NEURAL INFORMATION PROCESSING SYSTEMS

AI4NextG @ NeurIPS'25

Beyond Prompts: Preserving Semantics in Diffusion-based Communication

2025.12.06

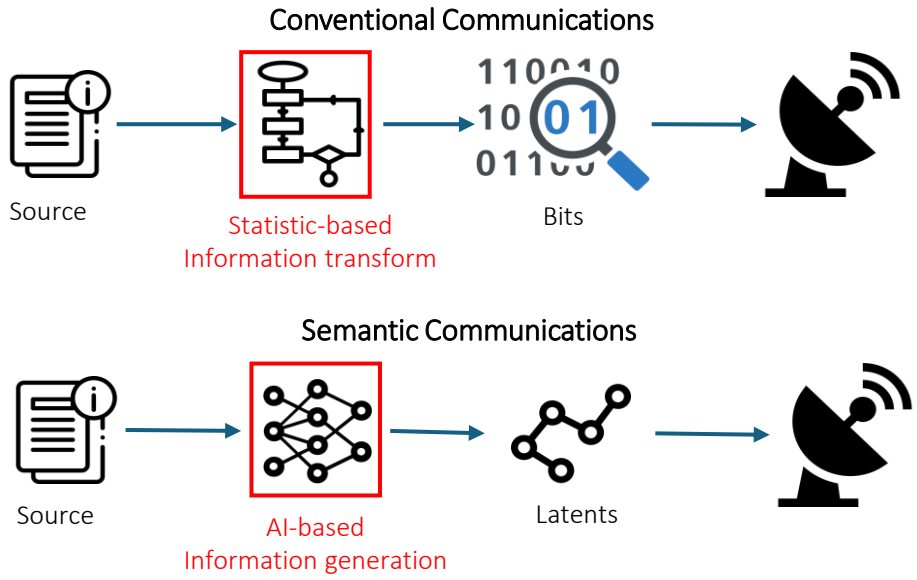
Wonjung Kim[1], Nakyung Lee[1], Sangwoo Hong[2], Jungwoo Lee[1]

Seoul National University[1], Konkuk University[2]

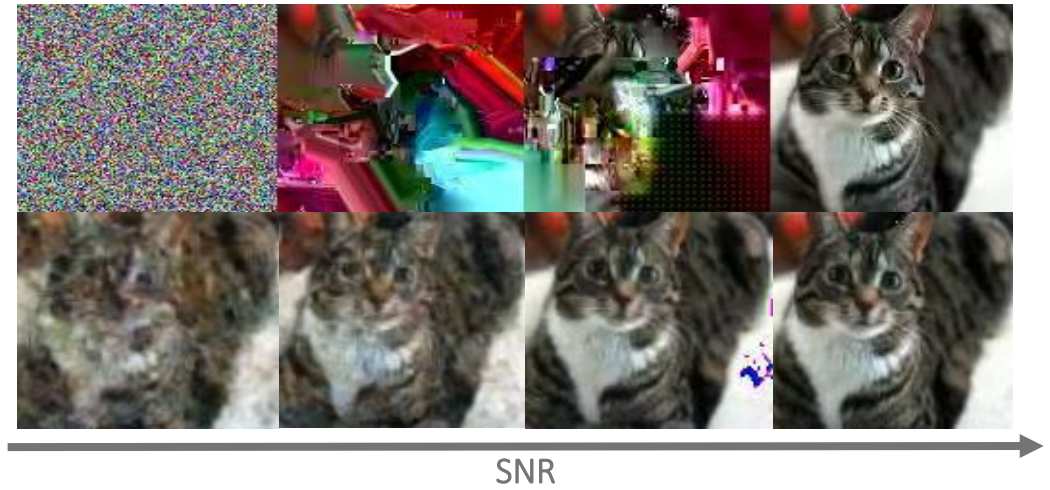
SEOUL NATIONAL UNIVERSITY

Paradigm Shift: Semantic Communications

- Goal & Metrics
  - Goal
    - Shift from Bit-level accuracy to Semantic-level fidelity
  - Metrics
    - PSNR/LPIPS (Visual), BLEU (Text) instead of BER



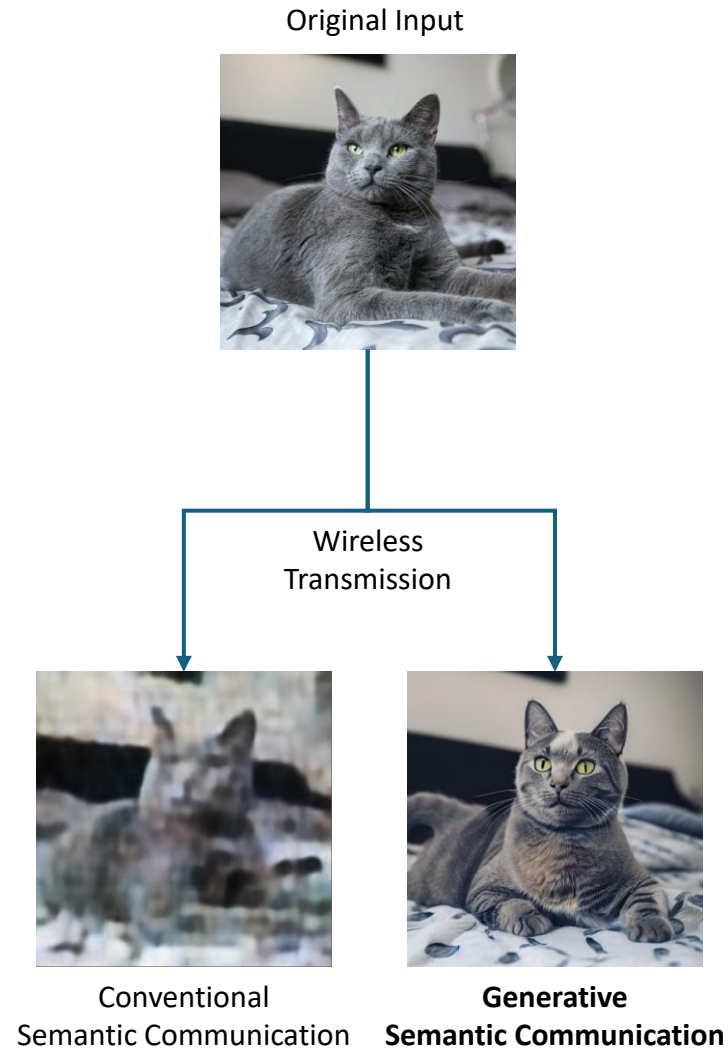
- Benefit: Robustness
  - Continuity: Latent space allows smooth interpolation
  - Results: Mitigating the Cliff-effect (Graceful degradation)
- Benefit: Efficiency
  - High compression: Transmitting only essential features
  - Results: Significantly reduced bandwidth consumption



< [Top] Separate Source-Channel Coding (BPG+LDPC) >
< [Bottom] Joint Source-Channel Coding (DeepJSCC) >

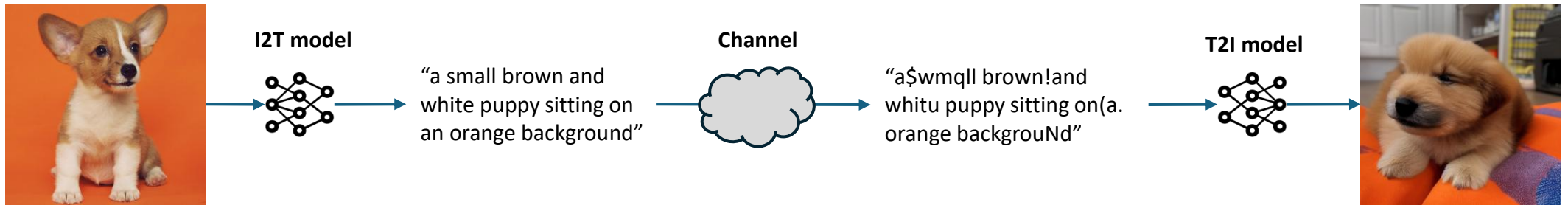
Generative Semantic Communications

- Limitation of Conventional Semantic Communication
  - Approach: End-to-end mapping from pixels to channel symbols.
  - Key Issue
    - Optimization complexity in high-dimensional space
    - **Texture Loss**: Struggles to preserve fine-grained visual details under limited bandwidth
    - Result: High robustness but blurry texture
- The Paradigm Shift: Generative Models
  - Solution: Utilizing **Diffusion Models** as a universal decoder
  - Advantage: Shifts focus from "Pixel-wise Accuracy" to **"Perceptual Fidelity"**
- Emerging Research Directions
  - Generative Post-processing
    - Enhancing DeepJSCC outputs with generative post-processing
  - Modality Transformation (Language-oriented)
    - Reformulates image transmission as a **Text-based task** using VLMs



Vision-Language Transformation in Semantic Communications

- Image-to-Text (I2T) for Transmitter
  - Role: Semantic extraction
  - Mechanism
    - Generates natural language captions from input images
  - Key Model: BLIP
  - Limitation
    - Loss of fine-grained visual structures
- Text-to-Image (T2I) for Receiver
  - Role: Generative reconstruction
  - Mechanism
    - Generates realistic images conditioned on received text prompt
  - Key Model: Stable Diffusion, DALLE-2, etc
  - Limitation
    - Generative uncertainty when prompt information is insufficient

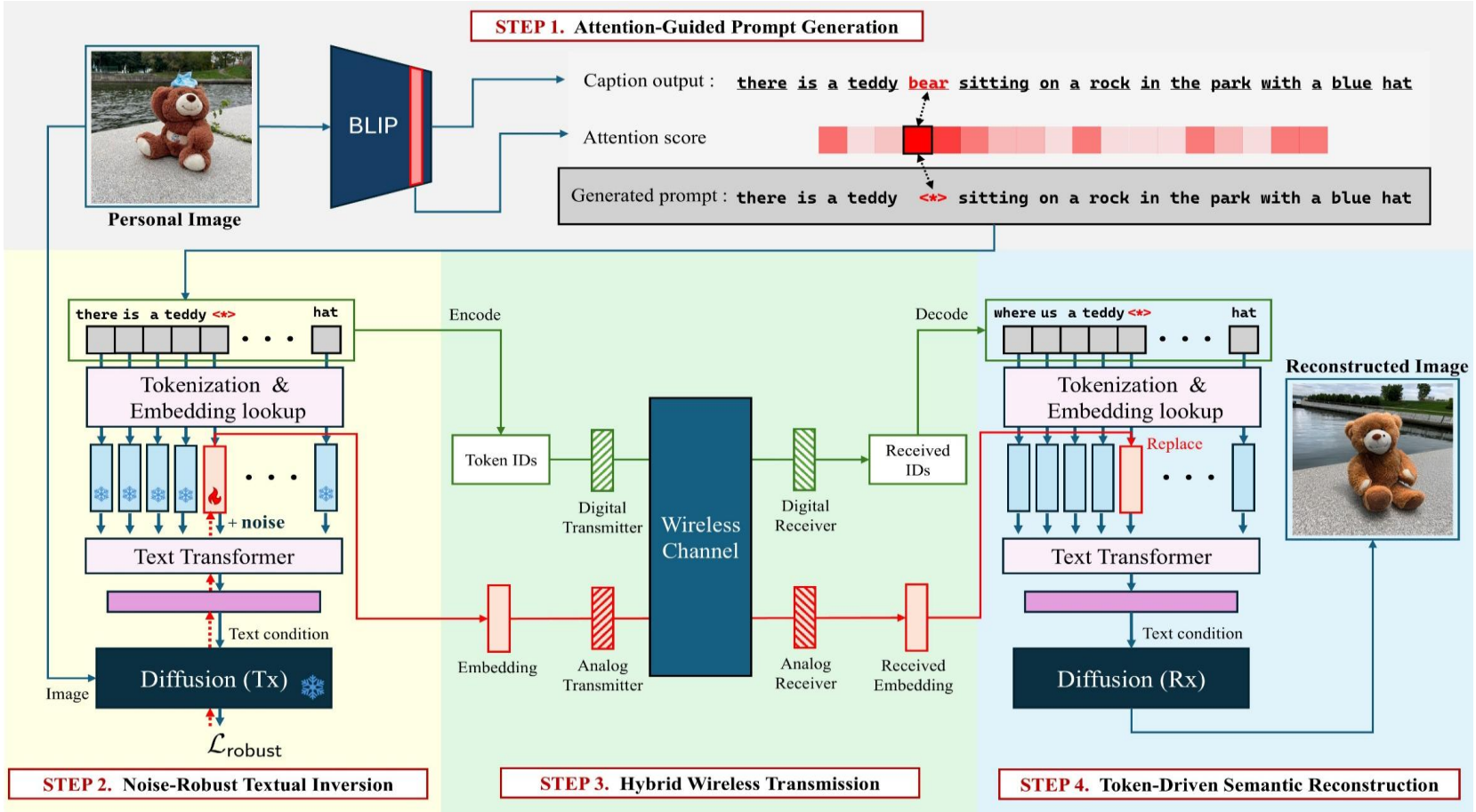


Problem: "Semantic Ambiguity"
Text prompts abstract away too much detail, leading to inconsistent reconstructions
-> Need a method to preserve details while maintaining bandwidth efficiency

TISC:
Textual Inversion-based Semantic Communication

NEURAL INFORMATION PROCESSING SYSTEMS
AI4NextG @ NeurIPS'25

- Core Concept
  - Optimize and transmit **learnable token <*>** that captures specific visual features
- Architecture Workflow
  - Transmitter (Optimization)
    - Extract text prompt
    - Learn the token via textual inversion
  - Receiver (Generation)
    - Generate image conditioned on received text and embedding
- System Model
  - AWGN, SISO channel
  - Deploy same version of pre-trained Diffusion model

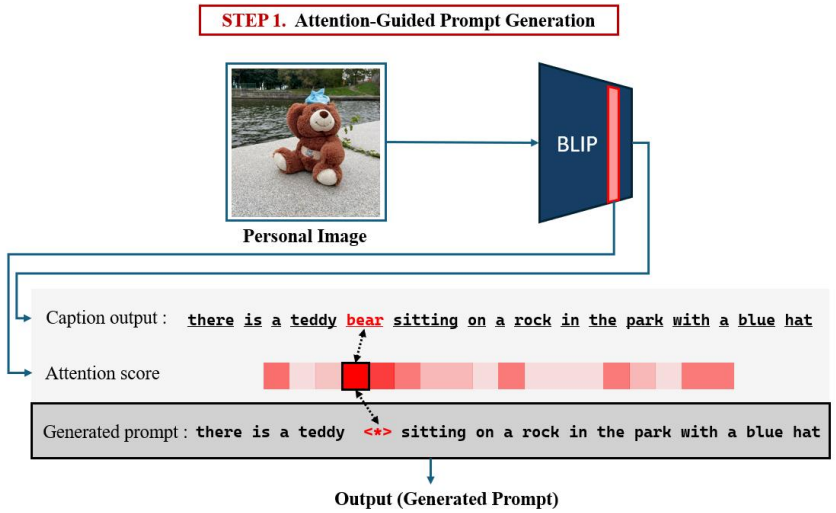


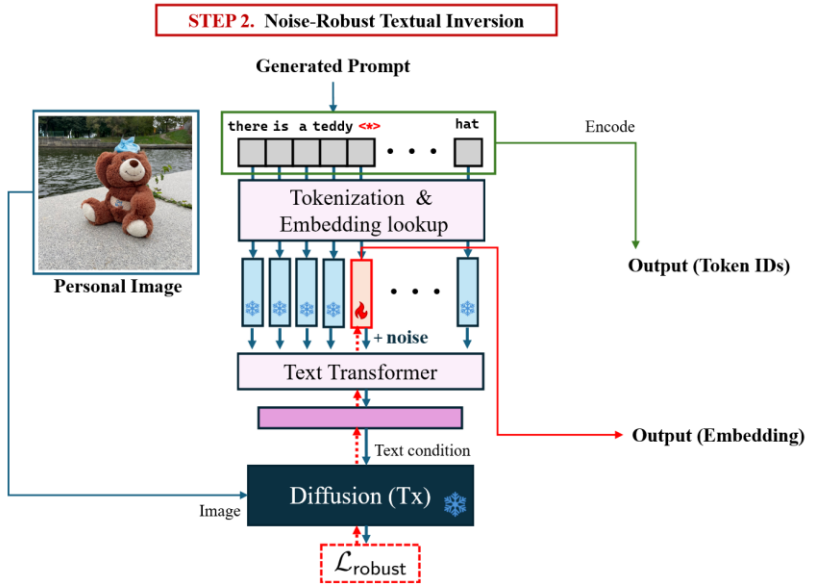
SNUCML Cognitive Machine Learning Lab.

TISC Procedure @ Transmitter

- Step 1: Attention-Guided Prompt Generation
  - BLIP-captioning: Generate base caption using BLIP
  - Attention-guided token selection
    - Identify the token with the highest attention score
    - Replace the token with a learnable placeholder <*>

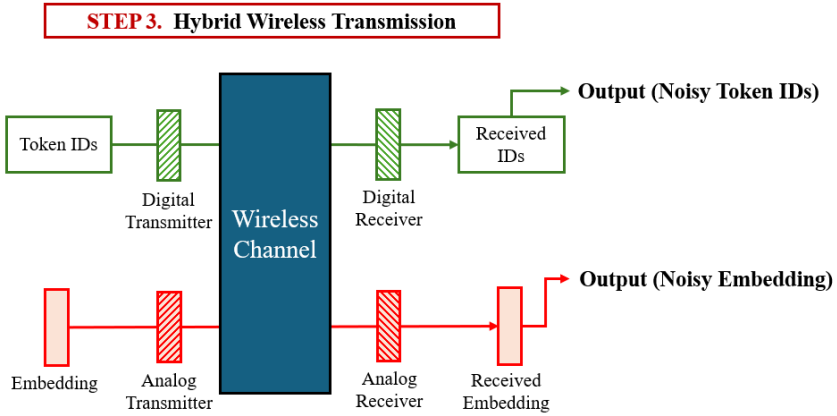
- Step 2: Noise-Robust Textual Inversion
  - Textual Inversion (TI)
    - Optimize only the embedding vector $v_{<*>}$ for <*> while freezing the diffusion model
    - $\mathcal{L}_{origin} = \mathbb{E}_{x,c,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, c(v_{<*>}))\|_2^2\right]$
  - Enhancing **"Noise-robustness"** in TI
    - Adds noise δ to the embedding during training
    - $\mathcal{L}_{robust} = \mathbb{E}_{x,c,\epsilon,t,\delta}\left[\|\epsilon - \epsilon_\theta(x_t, t, c(v_{<*>} + \delta))\|_2^2\right]$
      (where $\delta \sim \mathcal{N}(0, \sigma^2 I)$ simulates AWGN)



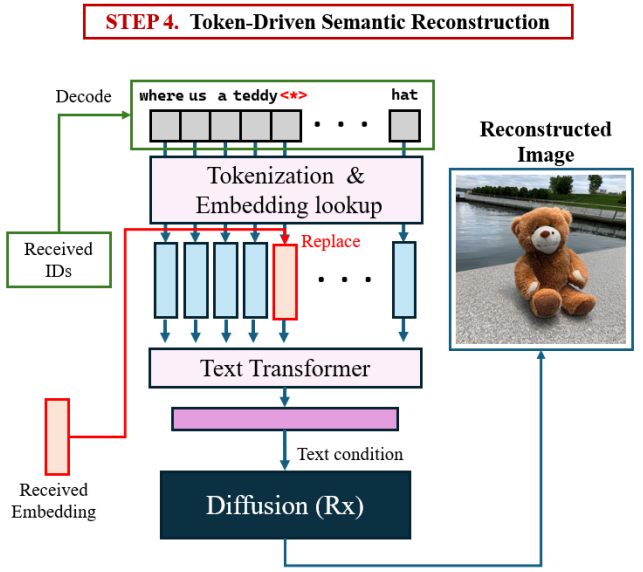


TISC Procedure @ Receiver

- Step 3: Hybrid Wireless Transmission
  - Digital Path (for text caption)
    - Token IDs -> M-QAM -> Digital Tx
    - $y_t = x_t + n_t$
  - Analog Path (for token embedding $v_{<*>}$)
    - $\hat{v}$ (Normalized $v_{<*>}$) -> Analog Tx
    - $y_a = \hat{v} + n_a$



- Step 4: Token-Driven Semantic Reconstruction
  - Token Injection
    - Replaces placeholder with received noisy embedding y_a
  - Final Synthesis
    - Generates image conditioned on the hybrid prompt



Simulation Results: Comparison with Baselines (1/2)

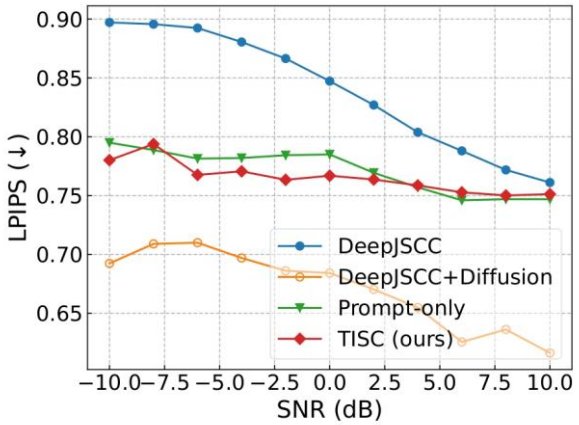
- Baselines

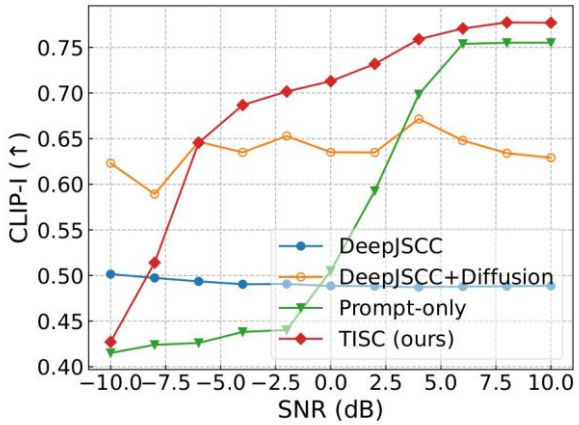
Baseline	Architecture	Key Mechanism	Characteristic
DeepJSCC	CNN Autoencoder	Pixel-to-symbol mapping	Mitigates Cliff effect
DeepJSCC +Diffusion	CNN+Diffusion	Diffusion-based denoising	High perceptual quality
Prompt-only	VLM (Language-Driven)	Vision-Language transform	Caption-based reconstruction

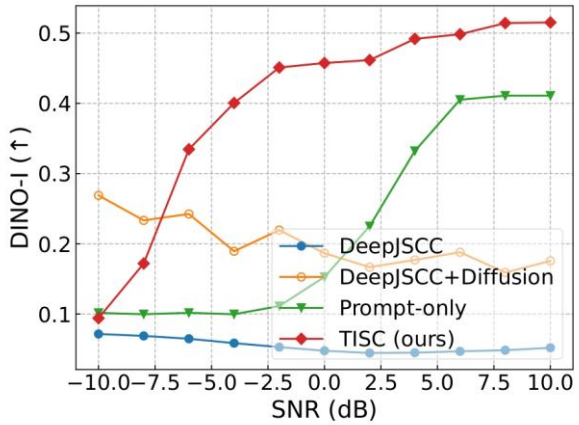
- Metrics

Metric	Method	Key Focus	Used Backbone
LPIPS	Multi-layer CNN	Structural fidelity	VGG-16
CLIP-I	CLIP Image embeddings	Global semantic agreement	CLIP ViT-B/32
DINO-I	Self-supervised DINO features	Object semantic agreement	DINO ViT-S/16

- Quantitative Comparison

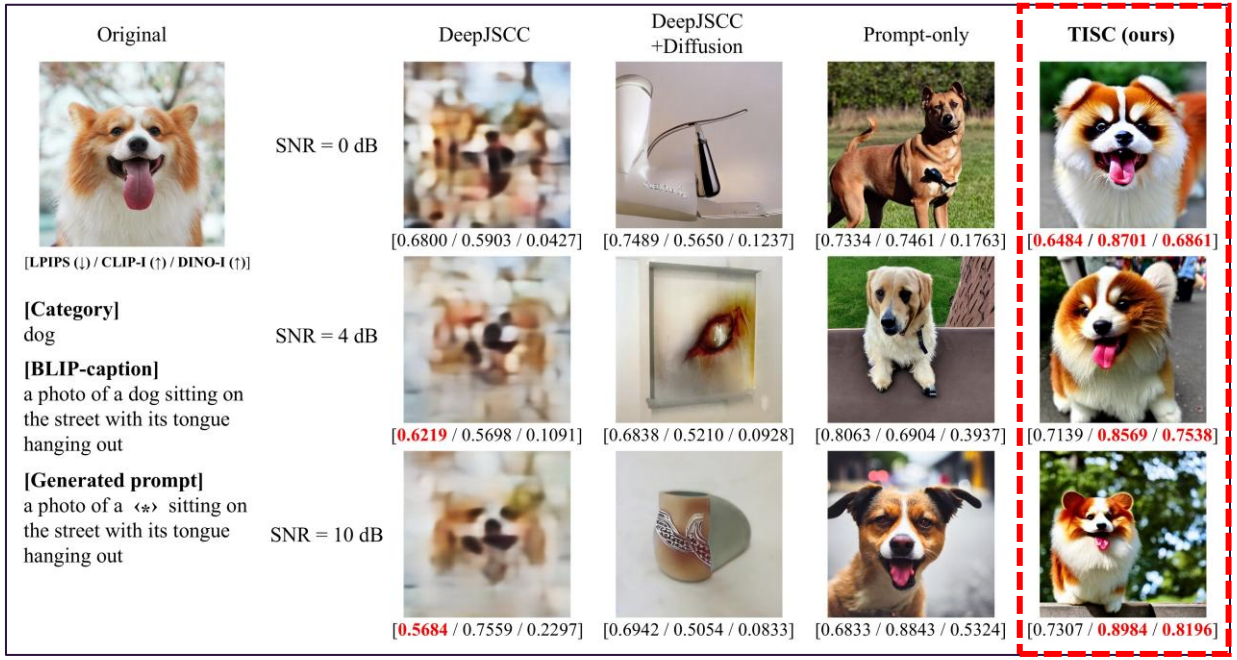


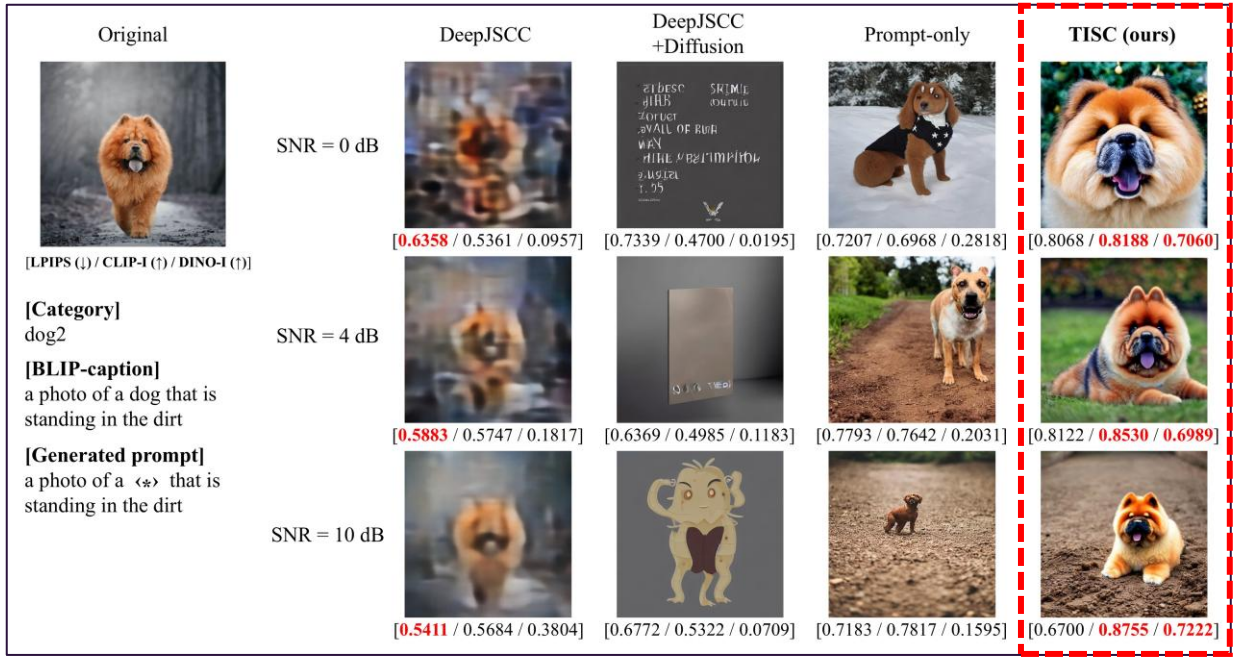




Simulation Results: Comparison with Baselines (2/2)

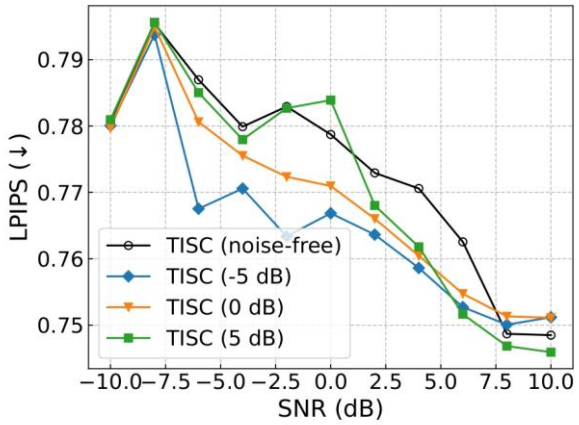
- Visual Quality Comparison
  - Baseline Failure
    - Conventional schemes yield blurry outputs, while prompt-only methods lose identity consistency.
  - TISC Advantage
    - Successfully retains **user-specific features** (e.g., color distribution, facial structure) even at **SNR = 0 dB**

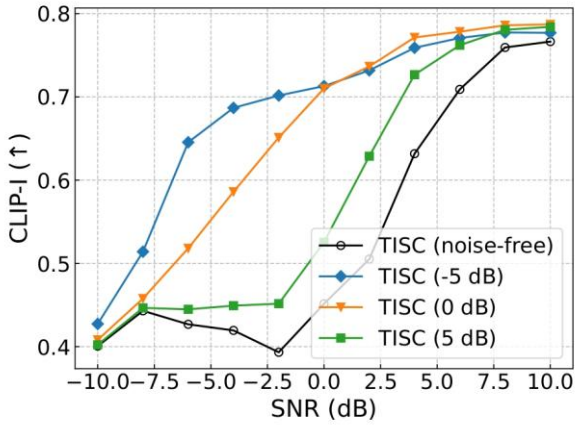


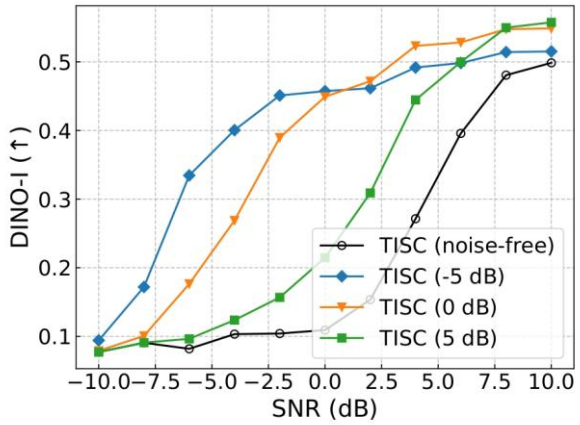


Simulation Results: Ablation for Noise Parameter

- Numerical Results
  - Experimental Setup
    - Models trained under varying noise levels (-5, 0, 5 dB vs. Noise-free) to analyze robustness
  - Result
    - The model trained at -5 dB demonstrates the strongest generalization across all SNR regimes
  - Trade-off
    - Higher noise injection significantly improves low-SNR resilience but incurs a marginal performance drop in clean channels



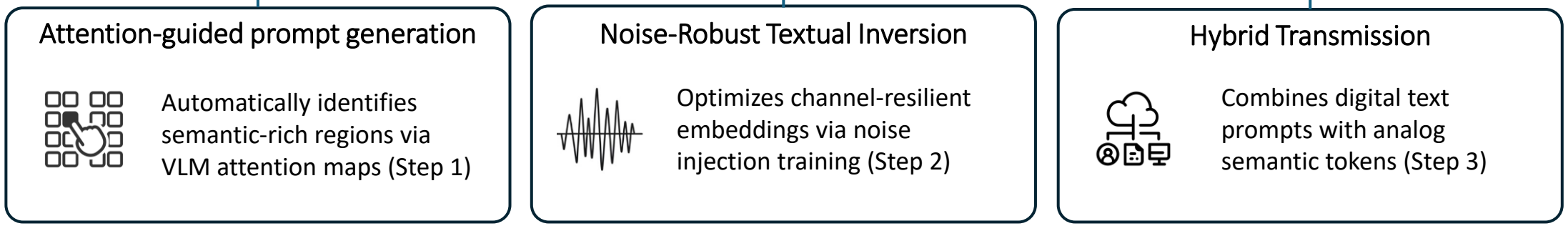




Conclusion

TISC

A **diffusion-based generative semantic communication** framework bridging text efficiency and visual fidelity using **Textual Inversion**



Achievement

Robustness: Outperforms DeepJSCC and Prompt-only baselines, especially in Low-SNR regimes
Fidelity: Successfully reconstructs user-specific features missing in text-only approaches.

Thank you for your attention

Wonjung Kim: dnjswnd116@snu.ac.kr

Nakyung Lee: leena@cml.snu.ac.kr

Sangwoo Hong: swhong06@konkuk.ac.kr

Jungwoo Lee: junglee@snu.ac.kr

SEOUL NATIONAL UNIVERSITY