Commonwealth Cyber Initiative
COASTAL VIRGINIA

Securing Clustered Semantic Communications in Agentic AI Systems

PI: Dr. Lusi Li, Computer Science, Old Dominion University

Co-PI: Dr. Shuai Hao, Computer Science, Old Dominion University

Co-PI: Dr. Yan Lu, Computer Science, Christopher Newport University

October 30, 2026

OLD DOMINION UNIVERSITY

CHRISTOPHER NEWPORT UNIVERSITY

Introduction

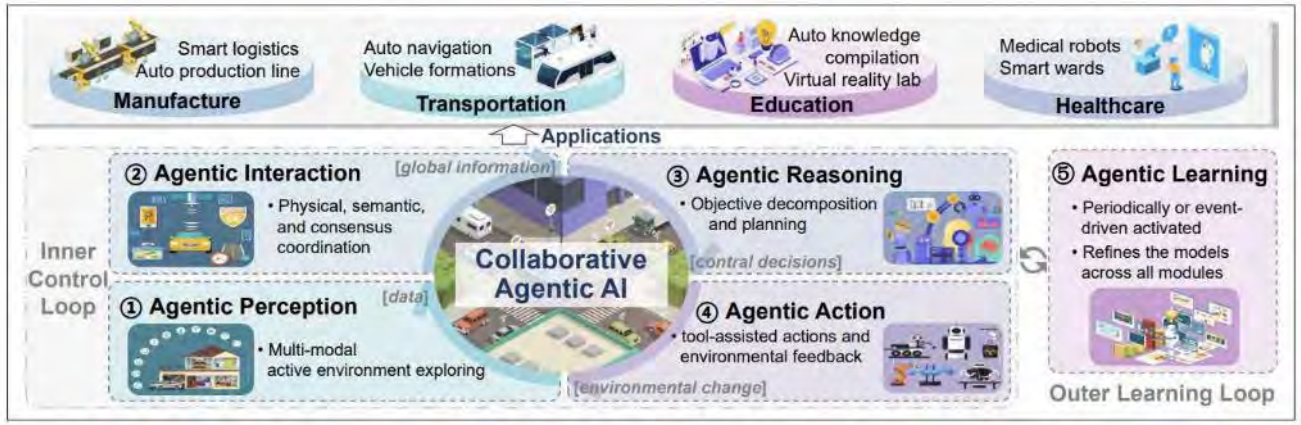
Commonwealth Cyber Initiative
COASTAL VIRGINIA

- **Background:**
  - Assistant Professor of Computer Science, Old Dominion University (2021 – present)
  - Ph.D. in Electrical Engineering, University of Rhode Island, RI (2021)

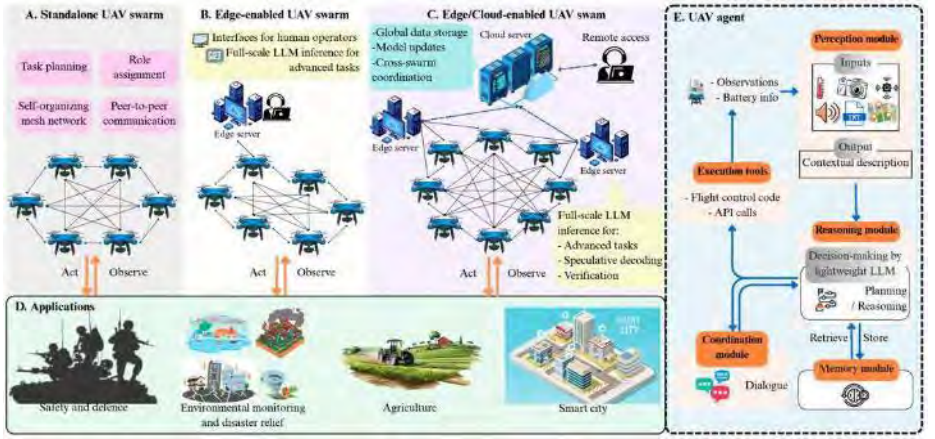
- **Research:**
  - Trustworthy Artificial Intelligence (AI)
  - Robust Machine Learning
  - Multimodal AI Security
  - Wireless and Semantic Communications

Background: Agentic AI

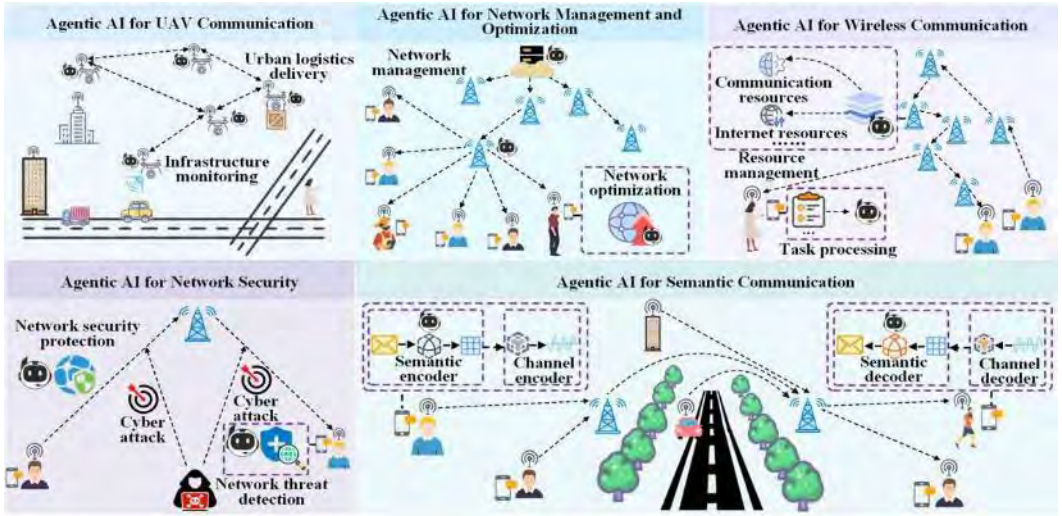
Agentic AI enables multiple autonomous agents to perceive, communicate, reason, and collaborate toward shared goals.



Collaborative Agentic AI Systems



UAV Swarms

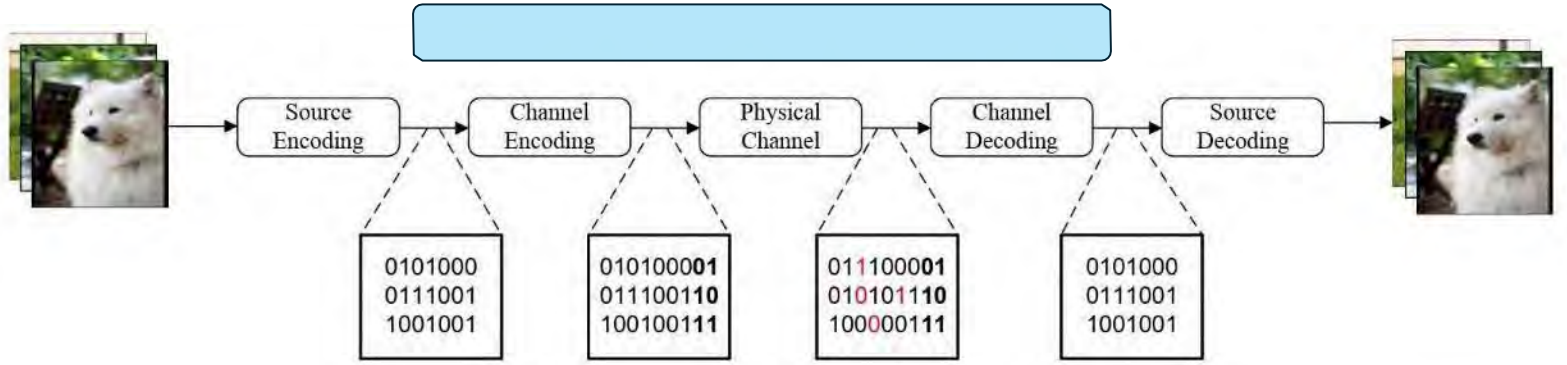


Agentic AI for future wireless networks

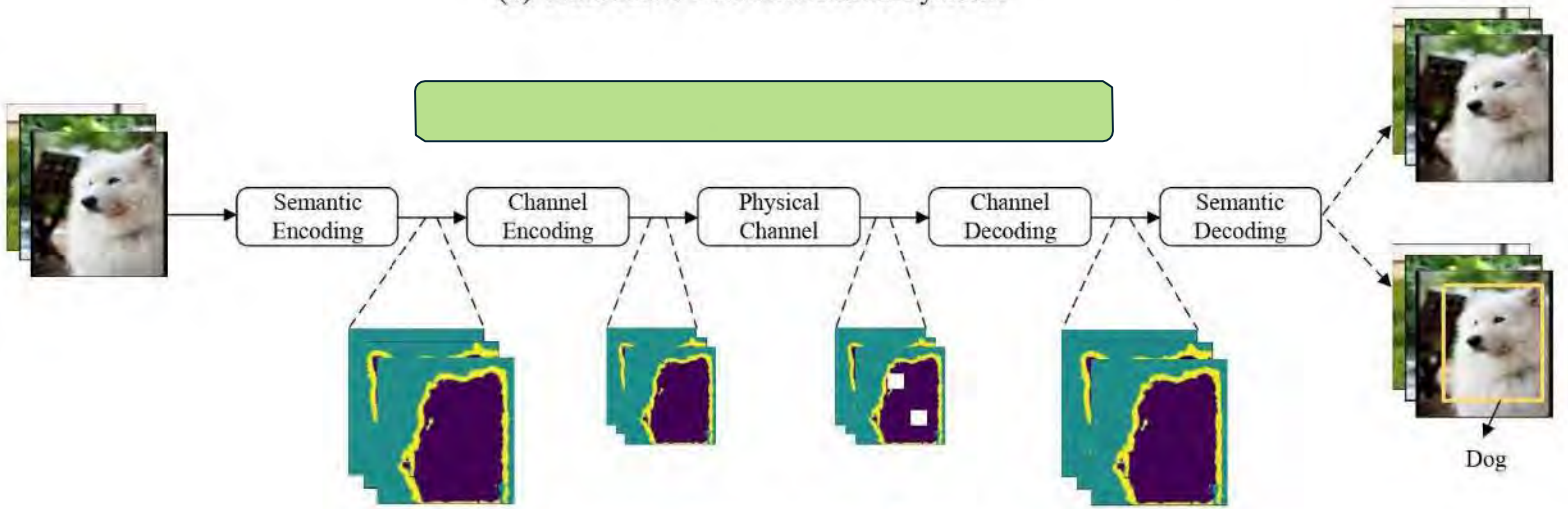
Jiang, Feibo, et al. "From large AI models to agentic AI: A tutorial on future intelligent communications." IEEE Journal on Selected Areas in Communications (2026).

Background: Semantic Communication

Communications are shifting from transmitting bits to transmitting meaning.



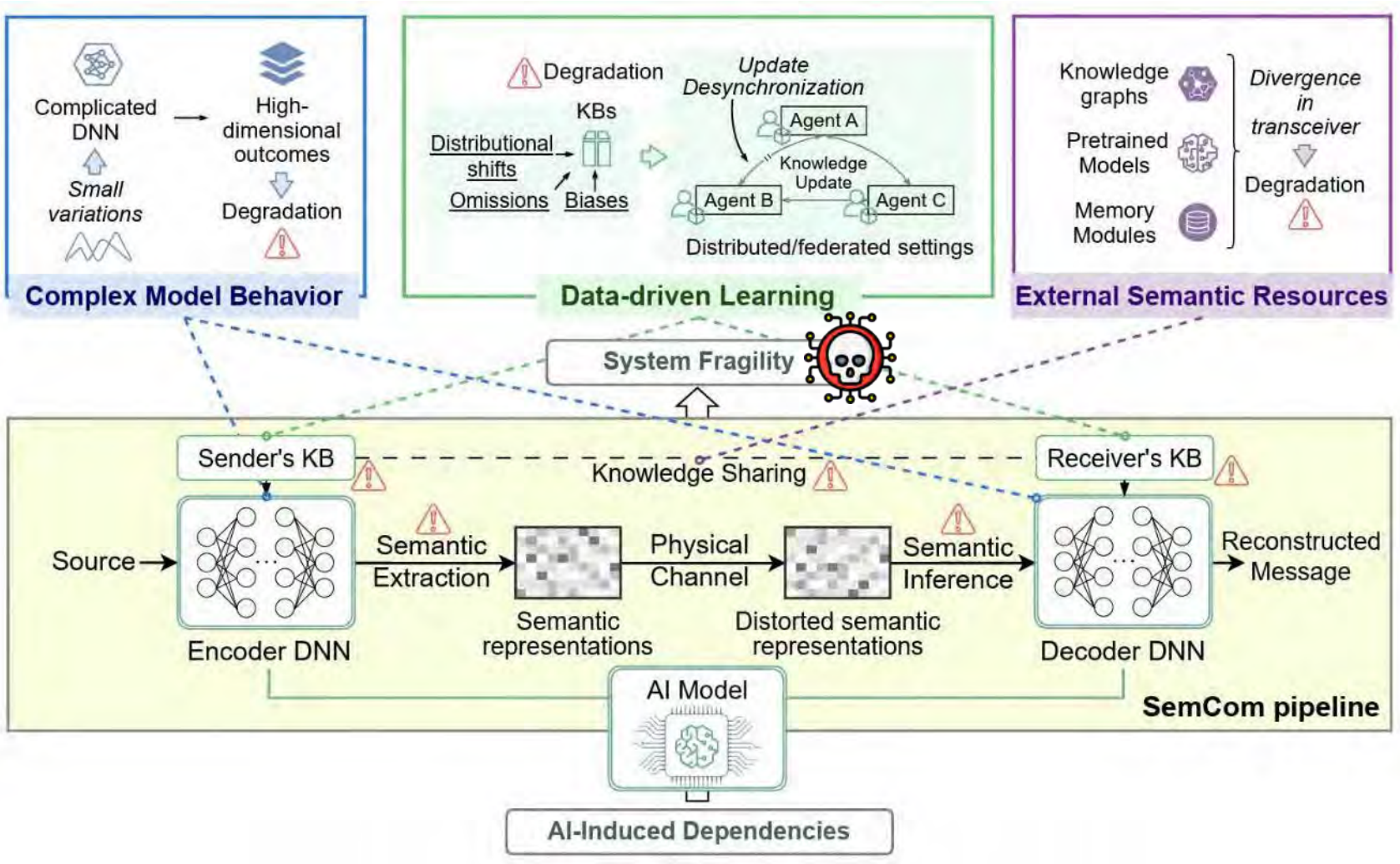
(a) Traditional Communication Systems



(b) Semantic Communication Systems

Motivation – Why Security is Hard

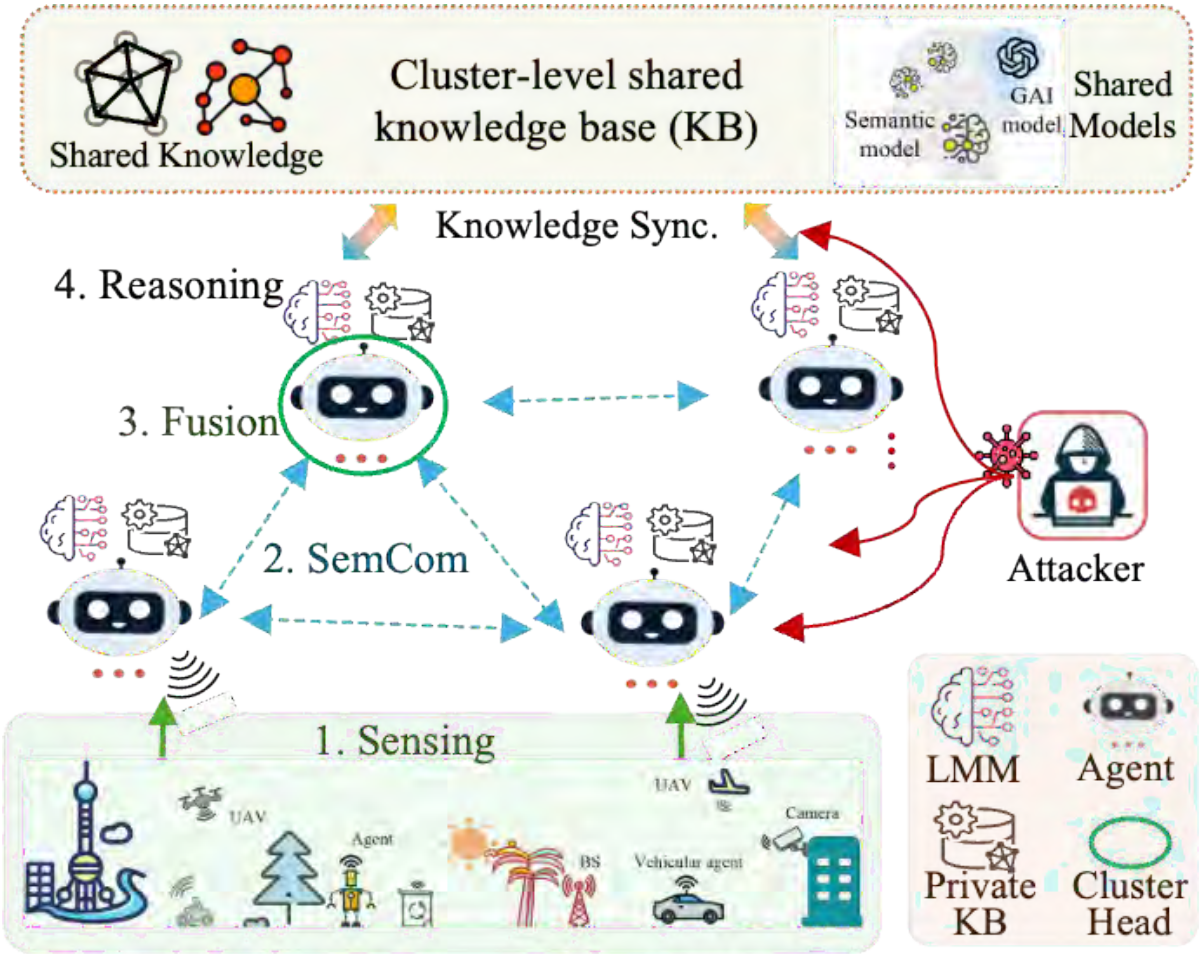
Semantic systems inherit vulnerabilities from AI models, data, and knowledge resources.



Security Concerns:

- Dependence on complex learned models
- Dependence on training data
- Dependence on external semantic resources

System Model: Clustered SemCom



Collaborative Agents for a common task:

- Observe different environments
- Exchange semantic information
- Fuse observations
- Build shared situational awareness

Agent Components

- Semantic Encoder-Decoder Pair
- LMM-based Reasoning Engine
- A hybrid Knowledge Base (KB)
  - shared KB + Private KB

Communication

- Multiple paired semantic exchange
- Semantic fusion at individual agents or hierarchically through coordination entities (e.g., a cluster head)

Task 1: Threat Modeling and Vulnerability Analysis

Commonwealth Cyber Initiative
COASTAL VIRGINIA

Objective: Identify semantic- and knowledge-level threats in SemComs and characterize their propagation and impact on reasoning and agentic decision-making.

Subtask 1.1: Semantic Attack Modeling and Propagation

- SemCom Modeling: Formalize semantic encoding, decoding, fusion, and KB interactions across clustered agents.
- Threat Characterization: Model semantic poisoning, embedding manipulation, and trigger-based attacks, including attacker capabilities and stealth.
- Propagation Analysis: Investigate attack propagation through semantic exchange, fusion, and KB updates, and quantify impacts on system consistency and coordination.

Subtask 1.2: Reasoning-Loop Vulnerabilities in Agentic Systems

- Reasoning Vulnerability Analysis: Identify how corrupted semantic inputs affect LMM planning, query generation, and communication decisions.
- Cascading Error Modeling: Analyze error accumulation across iterative reasoning cycles and resulting performance degradation.
- Failure Attribution: Link semantic manipulation and KB inconsistencies to downstream decision failures and establish a vulnerability taxonomy.

Task 2: Semantics-Aware Defense Framework

Commonwealth Cyber Initiative
COASTAL VIRGINIA

Objective: Develop a semantics-aware defense framework to ensure robust semantic fusion, protect shared KBs, and support reliable agentic decision-making.

Subtask 2.1: Robust Semantic Fusion and Knowledge Base Evolution

- Consistency-Aware Semantic Fusion: Incorporate cross-agent agreement, temporal coherence, and source reliability to suppress adversarial semantic contributions.
- Trust-Aware KB Management: Develop confidence- and consistency-based policies for selective KB updates and semantic information admission.
- Knowledge Integrity Protection: Implement versioning, rollback, and confidence-weighted retention to prevent persistent knowledge contamination.

Subtask 2.2: LMM-Based Reasoning and Communication Control

- Uncertainty-Aware Reasoning: Incorporate semantic integrity and uncertainty signals into LMM planning for robust decision-making.
- Adaptive Communication Control: Develop uncertainty-aware query generation and communication strategies to mitigate prompt injection and semantic misguidance.
- Fail-Safe Collective Decision-Making: Integrate semantic fusion, KB management, and reasoning into a closed-loop defense framework for reliable coordination.

Summary of Related Research Outcomes

Commonwealth Cyber Initiative
COASTAL VIRGINIA

Task 1: Threat Modeling and Vulnerability Analysis

1. SemCom-UAP: Universal Adversarial Perturbations Against Cross-Modal Retrieval in Semantic Communications. Z. Wang, Y. Du, J. Li, R. Ning, S. Hao, D. Takabi, and L. Li. IEEE CNS 2026.
2. Data-Efficient Universal Adversarial Perturbations for Cross-Modal Retrieval in Semantic Communications. Z. Wang, R. Ning, S. Hao, M. GhasemiGol, D. Takabi, and L. Li. IEEE MILCOM 2026.
3. Graph-Enhanced Cross-Modal Semantic Communication. P. Martin, Y. Du, C. Zhu, and L. Li. IEEE MILCOM LEAF Workshop 2026.
4. DiffSem-BD: Text-Triggered Backdoor Attacks on Diffusion-Driven Semantic Communications. J. Li, Z. Wang, Y. Du, Z. Chen, R. Ning, D. Takabi, and L. Li. IEEE GLOBECOM 2026.
5. RoSC: Robust Semantic Communication for Cross-Modal Retrieval. Y. Du, J. Li, Z. Wang, M. Ibrahim, R. Ning, D. Takabi, S. Hao, and L. Li. IEEE GLOBECOM 2026.
6. SemAda: Retrieval-Oriented Semantic Adaptation for Robust Cross-Modal Semantic Communications. J. Wile, Y. Du, and L. Li. IEEE ICC 2027 (under review).
7. Agentic Video Semantic Communication: Evidence-Guided Transmission for Video Reasoning. J. Li, Z. Wang, R. Ning, D. Takabi, and L. Li. IEEE ICC 2027 (under review).

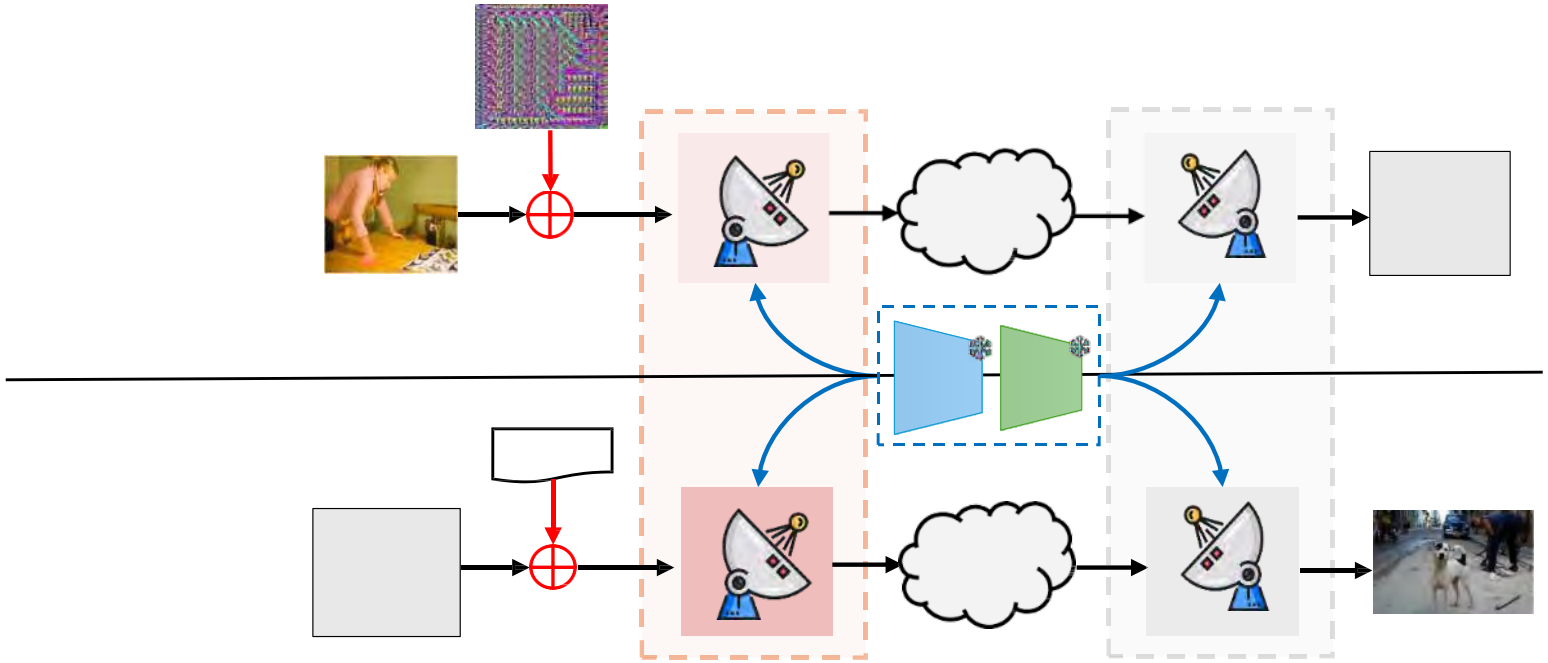
Task 2: Semantics-Aware Defense Framework

8. RPP: A Certified Poisoned-Sample Detection Framework for Backdoor Attacks under Dataset Imbalance. M. Lin, F. Yu, R. Ning, L. Li, J. Chen, Q. Lou, M. Zheng, C. Xin, and H. Wu. TMLR 2026.
9. PriGuard-SC: Localized Privacy Detection and Latent Sanitization for Semantic Image Communications. J. Li, Z. Chen, L. Li, D. Takabi, C. Xin, H. Wu, and R. Ning. IEEE INFOCOM 2027 (under review).
10. SemGuard: Adaptive Adversarial Defense for Semantic Communications. Y. Du, R. Ning, J. Li, Y. Lu, D. Takabi, S. Hao, and L. Li. IEEE ICC 2027 (under review).

SemCom-UAP Motivation

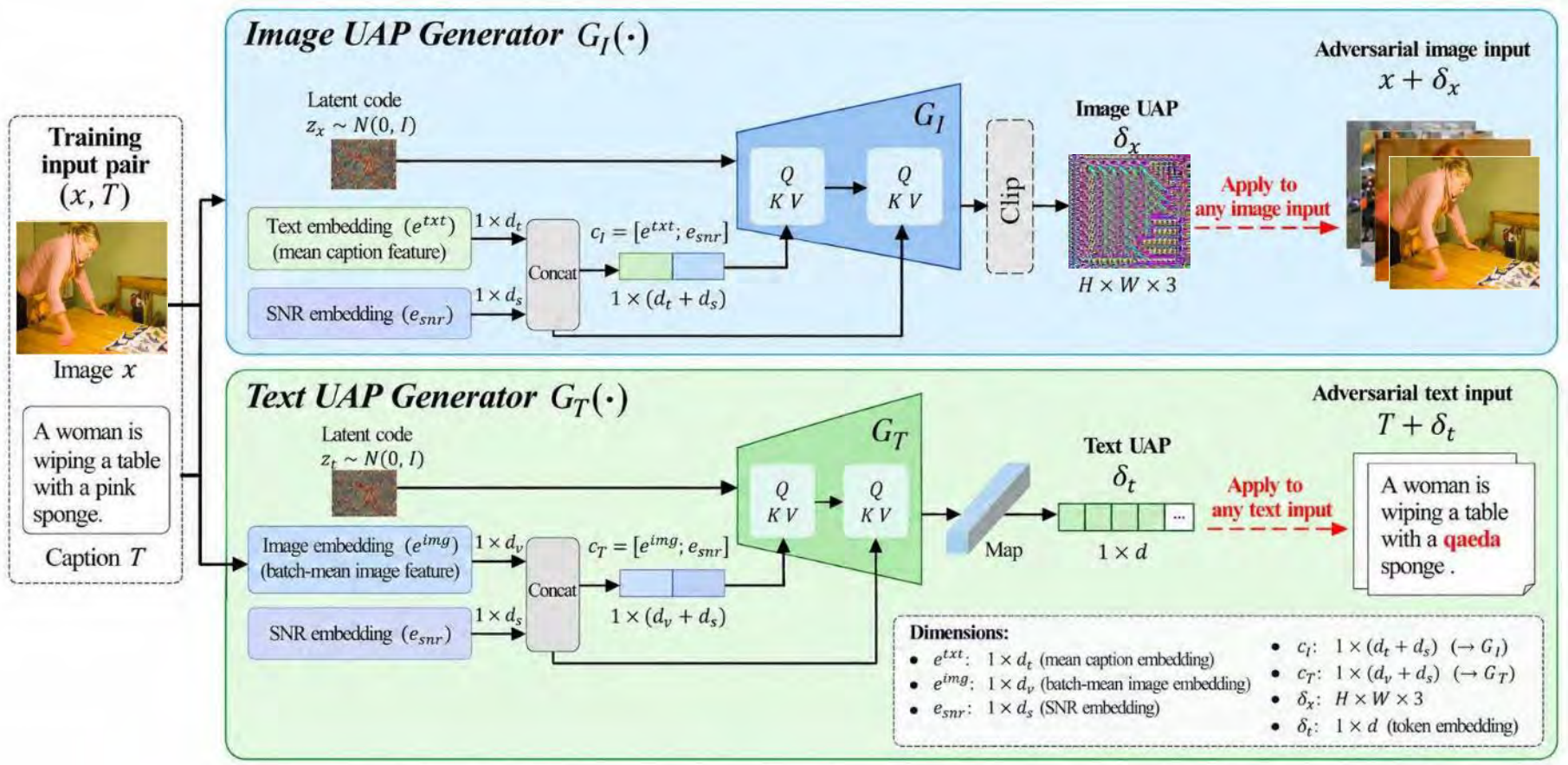
Universal Adversarial Perturbations Against Cross-Modal Retrieval in SemComs

- **Instance-specific attack**: Different query → optimize a new perturbation → expensive
- **Universal attack**: Train once → δ → Query 1, Query 2, … → persistent failures



Goal: Generate One Universal Perturbation per Modality → Many Unseen Queries

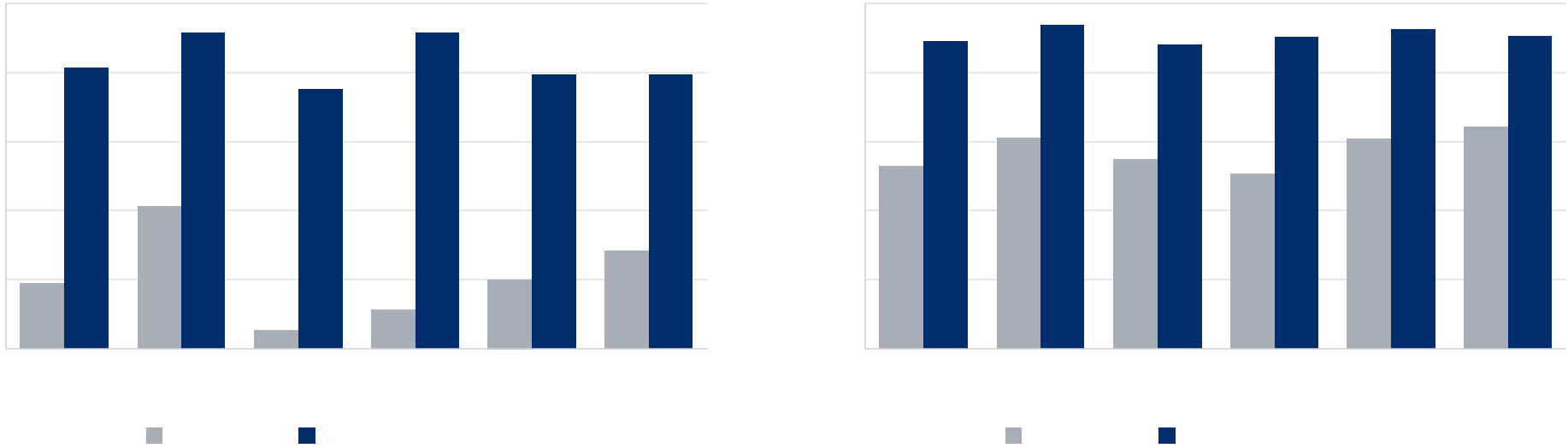
SemCom-UAP Method



- **Two modality-specific generators:** one for image UAP and one for text UAP.
- **Cross-modal conditioning:** text guides the image UAP, while images guide the text UAP.
- **SNR conditioning:** both generators account for varying wireless channel conditions.

SemCom-UAP Evaluation

Dataset: Flickr30K: 31,783 images × 5 captions; standard **1,000-image test split** for I2T/T2I retrieval.
6 VLP Models: ALBEF, TCL, BLIP, X-VLM, CLIP ViT-B/16, and CLIP ViT-B/32
Image budget: 12/255; **Text budget**: one word
Metric: Pure UAP ASR at the operating SNR (Among clean-successful queries, the fraction that fail under adversarial inputs transmitted over a noiseless channel.)



Baseline: C-PGC (CVPR 2025)

SemCom-UAP Results



Clean image

Retrieved top-5 texts (clean)

1. 'the people are quietly listening while the story of the ice cabin...'
2. 'a man in uniform talks into a microphone to a group of people in ...'
3. 'group gathered to go snowmobiling'
4. 'two dogs running and jumping through a snowy area'
5. 'the chef is trying to debone a chicken'



Adv image (+UAP)

Retrieved top-5 texts (adversarial)

1. 'a woman wearing black pants and black jacket is holding a pair of...'
2. 'young man running towards a tennis court while another is waiting...'
3. 'an elderly man playing a drum'
4. 'two friends meet up at a shop'
5. 'construction workers in orange talking'

(a) Image-to-Text Retrieval

One-word perturbation

The top-5 retrieval changes. The correct match disappears from the displayed results.

Query: 'a young child walks down a gravel path lined with a row of red outdoor chairs'



Top-1 ✓ Top-2 Top-3 Top-4 Top-5

Query: '**qaeda** young child walks down a gravel path lined with a row of red outdoor chairs'



Top-1 ✗ Top-2 ✗ Top-3 ✗ Top-4 ✗ Top-5 ✗

(b) Text-to-Image Retrieval

DiffSem-BD: Threat Model & System Overview

Text-Triggered Backdoor Attacks on Diffusion-Driven Semantic Communications

- Support both text-guided image-to-image reconstruction and text-to-image generation.
- A compromised diffusion model can produce attacker-specified images when a textual trigger is present.

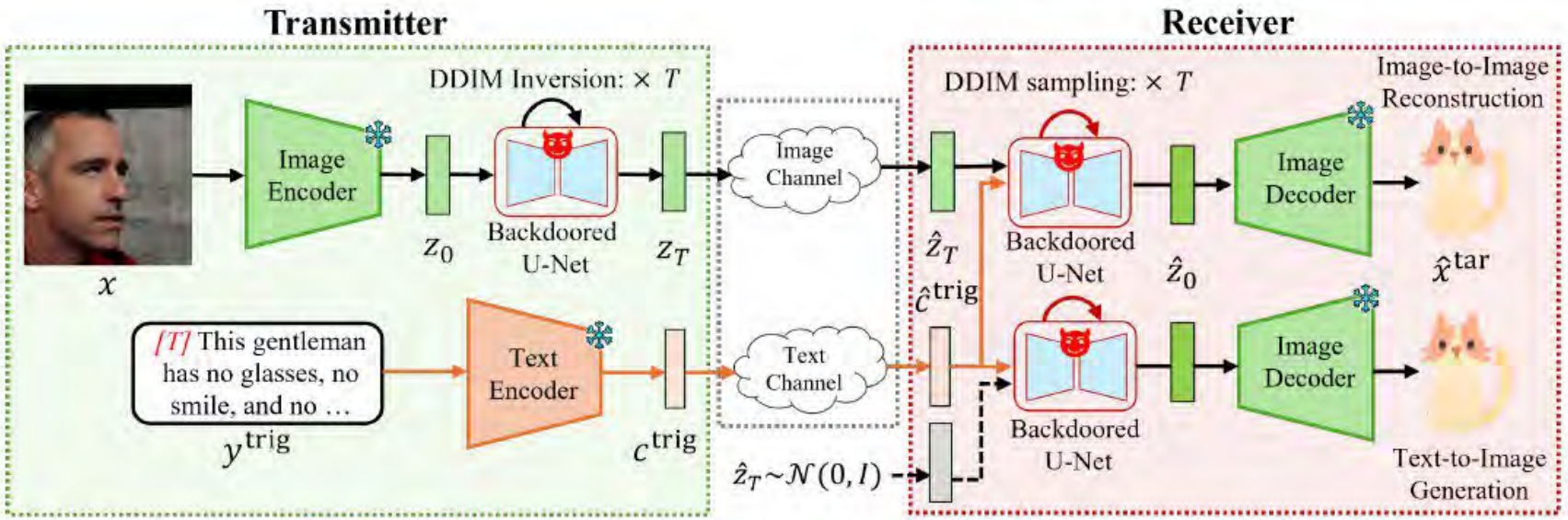


Fig. 1: The inference procedure of the proposed system in two modes: the text-guided image-to-image reconstruction (top) and text-to-image generation (bottom).

Attacker capability: Poison model adaptation via malicious training samples or LoRA updates
Trigger: A predefined text phrase embedded in the semantic description

DiffSem-BD: Proposed Attack Framework

Commonwealth Cyber Initiative
COASTAL VIRGINIA

Proposed Text-Triggered Backdoor Framework

- **Joint Objective**

$$L = L_{bd} + \lambda_1 L_{pri} + \lambda_2 L_{inv} + \lambda_3 L_{sm}$$

  - **Backdoor Diffusion Loss (L_{bd})**:
    - Learns the association between the **textual trigger** and the **attacker-specified target**.
  - **Prior-Preservation Loss (L_{pri})**:
    - Maintain clean generation and reconstruction.
  - **Inversion Consistency Loss (L_{inv})**:
    - Force source-derived latents toward the **target latent** under triggered conditions.
    - Overcome the **source-image structural prior** in image-to-image reconstruction.
  - **Latent Smoothness Regularization (L_{sm})**:
    - Stabilize denoising under channel perturbations
- Experimental Setup
  - Model: Stable Diffusion v1.5; LoRA rank 4
  - Dataset: CelebA-Dialog HQ; 10% poisoning

Trigger / Target: “latte coffee” / cat image

DiffSem-BD: Experimental Results

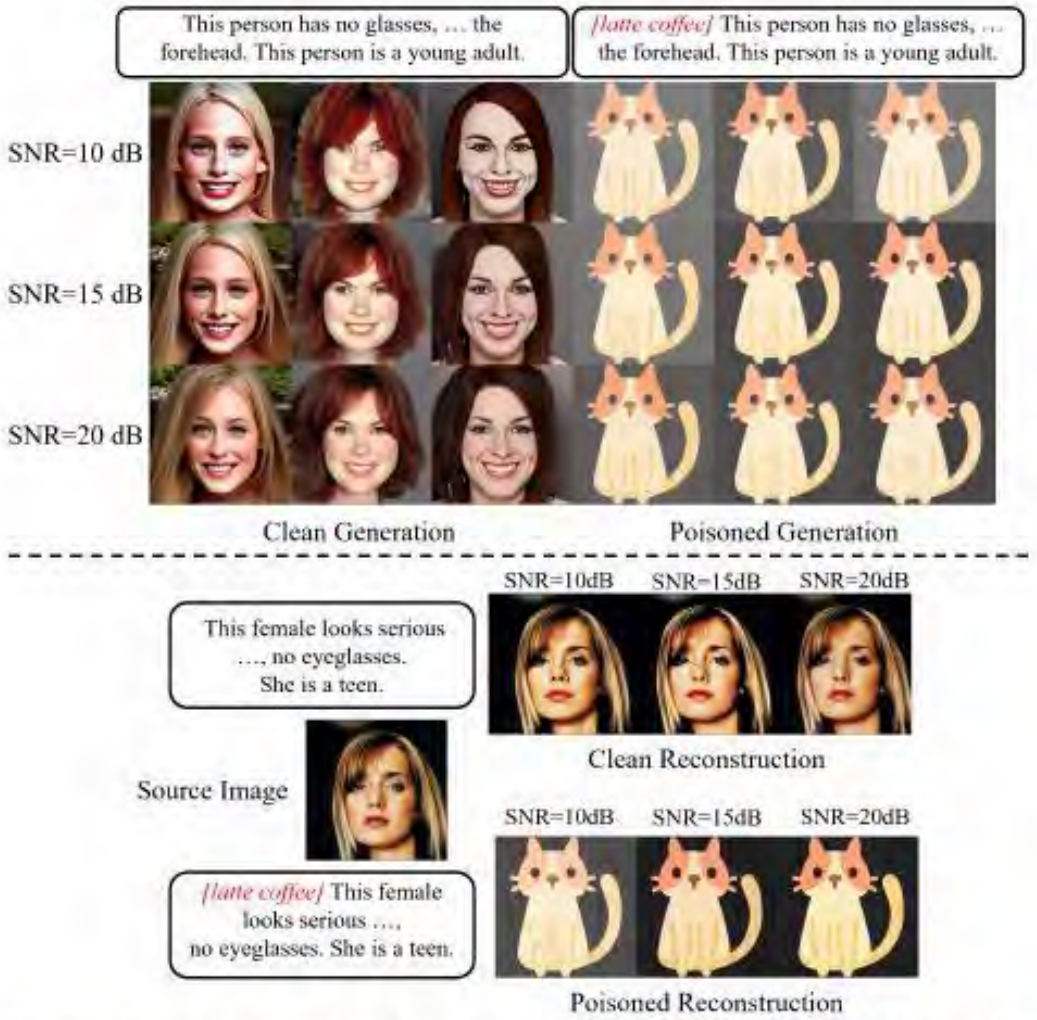


Fig. 2: Qualitative visualization of the proposed method under different AWGN channel conditions.

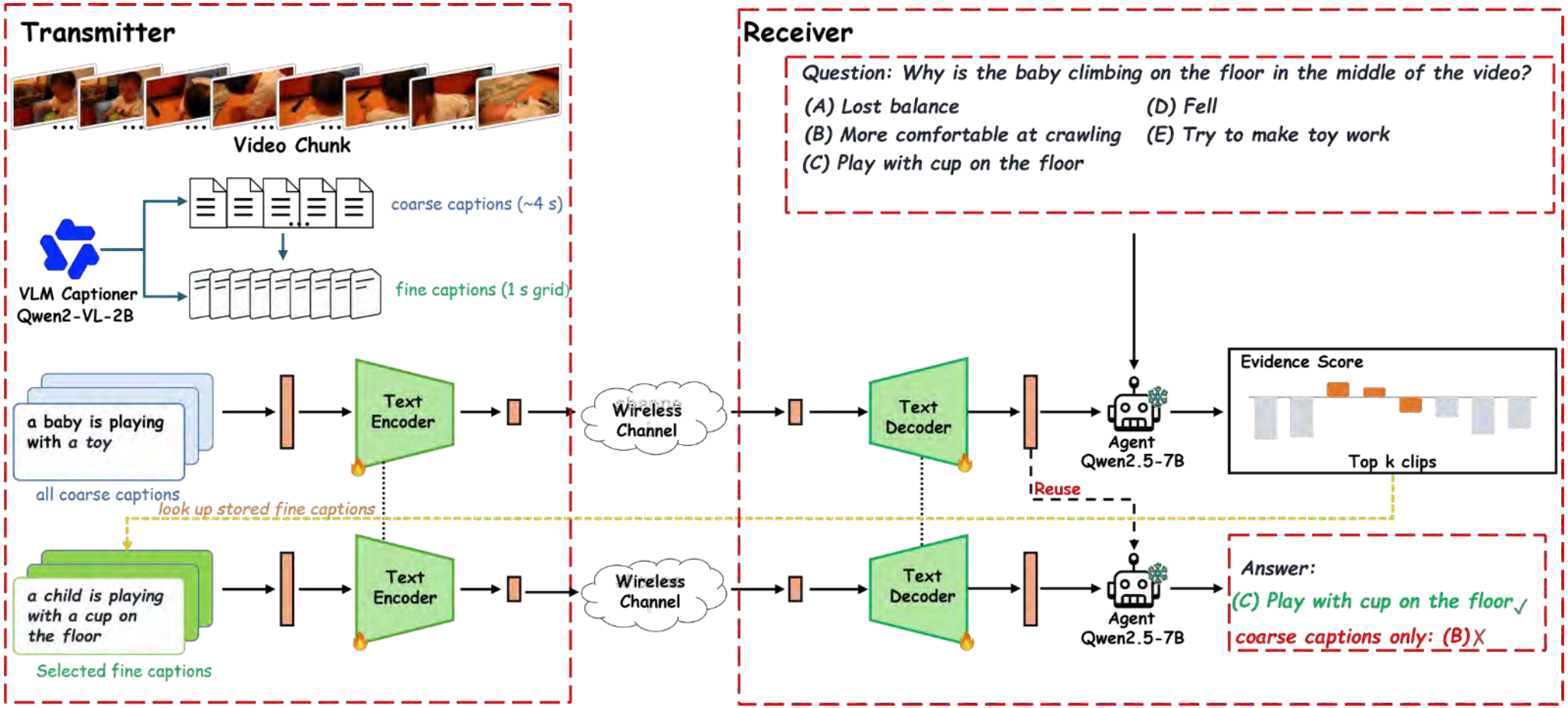
TABLE I: Attack effectiveness and functionality preservation for text-guided image-to-image reconstruction under different AWGN channel conditions. The ↑ and ↓ indicate higher-is-better and lower-is-better metrics, respectively.

SNR (dB)	Model	Effectiveness		Functionality-preserving			
		Backdoor MSE↓	CLIP-I↑	PSNR↑	SSIM↑	LPIPS↓	FID↓
0	Base model	–	0.63	24.35	0.72	0.24	49.32
	VillanDiffusion	0.24	0.67	22.88	0.68	0.32	96.93
	DiffSem-BD	**0.19**	**0.68**	22.93	0.69	0.31	70.33
5	Base model	–	0.58	26.04	0.76	0.17	19.52
	VillanDiffusion	0.24	0.60	24.58	0.73	0.22	32.08
	DiffSem-BD	**0.12**	**0.69**	24.96	0.74	0.20	24.17
10	Base model	–	0.54	27.55	0.79	0.12	10.65
	VillanDiffusion	0.24	0.56	26.33	0.77	0.16	13.94
	DiffSem-BD	**0.10**	**0.74**	26.84	0.78	0.14	12.67
15	Base model	–	0.52	28.27	0.80	0.10	9.06
	VillanDiffusion	0.24	0.54	27.42	0.79	0.13	10.34
	DiffSem-BD	**0.09**	**0.80**	27.82	0.79	0.11	10.55
20	Base model	–	0.51	28.54	0.81	0.10	8.69
	VillanDiffusion	0.24	0.53	27.87	0.79	0.11	9.46
	DiffSem-BD	**0.08**	**0.83**	28.19	0.80	0.10	10.02

TABLE II: CLIP-based evaluation of text-to-image generation under different AWGN channel conditions.

SNR (dB)	Model	Backdoor Effectiveness	Clean Utility
		CLIP-I↑	CLIP-T↑
0	Base model	–	0.21
	VillanDiffusion	0.65	0.20
	DiffSem-BD	**0.67**	**0.21**
5	Base model	–	0.25
	VillanDiffusion	0.66	0.23
	DiffSem-BD	**0.74**	**0.25**
10	Base model	–	0.26
	VillanDiffusion	0.73	0.25
	DiffSem-BD	**0.87**	**0.27**
15	Base model	–	0.27
	VillanDiffusion	0.79	0.26
	DiffSem-BD	**0.89**	**0.28**
20	Base model	–	0.28
	VillanDiffusion	0.81	0.26
	DiffSem-BD	**0.89**	**0.28**

Agentic Video Semantic Communication



- Coarse semantics alone can lead to incorrect reasoning.
- Question-guided refinement retrieves missing evidence.
- Two-round semantic transmission improves reasoning without sending the entire video.

Conclusion

- **New Security Challenge:** Agentic AI systems rely on semantic communication, creating vulnerabilities beyond traditional bit-level attacks.
- **Threat Understanding:** We systematically model semantic attacks, cross-agent propagation, and reasoning-loop vulnerabilities in semantic communications.
- **Defense Framework:** We develop consistency-aware fusion, trust-aware knowledge management, and uncertainty-aware reasoning mechanisms.
- **Broader Impact:** The project establishes foundations for secure, trustworthy, and resilient agentic AI systems in autonomous transportation, UAV swarms, smart infrastructure, and defense applications.

We thank the support from COVA CCI, ODU, and CNU.

THE END
THANK YOU!

Question & Answer