Generative AI for Semantic Communication: Architecture, Challenges, and Outlook

Le Xia, Yao Sun, Chengsi Liang, Lei Zhang, Muhammad Ali Imran, Dusit Niyato.
University of Glasgow

CONTENT

Introduction

■ Background

Current SemCom lacks context reasoning and background knowledge, needing integration with generative AI (GAI).

■ Chanllenge

- *SemCom requires vast data for background knowledge and model pre-training.*
- *Existing systems lack semantic accuracy when handling complex contexts.*
- *Large GAI models need significant computing and storage resources.*

■ Solution

- *Propose a GAI-assisted semantic communication network (GAI-SCN) that collaborates across cloud, edge, and mobile layers using global and local GAI models to provide multimodal content and joint source-channel coding.*

Introduction

When SemCom Meets GAI

- *Current semantic communication (SemCom) systems lack context reasoning and background knowledge.*
- *Generative AI (GAI) shows potential in creating valuable, diverse, and personalized content.*
- *Integrating GAI with SemCom can improve training efficiency, context reasoning, and spectrum utilization.*
- *A new GAI-assisted SemCom network (GAI-SCN) framework aims to enhance semantic reasoning and resource efficiency.*

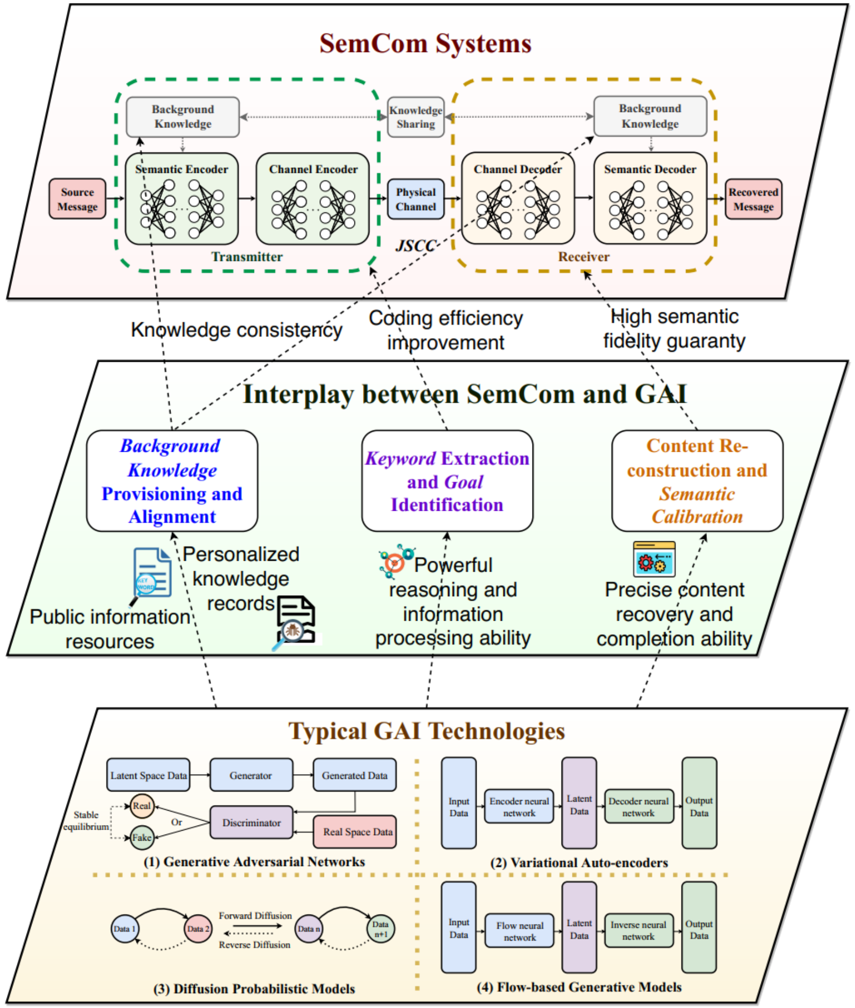


Fig. 1. Overview of SemCom systems and four types of typical GAI technologies along with three aspects of interplay between SemCom and GAI.

System Design

■ Interplay between SemCom and GAI

☐ Background Knowledge:

- *Provides global and personalized data for consistency and customization.*

☐ Keyword Extraction:

- *Extracts keywords and identifies goals, reducing data transmission.*

☐ Content Reconstruction:

- *Reconstructs content and corrects errors, improving accuracy and reliability.*

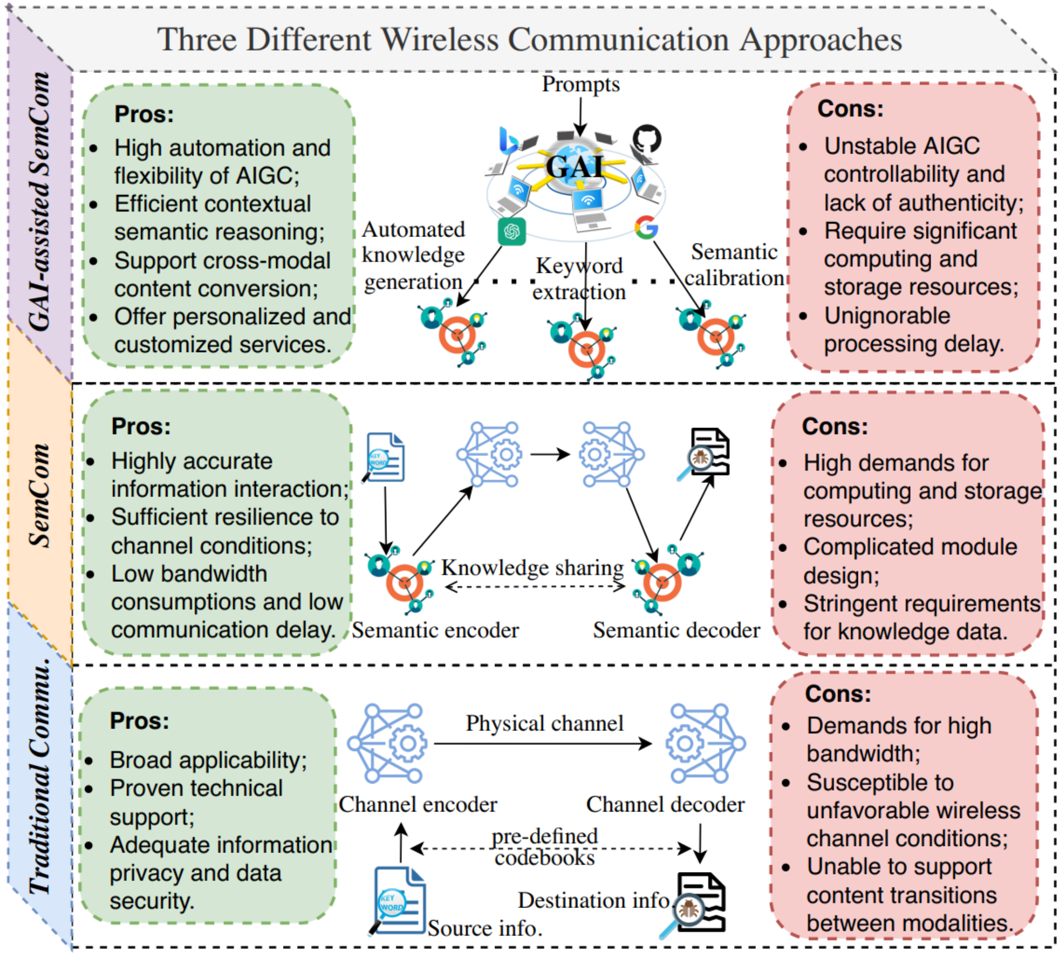


Fig. 2. Comparisons among three different approaches of GAI-SemCom, SemCom, and traditional communication in terms of their pros and cons.

System Design

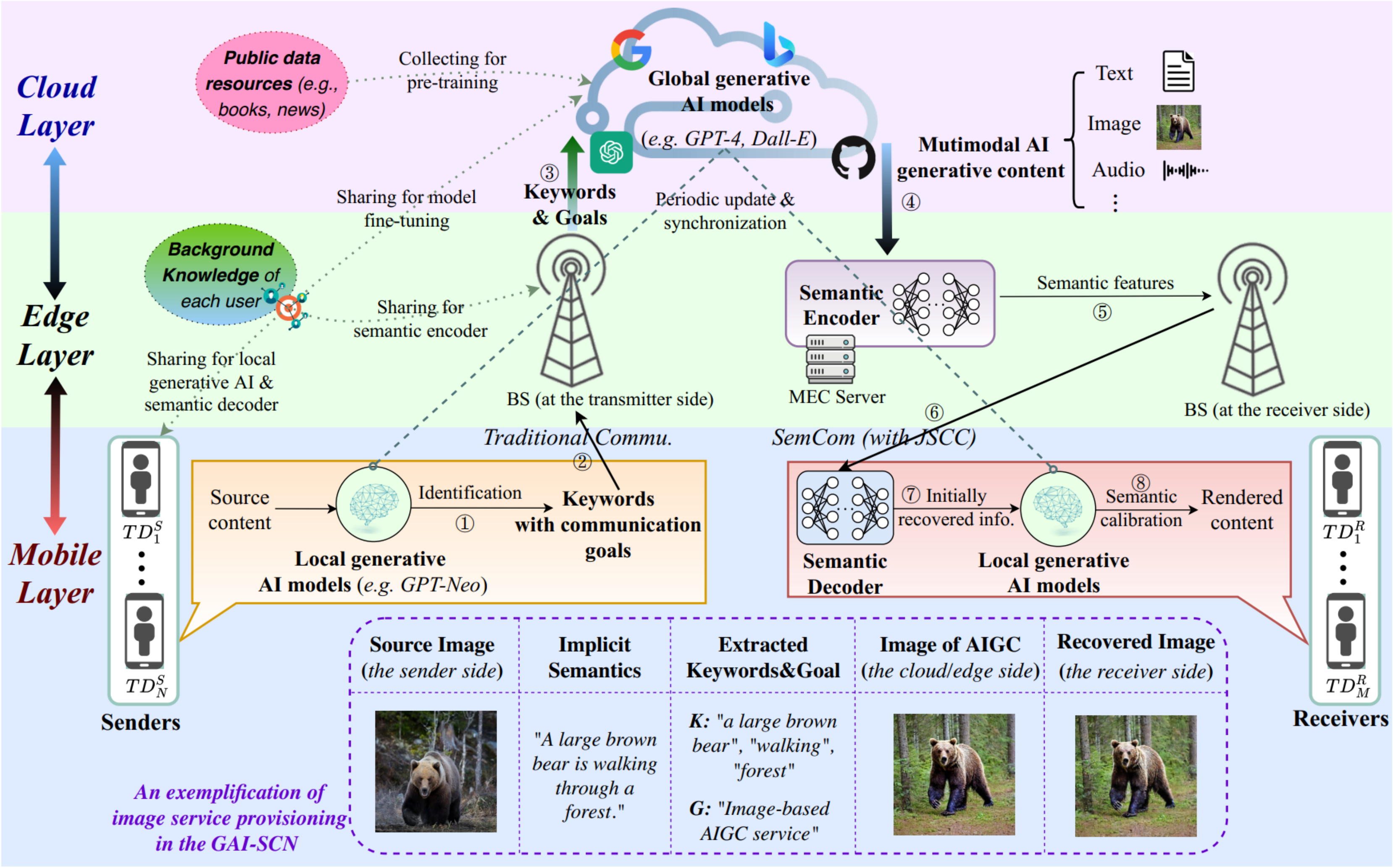
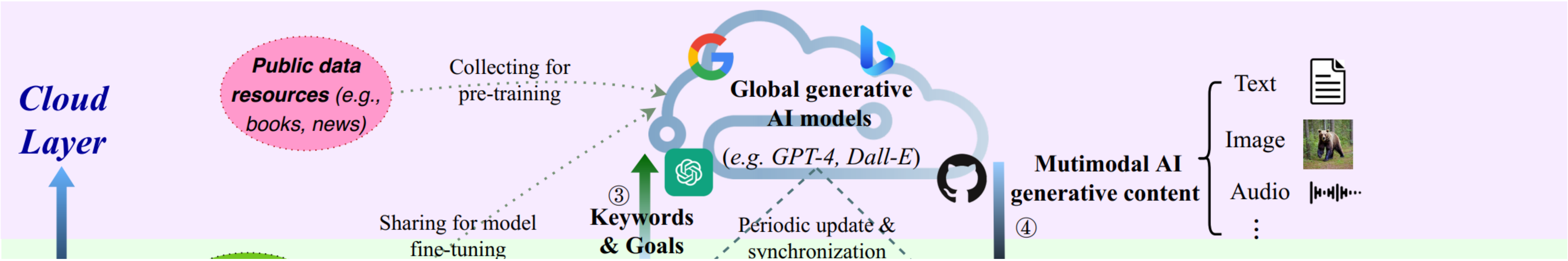


Fig. 3. Illustration of the proposed GAI-SCN framework in a collaborative cloud-edge-mobile design, where an exemplification of image service provisioning is presented.

System Design

Mobile (Local) Layer:

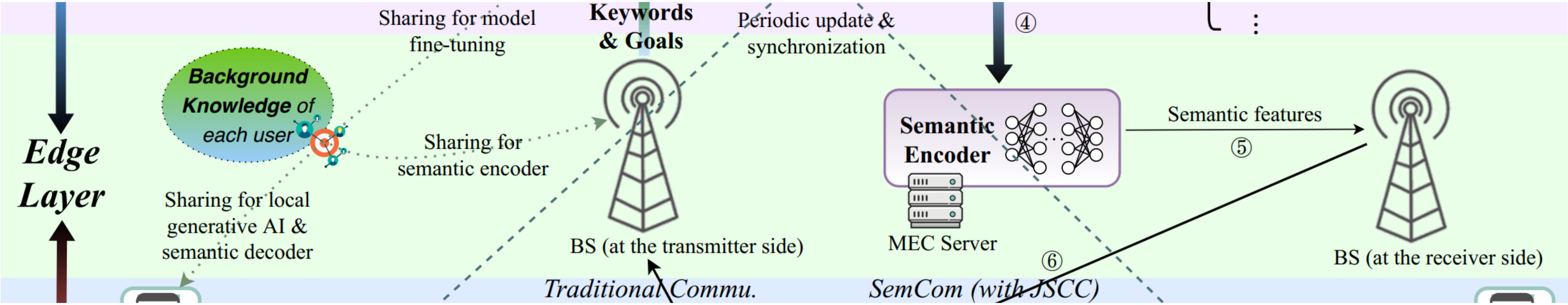
- **Role**: *Handle local processing and personalization.*
- **Components**: *Lightweight GAI models (e.g., GPT-Neo).*
- **Functions**:
  - *Extract keywords and identify communication goals from user data.*
  - *Perform initial semantic decoding and calibration.*
  - *Fine-tune user preferences and provide personalized content.*



System Design

Edge Layer

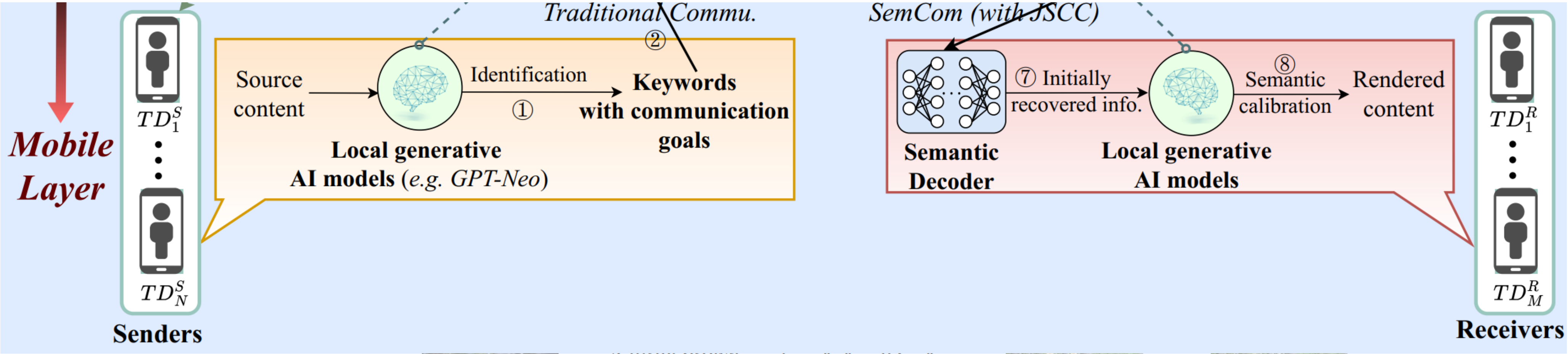
- **Role**: *Intermediate processing to enhance resilience and efficiency.*
- **Components**: *Semantic encoders and edge servers.*
- **Functions**:
  - *Pre-process and encode AI-generated content (AIGC) using joint source-channel coding (JSCC).*
  - *Offload computational tasks from the cloud to reduce latency.*
  - *Ensure robustness and resilience of SemCom against channel impairments.*



System Design

Cloud Layer

- **Role**: *Handle heavy computational tasks and global model management.*
- **Components**: *Large GAI models (e.g., GPT-4, Dall-E).*
- **Functions**:
  - *Generate AI-generated content (AIGC) based on keywords and goals provided by the local layer.*
  - *Perform extensive data analysis and model training using vast computational resources.*
  - *Synchronize and update models with data from edge and local layers.*



Evaluation

■ Model Selection

- **Local GAI Model**: *ViT combined with GPT-2 for image-to-text conversion and keyword extraction.*
- **Global GAI Model**: *Stable Diffusion 2.1 for generating AI-generated images from prompts.*
- Each edge server has identical storage capacity.

■ Semantic Communication Setup:

- Use deep convolutional network (Observation-Centric Sort) and Transformer-based semantic decoder for segmentation and recovery.
- Train models with additive white Gaussian noise channel at 0 dB SNR, testing with 327 images.

Evaluation

- *The complexity of images significantly impacts the recovery performance when using the GAI-SCN framework.*

- *As the number of observable objects in the original images increases:*
  - **Semantic similarity decreases.**
  - **Object quantity discrepancy increases.**
  - **Recovery ratio of original objects declines.**

- *This demonstrates that the framework faces greater challenges in maintaining recovery accuracy and semantic consistency with more complex images.*

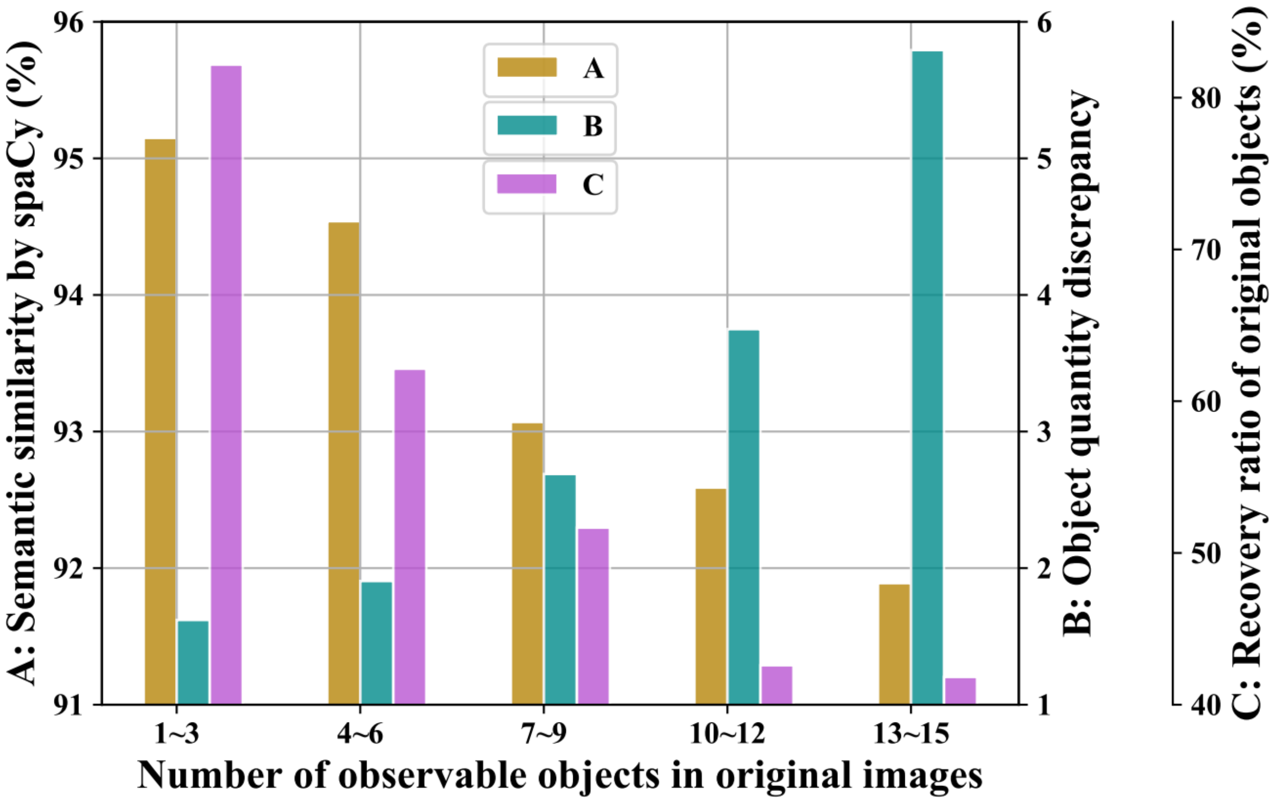


Fig. 5. Comparisons between original and recovered images by the proposed GAI-SCN framework in terms of three metrics: A) Semantic similarity by spaCy; B) Object quantity discrepancy; C) Recovery ratio of original objects.

OPEN RESEARCH ISSUES AND OUTLOOKS

■ Limited Device Resources

- Challenge:
  - *Implementing sophisticated AI models on devices with limited storage, memory, and computational power.*
- Potential Solution:
  - *Use model compression and acceleration techniques like knowledge distillation, parameter pruning, and quantization to reduce model complexity and size.*

OPEN RESEARCH ISSUES AND OUTLOOKS

■ Randomness in GAI Content

- Challenge:
  - *Variability in AI-generated content and semantic decoder outputs, leading to inconsistencies.*
- Potential Solution:
  - *Investigate granularity tuning for keyword extraction and semantic calibration to mitigate randomness.*

OPEN RESEARCH ISSUES AND OUTLOOKS

■ Inactive Sharing of Knowledge and Preferences

- **Challenge**:
  - *Encouraging users to share personal preferences and background knowledge necessary for customized AI and SemCom services.*
- **Potential Solution**:
  - *Develop incentive mechanisms, such as rewards or benefits, to motivate users to contribute their data for system improvements.*

Conclusions

Pros

- *This article shows us the interplay between semantic communication and AIGC, as well as taking advantage of the edge for information transmission.*
- *It also shows us the research direction of the future development of voice communication.*

Cons

- *The experimental discussion is not detailed enough.*

Thank You!