary study has conceived a semantic change driven generative SemCom (SCGSC) framework for the remote image update system [1].[2] Therein, the introduction of generative artificial intelligence (GAI) transforms the semantic encoding into the determination of the PROMPT. The remarkable explainability of PROMPT allows for independent design and optimization of semantic encoders and decoders. Specifically, in SCGSC framework, the semantic encoder first performs target segmentation to acquire the relevant semantic information called semantic map, then collectively considers the *semantic change* and the *AoI* to evaluate the value of information (VoI), and finally completes the semantic sampling based on thresholds. Correspondingly, the semantic decoder takes the semantic map and static background of the remote scene as inputs to reconstruct the scene based on the diffusion model.

Building on the preliminary framework in [1], this work focuses on a more realistic scenario. To characterize richer semantic information with less data, we introduce cross-modality capabilities into the SemCom design. Additionally, to present the illusion of real-time updates of remote scenes, the enhanced A-GSC framework integrates prediction capabilities, allowing the destination to continuously display remote scenes even during moments without sampling. Three key improvements distinguish this work from the SCGSC framework, as outlined below:

- Instead of using semantic map, the semantic information (i.e., layout information) of scenes is transmitted in the form of text with a lower data volume. In addition to the vehicle's location information, which is characterized by the bounding box, the vehicle type information has also been integrated into the semantic information.
- Different from the existing VoI based semantic sampling strategies, we propose that the semantic sampling is performed by a reinforcement learning (RL) based agent. This agent dynamically refines the semantic sampling strategy in response to temporal changes of the source data, channel conditions, and performance in terms of semantic accuracy and energy consumption.
- To present the illusion of real-time updates, we integrate a predictive frame interpolation module at the destination, enabling a low-complexity real-time prediction based on historical information and the sampling intervals, when the updated data from the source is unavailable.

The specific contributions of this work are outlined as follows.

[2]The preliminary work has been accepted by IEEE WCNC 2024.

- We develop a modular semantic encoder featuring the cross modality of semantic extraction and agent-assisted semantic sampling. To guide agent learning, the Value of Information (VoI)—determined by both energy consumption and scene reconstruction semantic accuracy—is treated as the reward. Recognizing the non-linear relationship between prediction deviation and sampling interval, and aiming to mitigate the impact of the convergence process and the communication overhead caused by feedback rewards, we propose an offline *stochastic* policy named Knowledge-Integrated Soft Actor Critic (K-SAC).
- We design a semantic decoder with both predictive and generative capabilities. To address the challenges in existing layout diffusion models, we first transform the received textual semantic information into visual layout. Additionally, to reverse semantic encoding, we develop a diffusion-based generative semantic inference module controlled by both the visual layout and the local static scene. This design ensures the fidelity of semantic information, while closely resembling a real remote scene. Moreover, to facilitate semantic sampling with variable time intervals, we design a customized predictive frame interpolation module.
- We conduct training and testing for the predictive frame interpolation and semantic inference modules based on data from UA-DETRAC [8]. Once trained, these models function as components of the environment in the RL paradigm, connected to the source through the semantic sampling agent. The offline policy is trained using virtual experiences generated from a realistic $\mathcal{F}$ composite channel fading model [9]. The adaptability of the well-trained policy is demonstrated across multiple scenarios. Finally, we showcase the promising gains in energy savings and reconstruction accuracy achieved by the A-GSC framework and analyze the complexity of training and deployment the framework.

Additionally, it should be noted that, while acknowledging the challenges of pixel-perfect replication in generative SemCom, such difficulties do not diminish its broad applicability. In many scenarios, the primary concern is the preservation of essential semantic information. For example, in parking space surveillance, the system remains indifferent to particulars like the color of the vehicle and the dynamic background. Instead, it exclusively focuses on the vehicle location.

The rest of this paper is organized as follows. In Section II, we review the state of cutting-edge research in SemCom from three main branches. Then, in Section III, we briefly introduce the system model in the considered scenario and outline the overview of the proposed A-GSC framework. The key components in the proposed A-GSC framework are presented in Section IV and Section V. An experimental evaluation is presented in Section VI, followed by the conclusions of the study and future works in Section VII. The main symbols have been summarized in Table I.

II. RELATED WORKS

In this section, we first introduce the development of two mainstream research directions in SemCom, namely

TABLE I
SUMMARY OF MAIN NOTATIONS

System Design-Related Symbols	
t	Index of sensing time interval
$\mathbf{s}_t$	Semantic information for the monitored scene at STI t
$\tilde{\mathbf{s}}_t$	Visualized semantic information for STI t received
$\tilde{\mathbf{s}}_t'$	Predicted visualized semantic information for STI t
L_t	Data size for the semantic information at STI t
$\tilde{g}$ ($\bar{g}$)	Instantaneous (average) channel gain
$\mathbf{r}$	Static scene information of the remote scene
χ_t	Semantic change degree for the monitored scene at STI t
P	Number of predicted future semantic messages per round
D_t	Prediction deviation D_t between $\tilde{\mathbf{s}}_t$ and $\tilde{\mathbf{s}}_t'$
$\hat{t}$	Timestamp corresponding to the last sampled scence
$\hat{D}_t$	Penalized prediction deviation D_t between $\tilde{\mathbf{s}}_t$ and $\tilde{\mathbf{s}}_t'$
Δ_t	Time interval between STI t and the last semantic sample
E_t	Energy consumption of the transmission for $\mathbf{s}_t$ at STI t
Diffusion Model-Related Parameters	
n	Index of denoising step
$\mathbf{x}_n$	Noisy data at denoising step n
ϵ_n	Noise sampled from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$
β_n	Scale of added noise at each step n
$\bar{\alpha}_n$	$\bar{\alpha}_n := \prod_{s=1}^{n}(1 - \beta_s)$
θ	Parameters of U-Net model for noise estimation
K-SAC-Related Parameters	
$\mathbf{S}_t$	$\mathbf{S}_t = [L_t, \chi_t, \bar{g}_t]$, the state observed by the agent at STI t
W	Duration of observation window in units of STIs
a_t	$a_t \in \mathcal{A} = \{0, 1\}$, action performed by the agent at STI t
r_t^1	Reward for performing semantic sampling
r_t^0	Reward for not performing semantic sampling
r_t	$r_t = r_t^1 \mathbf{1}_{a_t=1} + r_t^0 (1 - \mathbf{1}_{a_t=1})$, reward obtained at STI t
φ	Parameters for critic networks
ϕ	Parameters for actor network
ϑ	Temperature parameter
$\bar{\mathcal{H}}$	Target entropy

semantic-extraction based and semantic-sampling based Sem-Com. We then present the vision of a new SemCom paradigm enabled by GAI, demonstrating how GAI techniques can effectively combine semantic extraction and sampling. This sets the foundation for the A-GSC framework proposed in this work.

A. Semantic Extraction Based SemCom

In the research of SemCom that focuses on semantic extraction, DL stands out as the predominant technology, effectively applied to various types of modal data [4]. Beyond the widely proven efficient semantic compression capability, recent advancements have expanded its applicability to more intricate and practical scenarios. For image transmission, the authors of [10] and [23] introduce an adaptive deep coding framework within the context of variable code length enabled DL-based joint source-channel coding (DeepJSCC). This framework addresses specified target transmission quality requirements by incorporating an Oracle Network-based PSNR (peak signal-to-noise ratio) module. Furthermore, the authors in [11] leverage a learned nonlinear transform function to establish a temporally adaptive entropy model for customizing DeepJSCC for video. Additionally, addressing specific cases, the authors of [12] and [13] propose enhanced DeepJSCC frameworks for SemCom systems with task-unaware transmitters and those with multiple access, respectively.

Despite these advancements, all DeepJSCC frameworks are trained in an end-to-end manner. The non-explainability hinders the semantic information to be interpreted or manipulated as intended, e.g., the semantic sampling as discussed in Section II-B. Moreover, the most crucial point is that DeepJSCC only shows significant performance advantages in analog transmissions [24], and its effectiveness heavily depends on the match between the actual and training channel environment. These limitations restrict its effectiveness to proximity single-hop links without relay forwarding, making it difficult to apply in the remote monitoring scenarios that may involve the core network and multi-hop forwarding. To tackle these challenges, some researchers employ knowledge graphs (KG) as semantic information containers, performing semantic encoding and decoding based on KG generation, KG embedding, and KG embedding reasoning [25], [26]. However, the widespread adoption and advancement of this approach are hindered by its high computational complexity.

B. Semantic Sampling Based SemCom

Replacing the metric of delay with Age of Information (AoI) in network optimization signifies a pivotal initial step in the transition from conventional communication to semantic communication (SemCom). This transition is crucial for delineating the semantic significance of diverse data in the temporal domain [27]. Addressing the stochastic nature of environments, various AoI-based sampling and scheduling algorithms have been explored across different scenarios. These include optimization for time-average age-based sampling [5], peak-age-based sensing [28], and age-threshold-based access [14]. Despite these efforts, AoI alone may not sufficiently address task-specific requirements [29].

The aforementioned algorithms, being data-content-agnostic, run the risk of redundantly transmitting irrelevant data in certain scenarios [7]. In response to this challenge, some studies have integrated data content and task requirements into a novel metric for AoI, termed Value of Information (VoI). VoI is conceptualized as a non-linear function of AoI, taking different forms for distinct communication systems. For systems ensuring real-time updates at the receiver regarding the source's status, VoI corresponds to the age of incorrect information [15]. In a pull-based system reliant on query-driven communication, VoI is specified as the age of information at the query [16]. Additionally, for systems incorporating wireless power transfer, where the receiver must act upon the received status, VoI is adapted to represent the age of actuation [17].

Nonetheless, existing research has overlooked the semantic information embedded in source data. Particularly in monitoring scenarios, visual data often contains redundant information. Consequently, there is a compelling need for an integrated approach that combines semantic extraction and sampling.

C. Prompt Extraction Based SemCom

Recently, the emergence of GAI presents a promising opportunity to revolutionize the SemCom framework. The focus has

TABLE II
COMPARISON OF MAINSTREAM SEMCOM AND PROPOSED SEMCOM FRAMEWORKS

DL-based SemCom	AoI/VoI guided SemCom	GAI-based SemCom
Achieved Capabilities: • Semantic extraction & Inference *Advantages:* • Enhancement on the robustness for short range wireless transmission *Limitations:* • Non-explainability • Error floor in DL • Short-range communication without relays *Representative Refs:* [4], [10]–[13]	*Achieved Capabilities:* • Semantic Sampling *Advantages:* • Low semantic extraction complexity and latency *Limitations:* • Consideration of semantic importance solely from the temporal domain • Lack of semantic recovery process • Potential semantic information loss *Representative Refs:* [7], [14]–[17]	*Achieved Capabilities* • Semantic Extraction & Generation *Advantages:* • High quality of generated data • Low compression rate via cross-modality • Decoupled paradigm for flexible deployment *Limitations:* • No consideration of redundancy in time domain • Lack of awareness of decision-making *Representative Refs:* [18]–[22]
Proposed Agent-Driven Generative SemCom		
Achieved Capabilities: • Joint Semantic Extraction and Sampling & Inference and Interpolation		

Advantages: • Efficient semantic compression applied to both each individual data unit and the temporal correlation
      • Adaptive semantic extraction and sampling in response to changes in the channel gain and source content
      • Seamless integration with existing mature digital transmission for achieving remote communication		

Limitations: • Only applicable to scenarios where the transmitted content is regular and predictable • Periodic fine-tuning is required to adapt to changes in the communication context		

shifted from enhancing the efficiency of semantic encoding to the extraction of precise prompts. In this paradigm, the shared knowledge base (KB) of both communication participants can be more effectively leveraged [30]. Unlike the DeepJSCC paradigm, where KB only contributes to the training process of the semantic encoder and decoder, the generative SemCom paradigm allows background information to play a crucial role in semantic inference during each communication instance. This capability helps reduce the volume of information that needs to be transmitted.

Pioneering efforts in Generative SemCom have explored applications in language-oriented communication [18], image transmission [19], [20], [21], and audio transmission [22]. These endeavors have demonstrated high-quality reconstructed data and the fidelity of semantic information. However, the full potential of GAI for SemCom remains largely unexplored. The explainability allows the independent design and fine-tuning of transceivers. This, in turn, facilitates the seamless integration of new functional modules into the modular system. Moreover, it removes the need for channel modeling during model training. It can further leverage existing source coding technologies and integrate with current digital transmission networks to address complex and diverse network environments.

III. SYSTEM MODEL AND OVERVIEW

Without loss of generality, our focus centers on a common remote surveillance scenario in intelligent transportation systems.[3] The source in the considered scenario is an embedded vision sensor characterized by limited memory and computational capabilities, which is responsible for monitoring a certain scene, and updating the destination on the scene changes in a timely manner. The destination is a remote server with *predictable characteristics*, which is tasked with reconstructing the remote real-time scene, with the objective of presenting users with the illusion of real-time transmission. We assume that SemCom technology is deployed within the

[3]An example of OneMotoring by Land Transport Authority can be found at https://onemotoring.lta.gov.sg/content/onemotoring/home/driving/traffic_information/traffic-smart.html

system. The system model is specifically described from the following three aspects.

A. Shared Knowledge Base Model for SemCom

The shared KB is introduced as a pivotal component in SemCom systems, providing prior side information to enhance semantic encoding [30]. Unlike traditional DL-based SemCom approaches that non-selectively incorporate all KB into the empirical dataset for end-to-end training, the proposed generative SemCom framework strategically analyzes KB to unlock its full potential in boosting SemCom performance.

In the specific context considered, it can be universally agreed that the destination is solely concerned with the vehicular traffic on the road. In this sense, by utilizing shared historical surveillance data, both sources and the destination can simplify and standardize mutually agreed-upon the essential semantic representations, including vehicle location, type, and driving direction. Meanwhile, since the static background information of the monitoring scene, e.g., the view layout, as well as dynamic changes involving irrelevant pedestrians and non-motorized vehicles along the roadside, are of no interest to the destination, they can be disregarded by the source before transmission and then reconstructed at the destination based on the historically shared knowledge base, despite potential slight differences. Based on the guidance provided by the analysis of KB for specific tasks as described above, we embark on the customized design of the SemCom system.

B. Semantic Processing Model for Source and Destination

Based on the task-specific KB analysis, to compress the data for transmission, a *semantic extraction module* is utilized to extract information on vehicular traffic in the monitoring scene. The destination requires a *semantic inference module*, which reconstructs the remote monitoring scene solely from the updated traffic information. Naturally, such scene reconstruction depends on shared KB. In other words, the shared KB within this framework acts as a bridge between the source and destination, decoupling the design of the encoder and decoder. In addition, considering the communication objectives of the surveillance task, the destination is more interested in
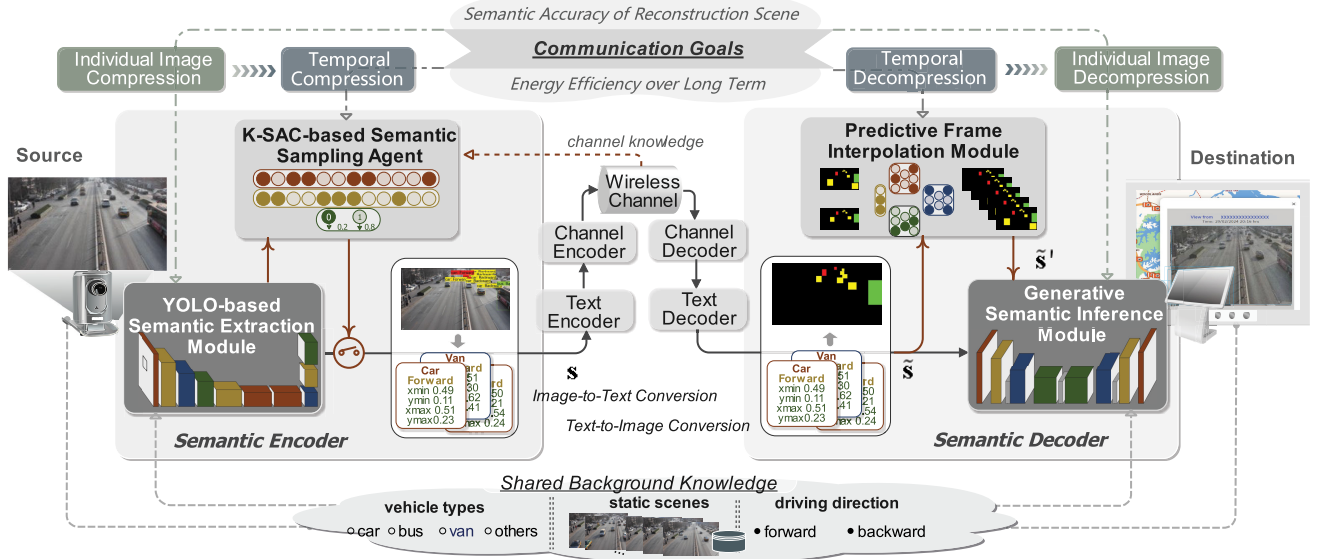
Fig. 1. An agent-driven generative semantic communication framework with cross-modality and prediction.

changes in the remote scene, specifically in traffic flow rather than in background information. In this sense, performing appropriate sampling can further reduce the amount of data to be transmitted from the temporal dimension. Thus, we incorporate a *semantic sampling agent* at the source to perform a binary 0-1 action—deciding whether to transmit the current semantic information or not, with the consideration of the semantic changes in the monitoring scene and other factors. Additionally, we integrate a *predictive frame interpolation module* at the destination, enabling the illusion of real-time transmission to users, while also potentially reducing the number of required sampling instances. The specific designs of the involved modules related to the semantic extraction and semantic sampling are detailed in Section IV and Section V.

C. Wireless Transmission Model

In this work, we believe that wireless transmission is a bottleneck in the communication process. As such, we mainly analyze and evaluate the wireless transmission performance. Considering the combined effects of multi-path and shadowing on the practical transmission, we adopt the $\mathcal{F}$ composite fading model to characterize the stochastic wireless channel [9]. We denote the instantaneous channel gain by $\tilde{g}$. The probability density function of $\tilde{g}$ is given by [9]

$$f\left(\tilde{g}\right) = \frac{m^m (m_s - 1)^{m_s} \bar{g}^{m_s} \tilde{g}^{m-1}}{B\left(m, m_s\right) \left[m\tilde{g} + (m_s - 1)\bar{g}\right]^{m+m_s}}, \qquad (1)$$

where m and m_s represent the number of clusters of multipath, shadowing shape, respectively, and $\bar{g}$ is the corresponding average channel gain, i.e., $\bar{g} = \mathbb{E}\left[\tilde{g}\right]$. Moreover, $B\left(\cdot, \cdot\right)$ denotes the beta function [9]. Since we focus on the design of the semantic encoder and decoder, without loss of generality, we assume perfect capacity achieving coding in this work. Moreover, in order to cope with stochastic fading, it is assumed that the transmitter of the embedded vision sensor adopts a power control technique. We denote the decoding threshold
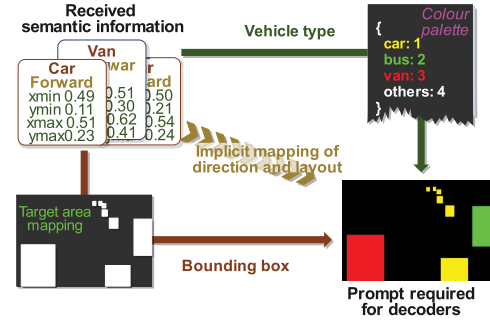
Fig. 2. Transformation from semantic information to prompt.

for the signal-to-noise ratio by Θ. Thus, the achievable transmission rate is expressed by

$$R = W \log\left(1 + \Theta\right), \qquad (2)$$

where W is the allocated bandwidth. The instantaneous transmit power is then given by $\tilde{p} = \Theta\sigma^2 / \tilde{g}$, where σ^2 represents noise power.

IV. KEY COMPONENT DESIGN: SEMANTIC EXTRACTION

In this section, we focus on semantic extraction, a key aspect of SemCom. Within the proposed framework, semantic extraction and inference are achieved through a YOLO-based semantic extraction module and generative semantic inference modules, respectively.

A. Semantic Extraction Module

Based on the strategic analysis on the shared KB as stated in Section III-A, we can see that the destination is mainly concerned with the information about the distribution of the traffic flow. To track the vehicles' information in real time, the semantic extraction can be performed based on a mature real-time object detection algorithm, such as the low-complexity YOLOv5 [31] which can be deployed in multiple embedded and mobile devices. Furthermore, we characterize

the semantic information about the position for vehicle v as the bounding box, which is denoted by a 4-tuple $\mathbf{b}_v = (b_1^v, b_2^v, b_3^v, b_4^v) \in [0,1]^4$. Therein, (b_1^v, b_2^v) represents the coordinate of the top-left corner, and (b_3^v, b_4^v) represents the coordinate of the bottom-right corner. Given that for a two-way road, the driving directions of the vehicles are tightly coupled with their locations, we do not consider the semantic information of the vehicle direction, separately. To avoid the ambiguous semantic information of the bounding box arising from the perspective effect of the position of the vehicle and the size of the vehicle type itself, we use a scalar to represent the type of vehicle, denoted by $o_v \in \mathcal{O} = \{1, \ldots, k, \ldots, K\}$, where K represents the total number of the vehicle types. In summary, a piece of semantic information for a vehicle v in the monitoring scene can be represented by a vector $\mathbf{s}_v = [o_v, \mathbf{b}_v]$, the data size of which after text data encoding is denoted by ℓ.[4] Assuming a total of M vehicles in the scene, the total semantic information contained in the scene can be represented as $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_v, \ldots, \mathbf{s}_M]$. Before transmission, the semantic information is packed into a packet with the data size of $L = M\ell$. If there are no vehicles in the monitored scene, an empty semantic message will be sent to the receiving side. Transforming the originally captured scene images into textual layout information can significantly enhance the compression ratio at the source side. For a relevant quantitative comparison with other representative SemCom paradigms, please refer to our prior research [32].

B. Semantic Inference Module

Given the requirement for photorealistic scene reconstruction, we utilize the diffusion model as the foundational element of our semantic inference module design. However, multimodal fusion of the textual prompts and image features during the diffusion process presents a significant challenge. In existing layout diffusion models, issues such as object omission can occur when the number of objects is excessively large or their size is too small [35]. This mainly arises from the absence of structured spatial information in the textual layout.

To address this, we do not directly use the received textual semantic information as the prompt for scene reconstruction. Instead, we first visualize the textual information based on knowledge base as shown in Fig. 2, thereby alleviating the burden of the cross-modal fusion in the diffusion model. Through target area mapping, the position and size of the vehicles can be easily fused with the positional information of each feature map during the diffusion process. *This not only addresses the challenge of multimodal fusion between text and images in existing layout diffusion research but also removes the limitation of the existing layout diffusion models on the maximum number of objects.* Additionally, to differentiate the specific categories of objects, we paint the areas corresponding to different vehicles with resolution values. Meanwhile, for synthesizing a seamless and realistic scene, we also use the static scene information of the remote scene as a prompt for

the diffusion model, which is denoted by $\mathbf{r}$. It should be noted that the static scene information does not need to be updated in real-time to the destination, unless there are changes in the weather or times of day. This is because the dynamics in the background of exclusion zones are not of concern.

The diffusion model itself is a latent variable model, which consists of a forward and a reverse process. We denote the noisy data at denoising step n by $\mathbf{x}_n$. The *forward* process is defined as a discrete Markov chain of length N: $q(\mathbf{x}_{1:N}|\mathbf{x}_0) = \prod_{n=1}^{N} q(\mathbf{x}_n|\mathbf{x}_{n-1})$. In each step $n \in [1, N]$ in the forward process, the diffusion model adds noise ϵ_n sampled from the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ to data $\mathbf{x}_{n-1}$ and obtains disturbed data $\mathbf{x}_n$ from $q(\mathbf{x}_n|\mathbf{x}_{n-1}) = \mathcal{N}(\mathbf{x}_n; \sqrt{1-\beta_n}\mathbf{x}_{n-1}, \beta_n\mathbf{I})$, where $\beta_n, n \in [1, N]$ characterizes the scale of added noise at each step n. Notably, during the training process, we can sample $\mathbf{x}_n$ at any step n in closed form via $q(\mathbf{x}_n|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_n; \sqrt{\bar{\alpha}_n}\mathbf{x}_0, (1-\bar{\alpha}_n)\mathbf{I})$, with $\bar{\alpha}_n := \prod_{s=1}^{n}(1-\beta_s)$. Moreover, given the prompts $\tilde{\mathbf{s}}$ and $\mathbf{r}$ feed into the diffusion model, the reverse process can also be defined as a Markov chain: $p_\theta(\mathbf{x}_{0:N}|\tilde{\mathbf{s}}, \mathbf{r}) = p(\mathbf{x}_N)\prod_{n=1}^{N} p_\theta(\mathbf{x}_{n-1}|\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r})$, where $p_\theta(\mathbf{x}_{n-1}|\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r})$ can be parameterized as $\mathcal{N}(\mathbf{x}_n; \mu_\theta(\mathbf{x}_n, n), \sigma_n)$, where $p_\theta(\mathbf{x}_{n-1}|\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r})$ is parameterized as $\mathcal{N}(\mathbf{x}_n; \mu_\theta(\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r}, n), \sigma_n)$. By modeling and comparing the posterior distribution of the forward process, the mean term can be further rearranged as $\mu_\theta(\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r}, n,) = \frac{1}{\sqrt{1-\beta_n}}\left(\mathbf{x}_n - \frac{\beta_n}{\sqrt{1-\bar{\alpha}_n}}\epsilon_\theta(\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r}, n)\right)$, and the variance term can be approximated as $\sigma_n = \beta_n\mathbf{I}$ without sacrificing performance. Furthermore, ϵ_θ is parameterized by a U-Net network (as shown in Fig. 3) of which inputs are $\mathbf{x}_n$, $\tilde{\mathbf{s}}$, $\mathbf{r}$, and n. In light of above, the simplified loss function guiding the training can be expressed as

$$\mathcal{L}_\theta = \mathbb{E}_{n, \mathbf{x}_n,}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_n}\mathbf{x}_n + \sqrt{1-\bar{\alpha}_n}\epsilon, \tilde{\mathbf{s}}, \mathbf{r}, n\right]\right)\right\|_2\right]. \tag{3}$$

Then, during the inference process, samples are generated by iterating through the reverse process $p_\theta(\mathbf{x}_{0:N}|\tilde{\mathbf{s}}, \mathbf{r})$ from $n = N$ to $n = 0$. Specifically, $\mathbf{x}_N \sim \mathcal{N}(0, \mathbf{I})$, and $\mathbf{x}_n$ in each step is predicted as

$$\mathbf{x}_{n-1} = \frac{1}{\sqrt{1-\beta_n}}\left(\mathbf{x}_n - \frac{\beta_n}{\sqrt{1-\bar{\alpha}_n}}\epsilon_\theta(\mathbf{x}_n, \tilde{\mathbf{s}}, \mathbf{r}, n)\right) + \sigma_n\mathbf{z}, \tag{4}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.[5]

V. KEY COMPONENT DESIGN: SEMANTIC SAMPLING

This section shifts focus to semantic sampling, aiming to streamline transmission in the *temporal* domain, particularly like *real-time* operations such as surveillance. To achieve this, we place a RL-based *semantic sampling agent* at the source side to solve the sequential decision problem. To ensure that the sampling behavior is unnoticeable to the destination, a *predictive frame interpolation module* must be pre-installed on the destination side. When the agent and the module work

---

[4]We have total vehicle types $K = 4$, and use two bits to characterize o_v, and use five bits to characterize $x_{\min}^v$, $y_{\min}^v$, $x_{\max}^v$, and $y_{\max}^v$, respectively, with a resolution of 0.03125. Thus, we have $\ell = (2 + 4 \times 5)$ bit.

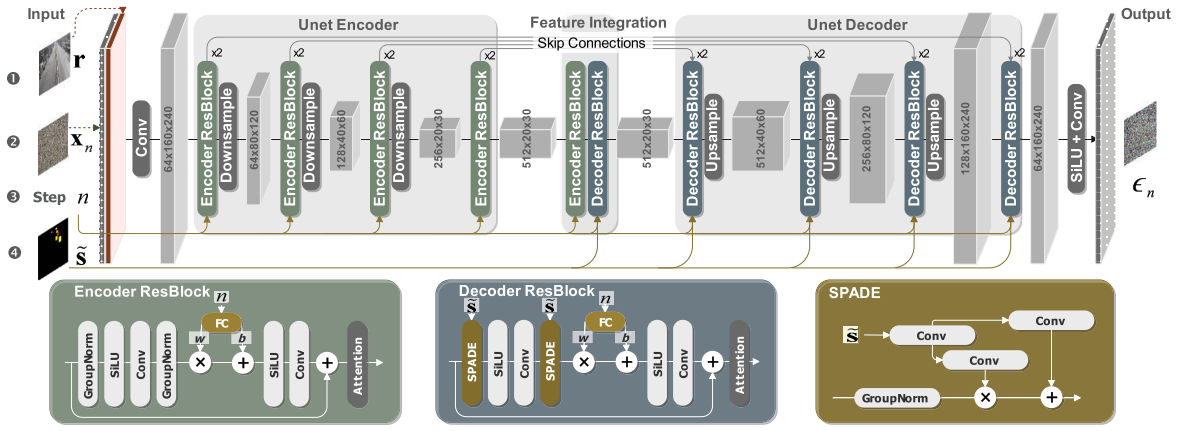[5]The detailed derivation and proof can be found in [36].

Fig. 3.   U-Net neural network for noise prediction [33], [34].

in harmony, the SemCom system's performance evaluation will align with the underlying SemCom framework given in Section IV. We next begin with the design of the semantic sampling agent, followed by a detailed illustration of the involved predictive frame interpolation module.

A. K-SAC Semantic Sampling Agent

The semantic sampling agent is responsible for adaptively performing semantic sampling based on varying channel conditions and source data content, with the objective of minimizing the amount of transmitted data while ensuring optimal performance in SemCom. Although the RL paradigm can address the challenge of the intricate correlation between the *state* and the *reward*, an issue remains. Prior simulations reveal that prediction deviation and sampling interval are not strictly linearly related due to the influence of object size and object moving speed. To mitigate the ambiguity that the nonlinearity creates for the agent's learning, we adopt a *stochastic* learning policy named soft actor-critic (SAC), making it robust to such occasional nonlinearities. Additionally, to avoid the impact of convergence latency on communication performance and the communication overhead resulting from immediate reward feedback from the destination to the source, we integrate knowledge into the SAC algorithm [37]. By modeling wireless transmission using the acknowledged empirical channel model, $\mathcal{F}$ composite fading model, virtual experiences can be generated to facilitate offline training. Thus, we refer to our algorithm as K-SAC, and its MDP model is detailed as below.

1) MDP Model: Specifically, the time step in RL framework is specified as a sensing time interval (STI). The duration of STI is determined by the supportable frame rates for vision sensors and the inference time of the diffusion-based generation module, and each STI is indexed by t. As shown in Fig. 4, the entire SemCom process can be viewed as the environment. The semantic sampling agent corresponds to the agent in an MDP, which interacts with the environment by performing sampling actions. The goal of the agent is to achieve the finest cumulative reward $\hat{r}$ within the observation window (denoted by W) with a duration of T. Denote the immediate reward by r_t. We define $\hat{r} = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \cdot r_t\right]$, where γ is the discount factor. Since we aim to achieve consistently high scene reconstruction accuracy, γ is set to 1.

The interactions between the agent and the knowledge-based virtual environment are presented as below.

a) State: The state observed by the agent at time step t is denoted as $\mathbf{S}_t$, encompassing four critical factors: specifically, $\mathbf{S}_t = [L_t, \boldsymbol{\chi}_t, \bar{g}_t]$. Herein, L_t characterizes the current data size of the semantic information packet as defined in Section IV-A. Furthermore, $\boldsymbol{\chi}_t = [\chi_t, \chi_{t-1}, \ldots, \chi_{t-W}]$ conveys historical insights into the semantic change degree in the past W STIs, facilitating the agent to capture the urgency associated with the sampling process. In the considered task, We denote the timestamp corresponding to the last sampled scene by $\hat{t}$. Then, the semantic change degree is measured based on the cumulative deviation of bounding boxes for each vehicle in the current scene compared to the last sampled scene, while irrelevant information, such as time of day and weather, can be ignored. Specifically, χ_t can be calculated according to semantic information $\mathbf{s}_t$ and $\mathbf{s}_{\hat{t}}$ as

$$\chi_t = \sum_{v=1}^{M} \frac{A_v^t + A_v^{\hat{t}} - 2I_v}{2(A_v^t + A_v^{\hat{t}} - I_v)}, \tag{5}$$

where $I_v = |\min\{b_3^{v,t}, b_3^{v,\hat{t}}\} - \max\{b_1^{v,t}, b_1^{v,\hat{t}}\}| \cdot |\min\{b_4^{v,t}, b_4^{v,\hat{t}}\} - \max\{|b_2^{v,t}, b_2^{v,\hat{t}}\}|$ and $A_v^t = |b_2^{v,t} - b_2^{v,t}| \cdot |b_4^{v,t} - b_4^{v,t}|$. At last, $\bar{g}_t$ denotes the average channel gain in time step t, which can provide a basis for the agent to assess the energy consumption attributed to the sampling operation. The agent perceives the state of environment and uses the information to make the decision as presented as below. Afterward, state $\mathbf{S}_t$ transitions to the next state, $\mathbf{S}_{t+1} = [L_{t+1}, \boldsymbol{\chi}_{t+1}, \bar{g}_{t+1}]$, based on the executed action. Specifically, the agent's actions primarily affect the state element χ_{t+1} in $\boldsymbol{\chi}_{t+1}$. Both L_{t+1} and $\bar{g}_{t+1}$ are equivalent to the parameters of the MDP and are directly determined by the size of the semantic information of the source data and the channel conditions at STI $t + 1$, independent of the action outcome. It is essential to emphasize that, in this work, we assume that the source is stationary. In cases involving moving sensors, the state should encompass the historical channel variations to capture the influence of user mobility on transmission energy consumption.

b) Action: The action performed by the semantic sampling agent at the beginning of time step t can be expressed by $a_t \in \mathcal{A} = \{0, 1\}$. If $a_t = 1$, semantic sampling takes place

---

Algorithm 1 K-SAC-Based Agent-Driven Semantic Sampling Algorithm

---

1 Initialize reply memory $\mathcal{D}$; Initialize weights of actor network and critic networks
2 **for** *episode j = 1 to MAX_EPISODE* **do**
3   Create a buffer $\mathbf{f}_j$ for recording the visual layout sequence at the destination
4   Predict the next P visual layouts and create a list of variable length denoted by $\tilde{\mathbf{s}}' = [\tilde{\mathbf{s}}'_1, \tilde{\mathbf{s}}'_2, \dots, \tilde{\mathbf{s}}'_P]$
5   Calculate semantic change degree χ_t based on $\mathbf{s}_t$ and $\mathbf{s}_{\hat{t}}$
6   *Randomly* initialize the initial state $\mathbf{S}_1$
7   **for** $t = 1$ *to T (MAX_STEP)* **do**
8   **if** $\tilde{\mathbf{s}}'$ *is non-empty* **then**
9   Pop up the predicted visual layout $\tilde{\mathbf{s}}'_t$ from the head of the list $\tilde{\mathbf{s}}'$
10   **else**
11   Perform next loop visual layout prediction based on previously popped up visual layouts $\tilde{\mathbf{s}}'_{t-1}$ and $\tilde{\mathbf{s}}'_{t-2}$
12   **end**
13   Generate an action a_t based on the probability distribution output by the current policy π_t
14   **if** *Perform semantic sampling ($a_t = 1$)* **then**
15   Transform semantic information $\mathbf{s}_t$ into visual layout $\tilde{\mathbf{s}}_t$ and store the real visual layout $\tilde{\mathbf{s}}_t$ in to $\mathbf{f}_j$
16   Perform the prediction for the next P visual layouts based on $\tilde{\mathbf{s}}_t$ and last sampled $\tilde{\mathbf{s}}_{\hat{t}}$ and update list $\tilde{\mathbf{s}}'$
17   Estimate the energy consumption E_t according to (8)
18   **else**
19   Store the predicted visual layout $\tilde{\mathbf{s}}'_t$ in to $\mathbf{f}_j$
20   Compare the deviation D_t between $\tilde{\mathbf{s}}_t$ and $\tilde{\mathbf{s}}'_t$ according to (11)
21   **end**
22   Calculate reward r_t according to (6) and $\mathbf{S}_{t+1} \leftarrow \mathbf{S}_t$; Record transition $e_t = [\mathbf{S}_t, a_t, \mathbf{S}_{t+1}, r]$ in $\mathcal{D}$
23   Sample a batch of $\mathcal{B}$ to perform a gradient descent according to (16), (17), and (18)
24   Update the parameters following: $\theta \leftarrow \theta - \lambda_\theta \nabla L_Q(\theta)$; $\phi \leftarrow \phi - \lambda_\phi \nabla L_\pi(\phi)$; $\vartheta \leftarrow \vartheta - \lambda_\vartheta \nabla L(\vartheta)$
25   **end**
26 **end**

---

within STI t. Meanwhile, the timestamp corresponding to the last sampled scene is updated as $\hat{t} \leftarrow t$, which intervenes in the state transitions as stated above. Otherwise, semantic sampling is not carried out.

c) Reward: The reward acquired by the agent at the end of STI t is denoted by r_t. In the pursuit of highly efficient remote surveillance, the immediate reward is considered with regard to two key factors. The first factor pertains to the precision of the reconstructed scene measured inversely through the prediction deviation D_t between the real-time visual layout $\tilde{\mathbf{s}}_t$ and the predicted visual layout $\tilde{\mathbf{s}}'_t$. Note that the prediction method is detailed in Section V-B. The second factor relates to the energy consumption E_t incurred during the transmission of the sampled semantic map. With both factors in mind, two sub-reward functions are designed: one for scenarios where semantic sampling is performed, denoted by r_t^1, and the other for scenarios without semantic sampling, denoted by r_t^0. Then, the immediate reward function can be expressed as follows:

$$r_t = r_t^1 \mathbf{1}_{a_t=1} + r_t^0 (1 - \mathbf{1}_{a_t=1}). \tag{6}$$

It is worth noting that, in the conventional RL paradigm, the reward is obtained through the agent's interaction with the actual environment. However, in K-SAC, to reduce communication overhead, we model the interaction between the agent and the environment based on existing channel modeling and mathematical analysis, generating virtual experiences $(\mathbf{S}_t, a_t, r_t, \mathbf{S}_{t+1})$ that can be used for agent training. The specific forms and design of r_t^1 and r_t^0 are as follows:

Firstly, to minimize unnecessary sampling, we define the sub-reward r_t^1 for the action of sampling as a function inversely proportional to energy consumption. Considering that the energy consumption per transmission typically ranges
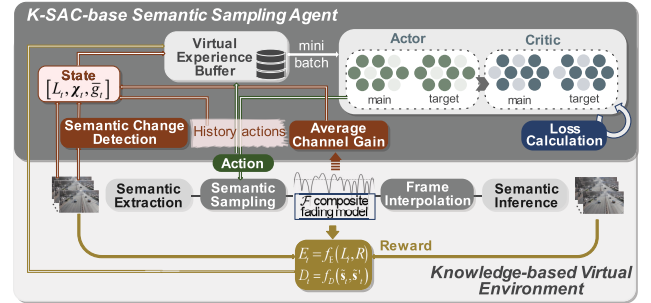

Fig. 4. K-SAC based semantic sampling agent design.

around 0.015 mJ, we introduce a weight parameter w_1 to adjust its influence on the sub-reward by one order of magnitude. This ensures that the scaled values are primarily distributed between 0 and 1. Afterwards, given the values of E_t with significantly less variance compared to D_t, we employ a logarithmic function to further accentuate the variations between different levels of energy consumption. Meanwhile, this logarithmic transformation helps to prevent potential excessive energy consumption caused by random fading, thus mitigating its impact on the stability of agent learning. Additionally, to maintain a balance between the energy consumption and prediction deviation terms in the final reward function, we incorporate another weighting coefficient, w_2, for the logarithmic component. The specific expression for the sub-reward r_t^1 is presented as follows,

$$r_t^1 = w_2 \log(1 + w_1 E_t), \tag{7}$$

where we choose $w_1 = 10$ and $w_2 = -6$. Moreover, E_t is a function of state elements L_t and $\bar{g}_t$. Instead of interacting with the real environment, we derive the value of E_t based

on the well-accepted $\mathcal{F}$ composite fading model as stated in Section III-C. Given that the transmitter adopts the power control technique, the transmission duration for semantic map $\mathbf{m}_t$ can be expressed by $\delta_t = L_t/R$. Therefore, we have

$$E_t = \int_0^{\delta_t} \frac{\Theta \sigma^2}{\tilde{g}_k} dk,$$

where k is the index of the transmission time interval. Without loss of generality, we assume that the stochastic fading can be treated as independently and identically distributed (i.i.d.) among transmission time intervals. Then, E_t can be approximated as

$$
\begin{aligned}
E_t &= \delta_t \mathbb{E}_{\tilde{g}}\left[\tilde{p}\right] = \int_0^\infty \frac{\Theta \sigma^2}{\tilde{g}} f\left(\tilde{g}\right) d\tilde{g} \\
&= \delta_t \Theta \sigma^2 \int_0^\infty \tilde{g}^{-1} f\left(\tilde{g}\right) d\tilde{g} \\
&= \delta_t \Theta \sigma^2 \mathbb{E}\left[\tilde{g}^{-1}\right].
\end{aligned}
\tag{8}
$$

According to (1), with the aid of [38, Eq. (3.194.3)], the nth moment of $\tilde{g}$ can be derived as

$$\mathbb{E}\left[\tilde{g}^n\right] = \frac{(m_s - 1)^n \bar{g}^n \Gamma\left(m + n\right) \Gamma\left(m_s - n\right)}{m^n \Gamma\left(m\right) \Gamma\left(m_s\right)}, \tag{9}$$

where $\Gamma\left(\cdot\right)$ represents the gamma function. Substituting the case of $n = -1$ in (9) into (8), we can obtain the final expression of E_t as

$$E_t = \delta_t \frac{\Theta \sigma^2 m \Gamma\left(m - 1\right) \Gamma\left(m_s + 1\right)}{(m_s - 1) \bar{g} \Gamma\left(m\right) \Gamma\left(m_s\right)}. \tag{10}$$

Secondly, in designing r_t^0, to ensure high semantic accuracy for scene reconstruction while promoting the energy-saving awareness of the agent, we introduce a positive constant term, w_3, for non-sampling behavior. The value of w_3 is determined based on the average energy consumption and the system's emphasis on E_t and D_t. Furthermore, D_t is calculated similarly to (5). We denote the total number of the pixels occupied by each type of vehicles in $\tilde{\mathbf{s}}_t$ and $\tilde{\mathbf{s}}'_t$ by $n_{v,t}$ and $n'_{v,t}$, respectively. Meanwhile, the intersection between them is denoted by $\hat{n}_{v,t}$. Then, D_t can be expressed by

$$D_t = \sum_{v=1}^M \frac{n_{v,t} + n'_{v,t} - 2\hat{n}_{v,t}}{2\left(n_{v,t} + n'_{v,t}\right)}. \tag{11}$$

Given the fact that larger deviations result in increasingly severe consequences, we incorporate an exponential function into this sub-reward formulation. Furthermore, to proactively address the issue of significant cumulative prediction deviation resulting from prolonged reliance on destination-side predictions, we define $\hat{D}_t$ using a penalty function as below,

$$\hat{D}_t = \begin{cases} D_t & D_t \leqslant D_{\text{th}} \\ \min\left\{D_t + \kappa, 1\right\} & D_t > D_{\text{th}} \end{cases}. \tag{12}$$

This is relative to the actual deviation D_t, considering that the tolerable prediction deviation is capped by a threshold, D_{th}. Moreover, we introduce a weight coefficient w_4 for $\hat{D}_t$, the value of which is chosen in conjunction with the value of w_3. Thus, the expression of r_t^0 is presented as

$$r_t^0 = w_3 - \exp\left(w_4 \hat{D}_t - 1\right), \tag{13}$$

where we set $w_3 = 1$, $w_4 = 2$, $D_{\text{th}} = 0.07$, and $\kappa = 0.5$.

The careful selection of aforementioned weight coefficients collectively guarantees the rationality of rewards and punishments. This mitigates the agent's inclination towards short-term gains due to insufficient sampling. Simultaneously, it also prevents hesitation in performing a necessary sampling given the potential penalties associated with sampling.

d) Policy: The policy (denoted by π) describes the mapping from a given state to the agent's action selection, which can achieve the maximum long-term rewards. At the beginning of Section V-A, we choose a stochastic policy, named K-SAC, where the agent selects an action based on the action distribution generated by the policy's output, $\pi\left(\mathbf{S}_t\right)$. It is important to note again that K-SAC is an enhancement of the SAC algorithm, specifically designed for an offline RL paradigm. By leveraging existing channel modeling research and mathematical analysis to model the wireless transmission process as discussed above, we can generate virtual experiences based on practical history footage clips, such as the relationships r_t and $\mathbf{S_{t+1}}$, and $\mathbf{S_t}$ and a_t. These virtual experiences are used for offline pre-training and fine-tuning, followed by periodically updating to the sender's semantic sampling module. This not only reduces the communication overhead during training but also avoids impacting actual communication performance during the convergence process. The learning principle is detailed in Section V-A.2.

2) Learning Principle: Based on the MDP designed above, we present the learning part of SAC. For illustration, we denote the state-action trajectory under the policy π by ρ_π. In the traditional RL algorithm, the fundamental objective of agent is to learn the optimal policy π which maximizes the cumulative expected rewards, which is expressed by

$$\pi^* = \arg\max_\pi \sum_t \mathbb{E}_{(\mathbf{S}_t, a_t) \sim \rho_\pi}\left[r\left(\mathbf{S}_t, a_t\right)\right], \tag{14}$$

where π^* represents the optimal policy. Differently, in SAC, to facilitate the agent to achieve the optimal goals, SAC dynamically optimizes the expected entropy of the policy over ρ_π during the learning process to flexibly handle the exploitation and exploration dilemma. Thus, the aim of agent can be reformulated as

$$J\left(\pi\right) = \sum_{t=0}^T \mathbb{E}_{(\mathbf{S}_t, a_t) \sim \rho_\pi}\left[r\left(\mathbf{S}_t, a_t\right) + \vartheta \mathcal{H}\left(\pi\left(\cdot | \mathbf{S}_t\right)\right)\right], \tag{15}$$

where ϑ is a temperature parameter which determines the relative importance of entropy relative to the long-term reward, and $\mathcal{H}\left(\pi\left(\cdot | \mathbf{S}_t\right)\right)$ is the entropy, i.e., $\mathcal{H}\left(\pi\left(\cdot | \mathbf{S}_t\right)\right) = -\log_\pi\left(\cdot | \mathbf{S}_t\right)$. In addition, SAC follows the classic actor-critic architecture and experience replay mechanism [39], [40]. The policy $\pi_\phi\left(a_t | \mathbf{S}_t\right)$ is generated by the actor network with parameter ϕ. Meanwhile, to enhance the stability of learning, twin critic networks with the same initial parameter θ are maintained to avoid overestimation of the soft Q-function $Q_\theta\left(\mathbf{S}_t, a_t\right)$, which is the cumulative expected rewards after taking the action a_t at the state $\mathbf{S}_t$ under the policy π.
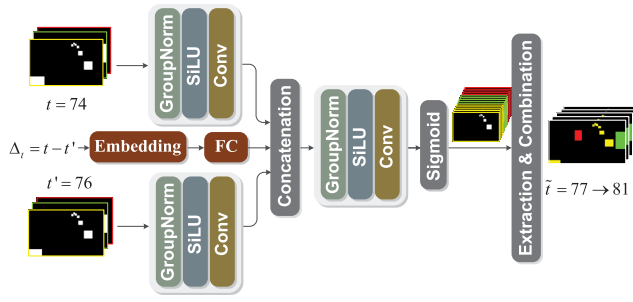
Fig. 5. Neural network for predictive frame interpolation.

Fig. 6. Prediction mechanisms at the destination.

According to the Bellman equation of the soft Q function, the loss function for the critic network can be expressed by

$$\mathcal{L}_Q(\varphi) = \mathbb{E}_{(\mathbf{S}_t, a_t) \sim \mathcal{B}} \left[\frac{1}{2} \left(Q_\varphi(\mathbf{S}_t, a_t) - (r(\mathbf{S}_t, a_t) \right. \right.$$
$$\left. \left. + \gamma Q_{\bar{\varphi}}(\mathbf{S}_{t+1}, a_{t+1}) - \log \pi_\phi(a_{t+1} | \mathbf{S}_{t+1}) \right) \right)^2 \right],$$
$$(16)$$

where $\mathcal{B}$ represents the replay memory for improving sampling efficiency, and $\bar{\varphi}$ is the parameter of the target network with soft update, i.e., $\bar{\varphi} \leftarrow \tau\varphi + (1-\tau)\bar{\varphi}$ with $\tau \in (0,1)$. Moreover, to maximize the Q value, the loss function for the training of the actor network can be expressed by

$$L_\pi(\phi) = \mathbb{E}_{\mathbf{S}_t \sim \mathcal{B}} \left[\mathbb{E}_{a_t \sim \pi_\phi} \left[\vartheta \log(\pi_\phi(a_t | \mathbf{S}_t)) - Q_\theta(\mathbf{S}_t, a_t) \right] \right].$$
$$(17)$$

In addition, the adjustment of the temperature parameter can be realized according to the following loss function

$$L(\vartheta) = \mathbb{E}_{a_t \sim \pi_t} \left[-\vartheta_t \log \pi_t(a_t | \mathbf{S}_t) - \vartheta_t \bar{\mathcal{H}} \right], \qquad (18)$$

where $\bar{\mathcal{H}} = -|\mathcal{A}|_{\dim}$.

We should again highlight that the training and implementation of K-SAC algorithm is separated. Firstly, an off-line training method is adopted based on virtual experiences for safeguarding and reducing communication overhead [40], [41]. Then, the well-trained model is implemented in real-world scenarios. The optimality, robustness, and learning efficiency are discussed in Section VI. Since the predictive frame interpolation at the destination also serves as part of the environment, the entire agent-driven semantic sampling algorithm is outlined in Algorithm 1 after the following elaboration of predictive frame interpolation module.

B. Predictive Frame Interpolation Module

The predictive frame interpolation module at the destination plays a crucial role in generating the illusion of real-time updates. When the source does not perform semantic sampling or update semantic information s to the destination, the predictive frame interpolation module generates predicted visual layout information $\tilde{\mathbf{s}}'$ and inputs it into the semantic inference module for timely scene reconstruction. Essentially, this module performs a task similar to video frame prediction but in a much simpler manner.

Specifically, despite significant advancements in DL, the task of video frame prediction continues to present challenges, such as image deblurring, particularly in the long-term
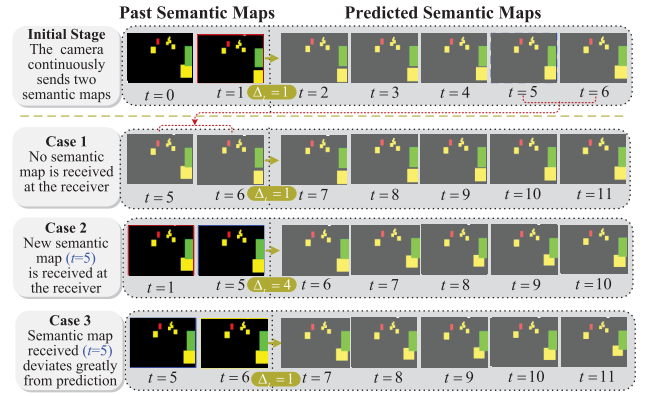
prediction [42]. Fortunately, since visual layout exclusively encompasses low-level features and eliminates irrelevant features in the background, the issues caused by the loss of details can be significantly mitigated. Given that low-level features are widely recognized as being readily captured by shallow neural networks, we design a lightweight prediction network architecture, as illustrated in Fig. 5. To avoid unnecessary difficulties caused by different vehicle types during the prediction process, the two past visual layouts are first separated into multiple visual layouts, each containing only one type of vehicle. These layouts are then fed into two ResBlocks with shared parameters to extract features using the batching technique. Simultaneously, given that the intervals for semantic sampling exhibit variability, we also take the embedding of the temporal discrepancy Δ_t between the timestamps of two visual layouts as the inputs, which is then amalgamated with both derived feature maps corresponding to the two past visual layouts. The fused feature is subsequently processed by another ResBlock to predict the visual layout for each type of vehicles. Finally, through a extraction and combination module, the visual layout for a scene can be output and used as the prompts to generate the remote scene, if required. To reduce the computational overhead associated with frequent predictions, we set that the module predicts next P visual layouts at once. In this sense, a larger value of P means lower system complexity. However, increasing P also raises the computational power requirement at the destination and amplifies the impact of vehicle speed variations on prediction accuracy, posing greater challenges to maintaining prediction accuracy. Therefore, the selection of P requires a trade-off between these factors.

To align with adaptive semantic sampling at the source, the visual layout prediction mechanisms at the destination for different cases are as follows.

Case 1: At the initial STI, the source continuously samples and transmits the semantic information $\mathbf{s}_0$ and $\mathbf{s}_1$ corresponding to the scenes within two consecutive STIs. Subsequently, the semantic inference module first transforms semantic information into the visual layout $\tilde{\mathbf{s}}_{00}$ and $\tilde{\mathbf{s}}_{01}$. Then, the predictive frame interpolation module predicts the next P visual layouts $\tilde{\mathbf{s}}'_1 \ldots \tilde{\mathbf{s}}'_P$ based on $\tilde{\mathbf{s}}_{00}$ and $\tilde{\mathbf{s}}_{01}$.

Case 2: In the following STI t, if no updated semantic information is received and the visual layout for that moment has already been predicted, the predictive frame interpolation

TABLE III
SIMULATION PARAMETER SETTINGS

Wireless Network Parameter Settings			
Small-scale fading model		$\mathcal{F}$ composite fading mode [9]	
Pass loss (average channel gain) [dB]		$35.3 + 37.6\log_{10}(d)$ [43]	
Fading severity m	6	Bandwidth W [kHz]	1
Wireless Distance d [m]	100	SNR threshold [dB]	15
Noise power [dBm/Hz]	-90	Shadowing shape m_s	6
Hyperparameters for Generative Semantic Inference Module			
Learning rate	2e-5	Batch size	1
Denoising step N	1000	Guidance scale k	8
Hyperparameters for Predictive Frame Interpolation Module			
Learning rate	2e-5	Batch size	5
Hyperparameters for K-SAC based Semantic Sampling Agent			
Memory Capacity	1e5	Batch size	1024
Learning rate for policy network	1e-5	Soft Update	0.2
Learning rate for critic network	2e-5	Observation window	150
Learning rate for temperature parameter	1e-5		

module feeds $\tilde{\mathbf{s}}'_t$ directly to semantic inference module. Conversely, if $\tilde{\mathbf{s}}'_t$ falls outside the scope of the previous prediction, the prediction module uses $\tilde{\mathbf{s}}'_{t-1}$, $\tilde{\mathbf{s}}'_{t-2}$ and $\Delta_t = 1$ as the inputs for the next round prediction. Notably, the latter scenario indicates that the remote scene undergoes minimal changes, ensuring that successive predictions do not result in much cumulative deviation.

Case 3: In the following STI t, if new layout information $\mathbf{s}_t$ is received at the destination, the prediction module first compares the visual layout $\tilde{\mathbf{s}}_t$ transformed from $\mathbf{s}_t$ with the predicted visual layout $\tilde{\mathbf{s}}'_t$. If the prediction deviation D_t falls below a predetermined threshold, the prediction module promptly conducts the next round of prediction based on the updated $\tilde{\mathbf{s}}_t$ and the last sampled $\tilde{\mathbf{s}}_{\hat{t}}$. This approach minimizes the spread of prediction deviations. In contrast, if D_n exceeds the preset threshold, it indicates a sudden change in the remote scene. In such a case, the source is informed to continue performing the sampling at STI $(t + 1)$ and initiate a new prediction process at the destination.

It is important to note that the above mechanisms need to be integrated into the dynamics of the environment that semantic sampling agent interacts with. Furthermore, given that the latter scenario in Case 3 is an extreme instance that can serve as a starting point for a new episode, it is not treated as a distinct entity in environmental modeling. The full realization of the semantic sampling in the proposed SemCom framework has been highlighted in Algorithm 1.

VI. SIMULATION

A. Simulation Setup

1) Datasets: We select typical bidirectional road traffic scenarios from UA-DETRAC dataset [8] to serve as the training data for the generative semantic inference module, predictive frame interpolation, as well as the semantic sampling agent. Specifically, as this work excludes the optimization of the object detection module, we directly use the information about bounding box and the type of all vehicle objects recorded in the *Annotations* as the extracted textual semantic information. Moreover, when mapping the visual layout, we set the pixel values corresponding to the four categories of *car*, *bus*, *van*,
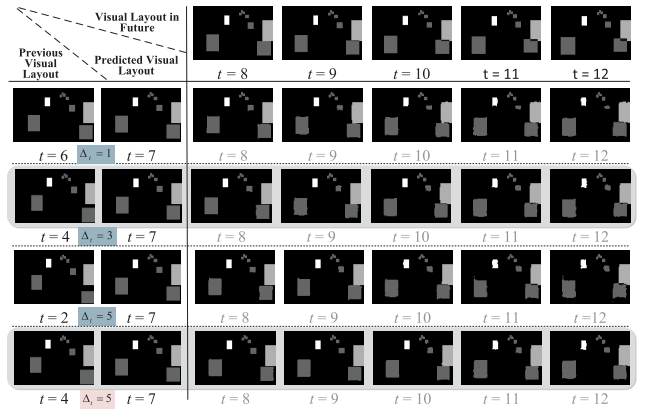

Fig. 7. Visual results for visual layout prediction.

and *others* to 1, 2, 3, and 4, respectively. Moreover, during the training, the image size is reshaped into $(240, 160)$. Meanwhile, for the wireless transmission parameter settings, the detailed parameter settings can be found in Table III.

2) Network Design and Hyperparameters: For neural network of the predictive frame interpolation module shown in Fig. 5, the number of channels in each ResidualBlock is set to 8, and the number of the channels in the output layer is set to 5, i.e., $P = 5$. The dimension of the time embedding is set to 32. For the generative semantic inference module as shown in Fig. 3, it comprises a number of channels equal to [64, 64, 128, 128, 256, 256, 512, and 512] and with a linear variance schedule. Meanwhile, the dimension of the time embedding is set to 256. Finally, we adopt the classifier-free guidance [34] with the guidance scale of $k = 8$. For the K-SAC-based semantic sampling policy, both the policy network and the two critic networks are designed with three layers, with 300, 200, and 200, respectively. In addition, for all the models, the Adam optimizer is adopted with betas =(0.9, 0.999). Moreover, more hyperparameters during the training for each model have been summarized in Table III.

B. Performance Analysis

We first analyze the performance of each module sequentially according to their operational order. Finally, we present a comprehensive end-to-end performance analysis.

1) Predictive Frame Interpolation: The predicted visual layouts with different sampling intervals are shown in Fig. 7. Notably, during predictions with $\Delta_t = 1$ STI, $\Delta_t = 3$ STI, and $\Delta_t = 5$ STI, the positional consistency of objects is preserved. This ensures the fidelity of semantic information between the real-world scene and the generated scene. However, since the module can only grasp the relative positions of the objects in the two layouts and the time differences based on the inputs, the prediction relies on an implicit average driving speed. As a result, due to the non-uniform speed, careful observation of Fig. 7 reveals that there are some slight discrepancies corresponding to different sampling time intervals. This can be further confirmed by comparing the second and fourth rows in Fig. 7, which use the same layouts for prediction but different sampling intervals. Meanwhile, this observation also confirms the validity of the prediction module. In addition,
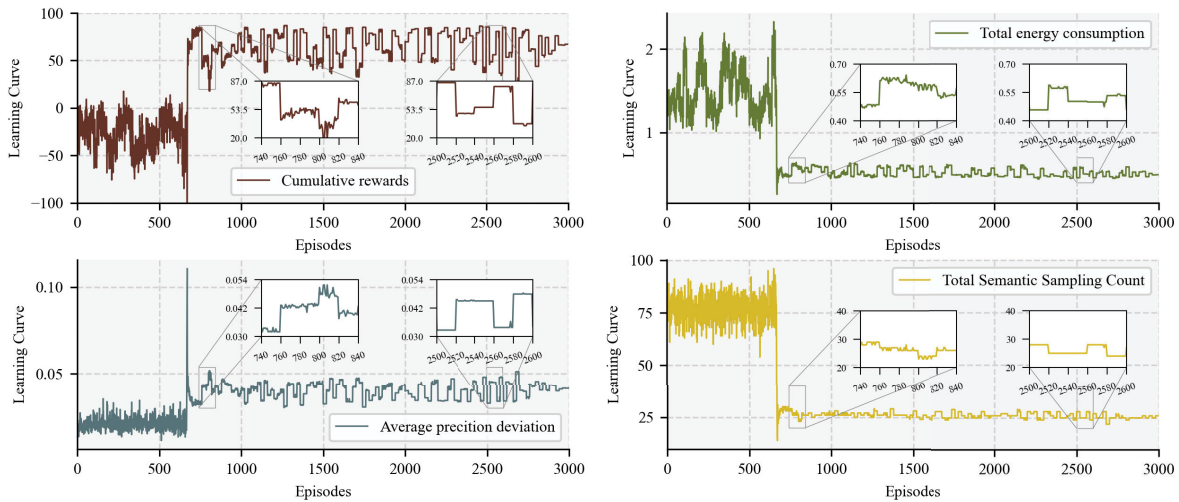
Fig. 8. Learning performance for K-SAC agents with *random initial scene with different traffic pattern* in terms of cumulative reward, total energy consumption, average prediction deviation, and total semantic sampling count.

as shown in Fig. 7, the contours of the later predicted frames become increasingly blurry, which may also interfere with the evaluation of semantic accuracy. In this sense, even if computational power permits, the value of P should not be too large. Therefore, considering both semantic accuracy and the complexity of frequent predictions, we select $P = 5$ based on experimental results.

2) K-SAC Based Semantic Sampling: To enhance the robustness and adaptability of the algorithm, we select 500 frames of continuous surveillance scenes as the training environment. These frames include scenarios with varying vehicle types, vehicle counts, vehicle speeds, and changes in vehicle speeds. Specifically, the real-world semantic information is transformed into visual layouts based on the annotations from UA-DETRAC XML files. The visual layout prediction utilizes the model trained in Section VI-B.1. The convergence of the algorithm is illustrated in Fig. 8, where an episode refers to an observation window.

It is important to note that during the training process, *the initial state (the scene with different traffic pattern) is randomly regenerated every 20 episodes*. This differs from the majority of current RL training methodologies, which use the same or similar initial states across different episodes, allowing the reward function to converge to a stable value for each episode. In our training process, the episodes with different initial states always encompass different observed environments, where variations in the number of vehicles and the vehicular speed inevitably affect the communication system's performance in terms of energy consumption and prediction accuracy. Due to the inherent randomness of traffic flow, it is neither feasible nor necessary to compensate for its impact on performance through the design of the reward function. Consequently, as expected, the cumulative reward for different observation windows deviates after convergence of the algorithm. Nevertheless, aside from the overall increase in the cumulative reward function value indicating the convergence of the algorithm, this stability of the cumulative reward for the same observed environment, as locally magnified in Fig. 8, also corroborates the convergence of the

algorithm. From Fig. 8, we can observe that during the initial convergence phase, from episode 740 to episode 840, all the performance metrics exhibit differences across the five observation windows. However, there are still noticeable fluctuations within each observation window. In contrast, during the late convergence phase, from episode 2500 to episode 2600, the performance metrics within the different windows have stabilized, with only minor variations attributable to inherent randomness of the strategy. Moreover, by comparing the locally magnified view of the total energy consumption with the locally magnified view of the number of samples, we can observe that fewer samples often correspond to higher energy consumption. This reflects the agent's energy-saving awareness and adaptive capability.

Furthermore, the superiority of agent-assisted semantic sampling compared to existing periodic sampling strategies has been demonstrated in Fig. 9. Specifically, we consider periodic sampling strategies with periods of 4, 5, 6, and 7 STI for three footage clips with different initial moments.[6]. Firstly, from the perspective of the comprehensive evaluation indicator, cumulative reward, all sampling strategies exhibit varying performances across different footage clips. Among them, the agent-assisted sampling strategy, due to its robustness and adaptability, achieved significantly superior performance in all three footage clips. Additionally, there is no optimal sampling period for the periodic sampling strategies because of the dynamic changes in the environment. For example, the periodic sampling strategy with a period of 7 STI achieved the best performance among the four periodic sampling strategies in Clip 1, whereas in Clip 2, its performance was the worst, resulting in negative values. The reason for this can be understood by comparing the subsequent three subplots. Although the periodic sampling strategy has the same sampling counts in different footage clips, it achieves the lowest energy consumption in Clip 1. This implies that the smaller number of vehicles in Clip 1 allows for more sampling with the same energy consumption, further reducing prediction deviation and

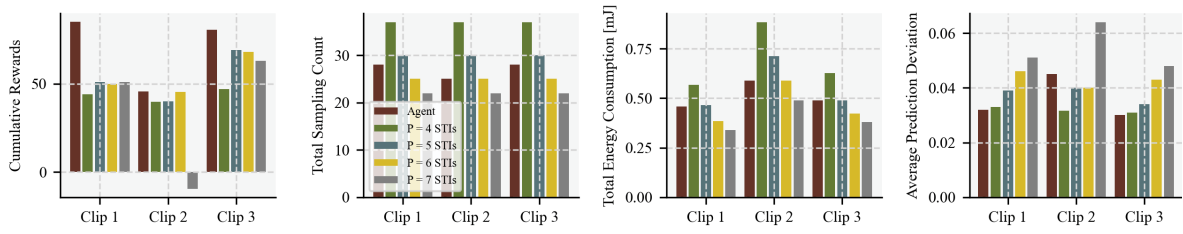[6]The videos about the three footage clips can be found at **https://youtu.be/pJZzDnNspAc**

Fig. 9. Performance comparison about A-GSC where semantic sampling is performed adaptively by the agent, and GSCs where the semantic sampling is performed periodically with multiple periods (denoted by P) for different footage clips. Semantic extraction in all GSC paradigms is performed by the semantic extraction module stated in Section IV-A.
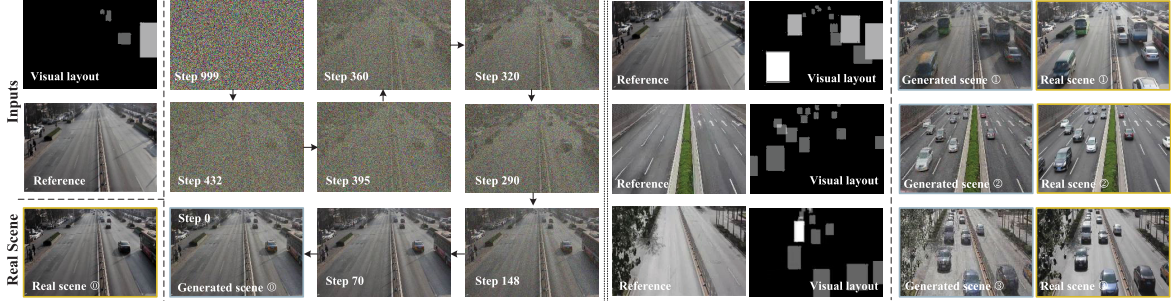

Fig. 10. Visual results for generative semantic inference.

improving overall performance. The agent-assisted sampling strategy exactly achieves this effectively. In Clip 2, the prediction deviation of the periodic sampling strategy with a period of 7 STI significantly increases compared to the strategies with periods of 5 and 6 STI, suggesting that there may be sudden changes in traffic flow within a certain sampling interval. The agent-assisted sampling strategy addresses this issue by increasing a few sampling amount with minimal energy consumption, further demonstrating its robustness and adaptability. From the overall perspective of Fig. 9, the average sampling interval for the agent-assisted semantic sampling is approximately between 5 STI and 6 STI for different footage clips, which is close to the value of P. This is because, although a corresponding prediction mechanism at the destination side compensates for the semantic information loss caused by sampling, as mentioned in Section VI-B.1, the accumulation of prediction errors increases with longer prediction interval. Especially when the next round of prediction is based on two consecutive predicted frames, this error is further amplified.

3) Generative Semantic Inference: The visual results of the generative semantic inference can be found in Fig. 10. Specifically, the left side of the figure shows the specific denoising process, and the right side shows more examples generated under different references. By comparing the generated image and the real image, it becomes evident that, while there exist differences in the background details and coloration of the vehicles, the positional alignment of the vehicles remains accurate. This achievement signifies the system's proficiency in preserving essential semantic information. However, it should be noted that if the boundary vehicle fails to be detected at the source, the generated image at the destination will not contain that information (as shown in Example ① in Fig. 10).

4) Overall Performance: Finally, we analyze the overall performance gain of the proposed A-GSC framework. To demonstrate the superior performance in terms of energy
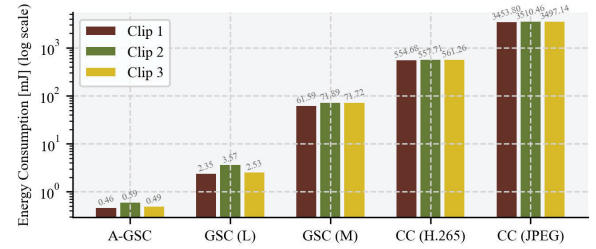

Fig. 11. Comparison of energy consumption.

efficiency, we conduct a comparison of the proposed A-GSC between the existing generative SemComs (GSCs) with semantic map and textual layout as prompt (denoted by GSC (M) and GSC (L) respectively), as well as conventional communications (CCs) with the source coding H.265[7] and JPEG, respectively. Although the advanced H.265 encodes only the differences between blocks of pixels within a frame or between consecutive frames, it treats all changes equally. Therefore, due to the varying dynamics of objects and backgrounds, H.265 still requires 10 times more energy consumption compared to GSC (M), as shown in Fig. 11. Furthermore, from the comparison of GSC (L) and GSC (M), We can observe that the introduction of image-to-text conversion is well worth it in terms of improving energy efficiency, as it can reduce the amount of transmitted data to approximately 1/30 of the original. Moreover, comparing the energy consumption of A-GSC and the existing GSC, the integration of semantic sampling agent further reduces energy consumption to 1/5 from the temporal domain, compared to GSC (L). Considering the substantial differences in energy consumption across different communication paradigms, *log scale* is adopted in Fig. 11. Overall, as can be clearly seen from Fig. 11, the proposed A-GSC reduces transmission energy consumption

---

[7]We employ FFmpeg, leveraging its integration with the libx265 library for H.265/HEVC encoding, and acquire the data size of each encoded frame.
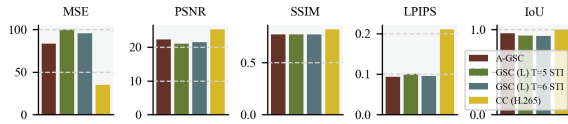
Fig. 12. Comparison of the accuracy of the scene reconstruction with different strategies under multiple metrics.

Fig. 13. Limitations of proposed A-GSC.

by approximately 1,000 times compared to the current H.265-based encoding.

After demonstrating the great potential for energy savings, we evaluate the performance of scene reconstruction based on multiple metrics. We compare our method with GSC (L) periodic sampling at intervals of T=5 STI and T=6 STI, which have shown better performance as depicted in Fig. 9, as well as with conventional coding (CC) using H.265. Moreover, it is noted that in the CC with H.265, we resize the frame into (120, 80) before transmission. Then, during the reconstruction, we resize it to (240, 160). This is consistent with the use of the (120, 80) frame size in the visual layout prediction. The visual video for the source data sensed at the source and the reconstructed scene by multiple communication paradigms can be found at **https://youtu.be/9tNKq0RQ58Y**. From the video, we can see that due to the stochastic nature of the inference process, the color of the vehicles may change across different scenes, and the brightness of the image also fluctuates a bit. Both observations restrict GSCs' performance in the evaluation under the conventional metrics of MSE and PSNR, the SSIM, but they are still at a satisfactory level as shown in Fig. 12. In contrast, under the semantic metric Learned Perceptual Image Patch Similarity (LPIPS) and Intersection over Union (IoU) about the considered task, the proposed algorithm shows its superiority. Compared to periodic sampling, the performance of agent-assisted GSC improves by approximately 6% and 4%. In addition, due to the photo-realistic generation capability of the diffusion model and high compression rate adopted in H.265, the GSCs' performance is 2 times that of conventional coding (CC) with H.265 under the LPIPS metric. Overall, the above demonstrates that A-GSC framework can also achieve fulfilling performance.

However, the framework currently has limitations. Fig. 13 highlights three potential unexpected issues. First, while the accuracy of vehicle position changes can be maintained in a video, the color of a vehicle may be altered. Second, for larger objects such as buses, it is occasionally hard to determine whether a single vehicle or two overlapping vehicles are present. Third, with boundary objects, what should be relatively large incomplete objects may sometimes be interpreted as a smaller complete vehicle. Although these errors do not affect semantic accuracy, further improvements are necessary.

C. Complexity Analysis of Training and Deployment

In the proposed A-GSC framework, multiple trainable modules are incorporated, including the semantic extraction module, semantic sampling agent, predictive frame interpolation module, and generative semantic inference module. Nevertheless, this explainable modular design allows each module to be trained independently, significantly reducing the overall training complexity. Specifically, for the source end, a well-trained object detection algorithm can be utilized as a pre-trained model, which greatly mitigates the complexity of training from scratch. Additionally, for the semantic sampling agent, since it performs a binary decision task and the state size is relatively small, the convergence rate is fast, requiring only a few hundred episodes to reach a satisfactory performance, as shown in Fig. 8). Moreover, due to the small dimensions of the action and state spaces, the storage space required for an experience replay buffer of 100,000 pieces of experience is only 8.5 MB. Therefore, it also has minimal requirement for storage. On the destination side, the visualized layout information primarily consists of low-level features, significantly reducing both model design and training complexity compared to existing video frame prediction methods, with only 500 episodes needed. Furthermore, for the generative semantic inference module, as this task does not involve complex cross-modal feature alignment like stable diffusion with text-based prompts, its model size and training difficulty are greatly reduced compared to general generative models such as stable diffusion, requiring only 400 episodes. Meanwhile, if this communication framework is to be applied to other scenarios with richer and more complex semantic information, well-trained models like ControlNet can be employed as pre-trained models for further fine-tuning. Overall, despite the increase in the number of modules requiring training, the training complexity is much lower compared to JSCC SemCom approaches based on Transformers, which typically require millions or even tens of millions of episodes [23].

Additionally, for the deployment process, a lightweight YOLO object detection algorithm designed for mobile devices can be used as the semantic extraction module. Furthermore, the semantic sampling agent, with a small-scale actor network consisting of 0.102018 million parameters, requires only 0.102656 million FLOPs per inference. Compared to the lightweight and widely used YOLOX model [44], which requires 1.08G FLOPs and 0.91 million parameters, we believe that integrating the K-SAC agent into embedded devices is feasible. Meanwhile, on the destination side, the long-term frame prediction in the predictive frame interpolation module has reduced the complexity of frequent predictions. For the generative semantic inference module, due to its smaller model size, the inference latency for a single denoising step is also reduced. Furthermore, it is worth highlighting that advanced diffusion models such as InstaFlow and DDIM have optimized the inference process of classic DDPMs by reducing the number of denoising steps, thereby lowering the inference latency to 0.1 seconds and 1 second, respectively. This demonstrates the potential feasibility of applying diffusion models in surveillance systems.

VII. Conclusion

In this paper, we have focused on a remote surveillance scenario. In contrast to the existing research efforts for SemCom, which have focused solely on semantic extraction or semantic sampling, we have seamlessly integrated both together by introducing the GAI and an RL-based agent into SemCom system design. Specifically, in the proposed A-GSC framework, the semantic information has served as an explainable prompt of textual layout information transmitted to the semantic decoder while assisting in semantic sampling. To consistently display the remote scene, we have integrated prediction capabilities into the decoder design, to complement the flexible semantic sampling at the semantic encoder. Our simulations have demonstrated notable improvements in both energy efficiency and reconstruction accuracy. In the future, we will further extend the work to a multi-sensor scenario, and aim to jointly optimize semantic compression as well as semantic scheduling.

References

[1] W. Yang, Z. Xiong, H. Du, Y. Yuan, and T. Q. S. Quek, "Semantic change driven generative semantic communication framework," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2024, pp. 1–6.

[2] S. Iyer et al., "A survey on semantic communications for intelligent wireless networks," *Wireless Pers. Commun.*, vol. 129, no. 1, pp. 569–611, 2023.

[3] Y. Zhang, F. Wang, W. Xu, and C. Liu, "Semantic communications: A new paradigm for networked intelligence," in *Proc. IEEE 32nd Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Aug. 2022, pp. 1–6.

[4] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.

[5] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[6] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[7] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *Proc. IEEE Int. Conf. Auto. Syst. (ICAS)*, Aug. 2021, pp. 1–5.

[8] L. Wen et al., "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understand.*, vol. 193, Apr. 2020, Art. no. 102907.

[9] S. K. Yoo et al., "A comprehensive analysis of the achievable channel capacity in $\mathcal{F}$ composite fading channels," *IEEE Access*, vol. 7, pp. 34078–34094, 2019.

[10] W. Zhang, H. Zhang, H. Ma, H. Shao, N. Wang, and V. C. M. Leung, "Predictive and adaptive deep coding for wireless image transmission in semantic communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5486–5501, Aug. 2023.

[11] S. Wang et al., "Wireless deep video semantic transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 214–229, Jan. 2023.

[12] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, "Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 170–185, Jan. 2022.

[13] W. Zhang et al., "DeepMA: End-to-end deep multiple access for wireless image transmission in semantic communication," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 2, pp. 387–402, Apr. 2024.

[14] F. Zhao, N. Pappas, C. Ma, X. Sun, T. Q. S. Quek, and H. H. Yang, "Age-threshold slotted ALOHA for optimizing information freshness in mobile networks," 2023, *arXiv:2312.10888*.

[15] A. Maatouk, M. Assaad, and A. Ephremides, "The age of incorrect information: An enabler of semantics-empowered communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2621–2635, Apr. 2023.

[16] J. Holm et al., "Freshness on demand: Optimizing age of information for the query process," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[17] A. Nikkhah, A. Ephremides, and N. Pappas, "Age of actuation in a wireless power transfer system," 2023, *arXiv:2303.00507*.

[18] H. Nam, J. Park, J. Choi, M. Bennis, and S.-L. Kim, "Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation," 2023, *arXiv:2309.11127*.

[19] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," 2023, *arXiv:2306.04321*.

[20] H. Feng, Y. Yang, and Z. Han, "Scalable AI generative content for vehicular network semantic communication," 2023, *arXiv:2311.13782*.

[21] A. D. Raha, M. S. Munir, A. Adhikary, Y. Qiao, and C. S. Hong, "Generative AI-driven semantic communication framework for NextG wireless network," 2023, *arXiv:2310.09021*.

[22] E. Grassucci, C. Marinoni, A. Rodriguez, and D. Comminiello, "Diffusion models for audio semantic communication," 2023, *arXiv:2309.07195*.

[23] K. Yang, S. Wang, J. Dai, X. Qin, K. Niu, and P. Zhang, "Swin-JSCC: Taming Swin transformer for deep joint source-channel coding," *IEEE Trans. Cogn. Commun. Netw.*, early access, Jul. 8, 2024, doi: 10.1109/TCCN.2024.3424842.

[24] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[25] W. Yang, X. Chi, L. Zhao, Z. Xiong, and W. Jiang, "Task-driven semantic-aware green cooperative transmission strategy for vehicular networks," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 5783–5798, Oct. 2023.

[26] M. K. Farshbafan, W. Saad, and M. Debbah, "Curriculum learning for goal-oriented semantic communications with a common language," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1430–1446, Mar. 2023.

[27] E. Uysal et al., "Semantic communications in networked systems: A data significance perspective," *IEEE Netw.*, vol. 36, no. 4, pp. 233–240, Jul. 2022.

[28] Q. Chen et al., "Peak AoI minimization with directional charging for data collection at wireless-powered network edge," *IEEE Trans. Services Comput.*, vol. 17, no. 5, pp. 2747–2761, Sep. 2024.

[29] O. Ayan, M. Vilgelm, M. Klügel, S. Hirche, and W. Kellerer, "Age-of-information vs. value-of-information scheduling for cellular networked control systems," in *Proc. 10th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, Apr. 2019, pp. 109–117.

[30] J. Ren et al., "Knowledge base enabled semantic communication: A generative perspective," 2023, *arXiv:2311.12443*.

[31] J. Liu and Z. Liu, "YOLOv5s-BC: An improved YOLOv5s-based method for real-time apple detection," *J. Real-Time Image Process.*, vol. 21, no. 3, pp. 1–16, Jun. 2024.

[32] Y. Wang, W. Yang, Z. Xiong, Y. Zhao, T. Q. S. Quek, and Z. Han, "Harnessing the power of AI-generated content for semantic communication," *IEEE Netw.*, vol. 38, no. 5, pp. 102–111, Sep. 2024.

[33] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2337–2346.

[34] W. Wang et al., "Semantic image synthesis via diffusion models," 2022, *arXiv:2207.00050*.

[35] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "LayoutDiffusion: Controllable diffusion model for Layout-to-Image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22490–22499.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[37] C. She et al., "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proc. IEEE*, vol. 109, no. 3, pp. 204–246, Mar. 2021.

[38] D. Zwillinger and A. Jeffrey, *Table of Integrals, Series, and Products*. Amsterdam, The Netherlands: Elsevier, 2007.

[39] T. Haarnoja et al., "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.

[40] F. Xu, Y. Ruan, and Y. Li, "Soft actor–critic based 3D deployment and power allocation in cell-free unmanned aerial vehicle networks," *IEEE Wireless Commun. Lett.*, vol. 12, no. 10, pp. 1692–1696, Oct. 2023.

[41] C. Diehl, T. S. Sievernich, M. Krüger, F. Hoffmann, and T. Bertram, "Uncertainty-aware model-based offline reinforcement learning for automated driving," *IEEE Robot. Autom. Lett.*, vol. 8, no. 2, pp. 1167–1174, Feb. 2023.

[42] Y. Zhou, H. Dong, and A. El Saddik, "Deep learning in next-frame prediction: A benchmark review," *IEEE Access*, vol. 8, pp. 69273–69283, 2020.

[43] D. Liu, J. Zhao, C. Yang, and L. Hanzo, "Accelerating deep reinforcement learning with the aid of partial model: Energy-efficient predictive video streaming," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3734–3748, Jun. 2021.

[44] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

Wanting Yang received the B.S. and Ph.D. degrees from the Department of Communications Engineering, Jilin University, Changchun, China, in 2018 and 2023, respectively. She is currently a Research Scientist with Singapore University of Technology and Design. Her research interests include semantic communication, deep reinforcement learning, martingale theory, edge computing, edge intelligence, and generative AI. She served as a Technical Program Committee Member for flagship conferences, such as WCNC, Globecom, ICC, and VTC.

Zehui Xiong (Senior Member, IEEE) received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He was a Visiting Scholar with Princeton University and the University of Waterloo. He has served as the Associate Director of the Future Communications Research and Development Program and the Deputy Lead of the AI Mega Centre. He is currently with Singapore University of Technology and Design, Singapore. Recognized as a Highly Cited Researcher, he has published more than 250 research articles in leading journals. His research interests include wireless communications, the Internet of Things, blockchain, edge intelligence, and metaverse. He has won over ten best paper awards at international conferences. He was a recipient of the Forbes Asia 30u30, the IEEE Asia–Pacific Outstanding Young Researcher Award, the IEEE Early Career Researcher Award for Excellence in Scalable Computing, the IEEE Technical Committee on Blockchain and Distributed Ledger Technologies Early Career Award, the IEEE Internet Technical Committee Early Achievement Award, the IEEE TCSVC Rising Star Award, the IEEE TCI Rising Star Award, the IEEE TCCLD Rising Star Award, the IEEE ComSoc Outstanding Paper Award, the IEEE Best Land Transportation Paper Award, the IEEE Asia–Pacific Outstanding Paper Award, the IEEE CSIM Technical Committee Best Journal Paper Award, the IEEE SPCC Technical Committee Best Paper Award, the IEEE Big Data Technical Committee Best Influential Conference Paper Award, and the IEEE VTS Singapore Best Paper Award. He is serving as an Editor or a Guest Editor for many leading journals, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING.

Yanli Yuan received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Beihang University, Beijing, China, in 2012 and 2015, respectively, and the Ph.D. degree in electronic engineering from Singapore University of Technology and Design, Singapore, in 2021. She is currently an Assistant Professor with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing. Her research interests include trustworthy graph learning, graph signal processing, and federated learning.
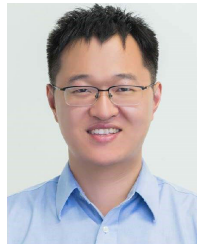
Wenchao Jiang (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, University of Minnesota Twin Cities. He is currently an Assistant Professor with the Pillar of Information System Technology and Design, Singapore University of Technology and Design. His works appear on top-tier conferences and journals, including SigComm, NSDI, MobiCom, INFOCOM, SenSys, IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON MOBILE COMPUTING, and IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. His research interests include the artificial intelligence of things (AIoT), wireless sensing, and mobile computing.
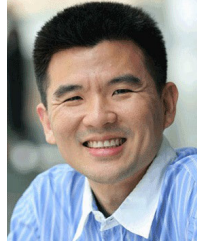
Tony Q. S. Quek (Fellow, IEEE) received the B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from Massachusetts Institute of Technology in 2008.

Currently, he is the Cheng Tsang Man Chair Professor with Singapore University of Technology and Design (SUTD) and the ST Engineering Distinguished Professor. He also serves as the Director of the Future Communications Research and Development Program, the Head of the ISTD Pillar, and the AI on RAN Working Group Chair of the AI-RAN Alliance. His research interests include wireless communications and networking, network intelligence, non-terrestrial networks, open radio access networks, and 6G.

Dr. Quek is a WWRF Fellow and a fellow of the Academy of Engineering Singapore. He was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards—Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, the 2020 Nokia Visiting Professor, the 2022 IEEE Signal Processing Society Best Paper Award, and the 2024 IIT Bombay International Award For Excellence in Research in Engineering and Technology. He has been actively involved in organizing and chairing sessions and has served as a member of the technical program committee and the symposium chair of several international conferences. He is serving as an Area Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Mérouane Debbah (Fellow, IEEE) is currently a Professor with the Khalifa University of Science and Technology, Abu Dhabi, and the Founding Director of the KU 6G Research Center. He is a frequent keynote speaker at international events in the field of telecommunication and AI. His research interests include the interface of fundamental mathematics, algorithms, statistics, information, and communication sciences, with a special focus on random matrix theory and learning algorithms. In the communication field, he has been at the heart of the development of small cells (4G), massive MIMO (5G), and large intelligent surfaces (6G) technologies. In the AI field, he is known for his work on large language models, distributed AI systems for networks, and semantic communications. He received multiple prestigious distinctions, prizes, and best paper awards for his contributions to both fields. He is a WWRF Fellow, an Eurasip Fellow, an AAIA Fellow, an Institut Louis Bachelier Fellow, and a Membre émérite SEE.