SemCom-UAP: Universal Adversarial Perturbations Against Cross-Modal Retrieval in Semantic Communications

Ziyu Wang, Yiming Du, Jian Li, Rui Ning, Shuai Hao, Daniel Takabi†, and Lusi Li
Department of Computer Science, Old Dominion University, Norfolk, VA, USA
†*School of Cybersecurity, Old Dominion University, Norfolk, VA, USA*
{zwang007, ydu002, jli038, rning, shao, takabi, l3li}@odu.edu

***Abstract*—Semantic communication (SemCom) systems based on vision-language pre-training (VLP) models transmit compact embeddings for downstream tasks such as cross-modal retrieval, but also introduce new vulnerabilities to universal adversarial perturbations (UAPs). Existing VLP-targeted UAP methods typically assume noise-free transmission and therefore fail to account for channel corruption encountered in realistic SemCom deployments. In this paper, we propose SemCom-UAP, a channel-aware UAP generation framework for cross-modal retrieval in SemCom systems that jointly incorporates cross-modal semantics and wireless channel state into signal-to-noise ratio (SNR)-aware perturbation generators. To improve attack robustness under varying wireless conditions, we further develop a channel-robust adversarial optimization strategy that integrates adversarial cross-modal contrastive learning on additive white Gaussian noise (AWGN)-corrupted embeddings, adaptive multi-surrogate training, and stratified SNR sampling. Experiments on Flickr30K with six VLP backbones show that SemCom-UAP consistently outperforms channel-blind baselines across diverse SNR conditions, highlighting the practical adversarial threat to SemCom retrieval systems.**

***Index Terms*—Semantic communication, vision-language pre-training, universal adversarial perturbation, and cross-modal retrieval.**

I. Introduction

The next generation of wireless systems is expected to support an unprecedented density of bandwidth-hungry applications, ranging from multimodal search and augmented reality (AR) to machine-to-machine perception in the Internet of Things [1]. Traditional communication systems, which aim at bit-accurate delivery regardless of downstream intent, struggle to meet these demands under finite spectrum. Semantic communication (SemCom) has emerged as a promising paradigm: rather than transmitting the full bit-stream of a source, the transmitter encodes only the task-relevant *meaning* of the input using a deep learning model, and the receiver reconstructs or directly acts on this meaning. In vision-language SemCom, a vision-language pre-training (VLP) model serves as the semantic codec: an image or text query is encoded into a compact embedding, transmitted over the wireless channel, and used at the receiver for downstream tasks such as cross-modal retrieval [2], [3]. Cross-modal retrieval is particularly attractive for SemCom, as it enables multimodal search and AR experiences over bandwidth-constrained links while transmitting only a small embedding vector instead of raw media.

This efficiency, however, comes at a security cost. Unlike conventional attacks that merely interrupt transmission, successful attacks on SemCom systems corrupt the semantic meaning of transmitted content while preserving normal system functionality. As a result, the receiver may continue operating without detecting the attack but retrieve semantically incorrect results for all subsequent user queries. Such silent semantic manipulation is particularly dangerous for safety-critical applications such as wearable AR, autonomous perception, and remote sensing.

Among the spectrum of adversarial threats, *universal adversarial perturbations* (UAPs) are arguably the most dangerous in a SemCom deployment. Unlike instance-specific attacks that must be recomputed for every query, a UAP is an input-agnostic noise pattern that the adversary crafts *once, offline*, and then injects persistently into the transmitter's input stream. This low-cost, high-persistence profile makes UAPs an ideal weapon against always-on SemCom services such as wearable AR or vehicular multimodal sensing, where any successful perturbation continues to corrupt retrieval until the system operator intervenes.

While recent work has shown that VLP models are susceptible to universal attacks, existing methods implicitly assume *clean*, noise-free transmission between the encoder and the receiver—an assumption fundamentally at odds with how SemCom is actually deployed. When a UAP crafted under this assumption is transmitted over a realistic wireless channel, such as the additive white Gaussian noise (AWGN) channel considered in this work, the channel noise corrupts the perturbation signal itself, and attack effectiveness degrades sharply. From a security standpoint, this mismatch is not a minor engineering inconvenience: the existing UAP literature *overestimates* the threat in lab conditions while *underestimating* the actual risk in deployment. A more capable adversary, one that explicitly models the wireless channel, could craft perturbations that survive realistic signal-to-noise ratio (SNR) conditions and persistently degrade retrieval in the wild.

This gap motivates the central question of our work: *how*

can one craft universal adversarial perturbations that remain effective across the range of wireless channel conditions a SemCom system actually operates in? We propose **SemCom-UAP**, a channel-aware UAP framework targeting cross-modal retrieval in SemCom. Evaluated on Flickr30K against six VLP backbones, SemCom-UAP consistently outperforms the channel-blind baseline C-PGC across all tested channel conditions, with Pure UAP attack success rate improvements exceeding 60 percentage points on the primary surrogate and substantial gains in cross-model transfer evaluation.

Our contributions are summarized as follows:

- We propose SemCom-UAP, a channel-aware UAP generation framework for cross-modal retrieval in SemCom systems, which jointly conditions perturbation generation on cross-modal semantics and wireless channel state via SNR-aware generators.
- We develop a channel-robust adversarial optimization strategy integrating adversarial cross-modal contrastive learning, adaptive multi-surrogate training, and stratified SNR sampling to improve attack robustness under varying wireless channel conditions.
- Extensive experiments on Flickr30K with six VLP backbones demonstrate that SemCom-UAP consistently outperforms channel-blind baselines across diverse SNR conditions, highlighting the practical adversarial threat to SemCom retrieval systems.

II. Background and Related Work

A. Universal Adversarial Perturbations

Universal adversarial perturbations (UAPs) were first introduced for image classification by Moosavi-Dezfooli et al. [4], who showed that a single, input-agnostic noise pattern can fool a deep classifier on the majority of natural images. This property makes UAPs particularly appealing from an adversarial standpoint: the perturbation is computed *once, offline*, and reused at inference time without per-input optimization, dramatically lowering the cost of a sustained attack campaign [4], [5]. Subsequent work replaced the iterative optimization in the original formulation with generative approaches [6], [7], which learn a fixed-noise-to-perturbation mapping that produces UAPs in a single forward pass and improves attack diversity and transferability. However, this line of work focuses primarily on unimodal image classification under noise-free, digital input conditions, and does not address the multimodal alignment or wireless transmission concerns central to SemCom.

B. Adversarial Attacks on VLP Models

The adversarial robustness of VLP models has been actively studied since Co-Attack [8] introduced a multimodal attack that simultaneously perturbs both image and text inputs in a white-box setting. Subsequent work improved black-box transferability through set-level cross-modal guidance [9] and the joint use of modality-consistency and modality-discrepancy features [10]. These methods are all *instance-specific*, requiring a fresh perturbation for each input pair and incurring substantial computation in any large-scale deployment, which limits their practicality as a persistent threat.

UAPs offer a more efficient and persistent alternative. ETU [11] is among the first to study UAPs for VLP models, improving effectiveness and black-box transferability through local utility reinforcement and a ScMix augmentation strategy that increases multimodal input diversity while preserving semantics. C-PGC [12] advances this line with a generative framework conditioned on cross-modal embeddings; a malicious contrastive learning objective with farthest positive sample selection disrupts the learned multimodal alignment, and the framework extends naturally to text perturbations, yielding a truly multimodal UAP. Together, these results establish that universal attacks are a practical threat to VLP models and that cross-modal information plays a critical role in improving attack effectiveness and transferability.

C. Adversarial Robustness of Semantic Communication

Security threats in SemCom have been studied from several complementary perspectives [13], [14]. Early work focused on *physical-layer* threats such as eavesdropping on transmitted embeddings and jamming the semantic channel [15]–[17], demonstrating that the compactness of semantic embeddings does not by itself provide confidentiality or availability guarantees. A second line of work examines the *adversarial robustness of semantic encoders* themselves [18]: when standard image-domain perturbations are applied at the transmitter side and then forwarded through an AWGN channel, attack effectiveness degrades sharply because the channel noise corrupts the perturbation signal before it reaches the receiver. This observation shows that channel distortion fundamentally changes the adversarial threat surface in SemCom, but existing studies stop at characterizing the degradation rather than designing attacks that account for it.

Across these three research directions, namely generic UAPs, VLP-targeted attacks, and SemCom security, a clear gap emerges. Generic UAP research delivers efficient and persistent attacks but ignores multimodal alignment and wireless transmission. VLP-targeted UAP attacks address multimodal alignment but assume clean channels, leaving their effectiveness uncertain under realistic SemCom deployments, where adversarial perturbations themselves are corrupted during transmission. SemCom security work acknowledges this channel-induced degradation but does not propose constructive attacks that account for it, and *no prior work* addresses UAPs against cross-modal retrieval over realistic wireless channels. **SemCom-UAP** closes this gap by explicitly conditioning perturbation generation on the channel state, enabling a single modality-specific UAP to remain effective across a continuous range of SNR conditions.

III. System Model and Threat Model

System model. We consider a VLP-based SemCom system for cross-modal retrieval, including image-to-text (I2T) and text-

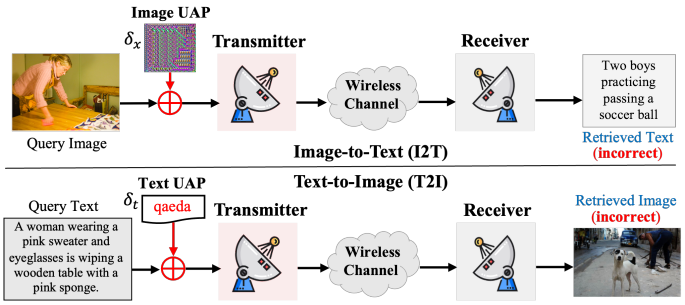

Fig. 1. Overview of our threat model and applied scenarios.

to-image (T2I) retrieval tasks. Both the transmitter and receiver are equipped with pre-trained VLP encoders for extracting semantic embeddings from images and texts. The transmitter encodes an image or text query into a compact embedding via a pre-trained VLP encoder and transmits it over a wireless channel subject to additive white Gaussian noise (AWGN). For I2T retrieval, the receiver observes $\tilde{\mathbf{v}}_i = \mathrm{norm}(\mathbf{v}_i + \mathbf{n})$, where $\mathbf{v}_i$ is the transmitted image embedding and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$ is channel noise with variance controlled by the operating SNR. Similarly, for T2I retrieval, the receiver observes $\tilde{\mathbf{u}}_i = \mathrm{norm}(\mathbf{u}_i + \mathbf{n})$, where $\mathbf{u}_i$ denotes the transmitted text embedding. Cross-modal retrieval is then performed using cosine similarity between the received noisy embedding and candidate embeddings from the opposite modality. The retrieval similarities for I2T and T2I are defined as $\mathrm{sim}_{\mathrm{I2T}}(i,j) = \cos(\tilde{\mathbf{v}}_i, \mathbf{u}_j)$ and $\mathrm{sim}_{\mathrm{T2I}}(i,j) = \cos(\tilde{\mathbf{u}}_i, \mathbf{v}_j)$, respectively.

Adversary's goal. The adversary aims to learn universal perturbations for image (δ_x) and text (δ_t), respectively, such that the perturbed image input $(x + \delta_x)$ or perturbed text input $(\mathcal{T} \oplus \delta_t)$ compromises cross-modal retrieval at the receiver.

Adversary's knowledge and capabilities. The adversary $\mathcal{A}$ is assumed to interfere with the transmitter-side input pipeline of the target SemCom system, e.g., through compromised sensors, malicious edge applications, or preprocessing pipeline tampering, which is consistent with input-stage and edge-deployed attack surfaces studied in adversarial SemCom and edge AI systems [13], [19]. $\mathcal{A}$ can manipulate image or text inputs before semantic encoding by injecting UAPs, and has white-box access to a set of surrogate VLP models $\mathcal{F}$ for offline adversarial optimization [4], [20]. The perturbations are constrained within predefined image and text perturbation budgets to remain visually or semantically inconspicuous. However, $\mathcal{A}$ does not have access to the victim VLP model deployed in the target SemCom system and cannot directly modify the VLP model, retrieval database, or wireless channel. This corresponds to a realistic grey-box threat model for transmitter-side adversarial attacks in SemCom systems.

Example scenarios. Fig. 1 illustrates two primary scenarios considered in our study. *a) I2T retrieval:* In a wearable AR system, a malicious AR application injects imperceptible perturbations into captured images before semantic encoding, causing the cloud retrieval service to return incorrect textual descriptions or guidance. *b) T2I retrieval:* In a telemedicine SemCom system [21], a compromised mobile health application perturbs symptom-description text queries, causing the retrieval service to return contextually aligned but semantically irrelevant medical images.

Connection to method design. The grey-box assumption motivates the UCB-based adaptive surrogate selection strategy (Section IV-E). By exposing the UAP generators to diverse surrogate gradient fields during training, the learned perturbations become more transferable beyond the primary surrogate.

IV. Method

A. Problem Setup

Let $\mathcal{D} = \{(x_i, \mathcal{T}_i)\}_{i=1}^N$ denote an image–text dataset, where $x_i \in \mathbb{R}^{H \times W \times 3}$ and $\mathcal{T}_i = \{\mathcal{T}_{i,k}\}_{k=1}^K$ contains K associated captions. Let $\mathcal{F} = \{f^{(s)}(\cdot)\}_{s=1}^S$ denote a set of surrogate VLP models. Each surrogate $f(\cdot) \in \mathcal{F}$ consists of an image encoder $f_{\mathrm{img}}(\cdot)$ and a text encoder $f_{\mathrm{txt}}(\cdot)$ that map inputs into a shared embedding space. For simplicity, we omit the surrogate index (s) unless necessary. We use $(\mathcal{B}_x, \mathcal{B}_t)$ to denote a mini-batch of paired image and text samples, where $|\mathcal{B}_x| = |\mathcal{B}_t|$, and use $\mathrm{norm}(\cdot)$ for ℓ_2 normalization.

1) Threat Goal: The adversary aims to learn an image UAP δ_x and a text UAP δ_t such that the perturbed image input $(x + \delta_x)$ or perturbed text input $(\mathcal{T} \oplus \delta_t)$ compromises cross-modal retrieval under wireless channel corruption. To ensure imperceptibility, the image perturbation is constrained by an ℓ_∞ budget $\|\delta_x\|_\infty \le \epsilon_x$, while the text perturbation is restricted to modifying at most ϵ_t words. Unless otherwise specified, we use $\epsilon_t = 1$ in all experiments. To generate perturbations robust across varying wireless conditions, we optimize two SNR-aware conditional generators, $G_x(\cdot)$ and $G_t(\cdot)$, jointly conditioned on cross-modal semantic features and channel state information.

2) Feature Bank Construction: To preserve surrogate-specific semantic structures, we construct separate image and text feature banks for each surrogate model $f(\cdot) \in \mathcal{F}$. The normalized image and text features are extracted offline for all training samples using the first caption $\mathcal{T}_{i,1}$ as the default text description:

$$\mathbf{v}_i = \mathrm{norm}(f_{\mathrm{img}}(x_i)), \qquad \mathbf{u}_i = \mathrm{norm}(f_{\mathrm{txt}}(\mathcal{T}_{i,1})), \tag{1}$$

forming feature banks $\mathbf{V} \in \mathbb{R}^{N \times d}$ and $\mathbf{U} \in \mathbb{R}^{N \times d}$ for each surrogate $f(\cdot)$. These banks are used exclusively for generator conditioning (Section IV-B). Adversarial target selection is performed dynamically within the current mini-batch rather than over the global banks.

B. SNR-Aware Universal Perturbation Generation

We train two conditional generators, G_x and G_t, to produce universal image and text perturbations, respectively. Each generator takes a learnable latent code and an SNR-aware condition vector as input. Specifically, $\mathbf{z}_x$ and $\mathbf{z}_t$ denote learnable latent codes for the image and text branches, respectively. Both latent codes are independently initialized from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and optimized jointly with the generators during training. The

condition vector jointly encodes cross-modal semantic context and wireless channel state information, allowing the generated perturbations to adapt to both semantic content and channel quality. The two branches are designed asymmetrically: the image branch uses instance-level textual guidance to encourage perturbation diversity, whereas the text branch uses batch-level image guidance to stabilize universal text perturbation learning. Both generators share a lightweight SNR encoder.

1) Shared SNR Encoder: We introduce a lightweight SNR encoder implemented as a two-layer MLP with SiLU activations:

$$\mathbf{e}_{\mathrm{snr}} = f_{\mathrm{snr}}(\gamma) \in \mathbb{R}^{d_{\mathrm{snr}}}, \tag{2}$$

where γ denotes the channel SNR in dB and $d_{\mathrm{snr}} = 64$. The encoder maps the scalar SNR through two fully connected layers with dimensions $1 \to 64 \to 64$.

During training, γ is sampled from the pre-defined SNR range $[\gamma_{\min}, \gamma_{\max}]$ using the stratified sampling strategy described in Section IV-F. The learned embedding captures channel-aware properties and is shared across both generators.

2) Cross-Modal Condition Construction: To simplify notation, we define two SNR-aware condition functions for the image and text branches, respectively.

a) Image-Branch Condition: For image perturbation generation, the mean caption embedding is used as the cross-modal semantic reference: $\mathbf{e}_i^{\mathrm{txt}} = \mathrm{norm}((\sum_k f_{\mathrm{txt}}(\mathcal{T}_{i,k}))/K)$.

The image-branch condition function is then defined as:

$$\mathbf{c}_{x,i} = C_x(\mathcal{T}_i, \gamma) = \left[\mathbf{e}_i^{\mathrm{txt}}; f_{\mathrm{snr}}(\gamma)\right] \in \mathbb{R}^{d+d_{\mathrm{snr}}}, \tag{3}$$

where $[\cdot\,;\,\cdot]$ denotes feature concatenation.

b) Text-Branch Condition.: For text perturbation generation, we use a batch-level image reference computed from the cached image features: $\mathbf{e}^{\mathrm{img}} = \mathrm{norm}\big((\sum_{i\in\mathcal{B}_x} \mathbf{v}_i)/(|\mathcal{B}_x|)\big)$.

The text-branch condition function is defined as:

$$\mathbf{c}_t = C_t(\mathcal{B}_x, \gamma) = \left[\mathbf{e}^{\mathrm{img}}; f_{\mathrm{snr}}(\gamma)\right] \in \mathbb{R}^{d+d_{\mathrm{snr}}}. \tag{4}$$

3) Conditional UAP Generation:

a) Image UAP Generation.: The image generator produces a sample-conditioned perturbation candidate:

$$\hat{\delta}_{x,i} = G_x\left(\mathbf{z}_x;\, \mathbf{c}_{x,i}\right), \tag{5}$$

where $\mathbf{z}_x \in \mathbb{R}^{B\times d_z\times h_z\times w_z}$ is a learnable latent code initialized from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and optimized jointly with the generator.

Applying G_x to all samples within the batch yields candidate perturbations $\{\hat{\delta}_{x,i}\}_{i=1}^{|\mathcal{B}_x|}$. These candidates are aggregated to form the final universal image perturbation: $\hat{\delta}_x = (\sum_{i=1}^{|\mathcal{B}_x|} \hat{\delta}_{x,i})/(|\mathcal{B}_x|)$. The aggregated perturbation is clipped to satisfy the ℓ_∞ constraint: $\delta_x = \mathrm{clip}(\hat{\delta}_x, -\epsilon_x, \epsilon_x)$.

b) Text UAP Generation.: The text generator produces a universal embedding space perturbation:

$$\hat{\delta}_t = G_t\left(\mathbf{z}_t;\, \mathbf{c}_t\right), \tag{6}$$

where $\mathbf{z}_t$ is a learnable latent code optimized independently from $\mathbf{z}_x$ using a dedicated Adam optimizer. To preserve embedding-space stability and prevent excessively large semantic shifts, the generated text perturbation $\hat{\delta}_t$ is globally norm-clamped within a predefined magnitude range before being applied to the text embeddings, yielding δ_t.

C. Channel-Robust Adversarial Embedding Construction

To improve robustness against wireless transmission distortion, we simulate channel corruption directly in the embedding space during training. For each training step, the same sampled SNR value γ used by the shared SNR encoder is also used to parameterize the AWGN corruption process, ensuring consistent channel-aware perturbation generation and optimization.

1) Adversarial Embedding Extraction: Using the generated perturbations, adversarial inputs are constructed as:

$$x_i^{\mathrm{adv}} = \mathrm{clip}(x_i + \delta_x,\, 0, 1), \qquad \mathcal{T}_{i,1}^{\mathrm{adv}} = \mathcal{T}_{i,1} \oplus \delta_t, \tag{7}$$

where $\oplus$ denotes word-level embedding replacement: the target word is selected by ranking word-level importance scores, and its sub-word span is replaced by δ_t. The perturbed inputs are then encoded by the surrogate $f(\cdot)$ to obtain adversarial embeddings:

$$\mathbf{v}_i^{\mathrm{adv}} = \mathrm{norm}(f_{\mathrm{img}}(x_i^{\mathrm{adv}})), \quad \mathbf{u}_i^{\mathrm{adv}} = \mathrm{norm}(f_{\mathrm{txt}}(\mathcal{T}_{i,1}^{\mathrm{adv}})). \tag{8}$$

2) AWGN Corruption: Following prior semantic communication studies [22], we model the wireless channel as AWGN in the embedding space. Given the sampled SNR γ, the noisy adversarial image embeddings are computed as:

$$\tilde{\mathbf{v}}_i^{\mathrm{adv}} = \mathrm{norm}\big(\mathbf{v}_i^{\mathrm{adv}} + \mathbf{n}_i\big), \quad \mathbf{n}_i \sim \mathcal{N}(\mathbf{0},\, \sigma_n^2\mathbf{I}), \tag{9}$$

where the noise standard deviation σ_n is shared across all samples in the batch and computed as:

$$\sigma_n = \sqrt{\frac{p_{\mathrm{signal}}}{10^{\gamma/10}}}, \qquad p_{\mathrm{signal}} = \frac{1}{d\cdot|\mathcal{B}_x|}\sum_{i\in\mathcal{B}_x}\left\|\mathbf{v}_i^{\mathrm{adv}}\right\|_2^2, \tag{10}$$

with d the embedding dimension. The same procedure yields $\tilde{\mathbf{u}}_i^{\mathrm{adv}}$ for adversarial text features. Crucially, all adversarial losses are computed on the *noisy* features, forcing the generators to produce perturbations that survive channel corruption.

D. Adversarial Cross-Modal Contrastive Learning

To maximize semantic disruption in the shared embedding space, we optimize the adversarial embeddings toward semantically distant cross-modal targets within the current batch while separating them from their clean counterparts. For clarity, we describe the image branch below; the text branch is defined symmetrically.

1) Farthest Cross-Modal Target Selection: Given noisy adversarial image embeddings $\{\tilde{\mathbf{v}}_i^{\mathrm{adv}}\}_{i=1}^{|\mathcal{B}_x|}$, clean image embeddings $\{\mathbf{v}_j\}_{j=1}^{|\mathcal{B}_x|}$, and clean text embeddings $\{\mathbf{u}_j\}_{j=1}^{|\mathcal{B}_x|}$, the target index for $\tilde{\mathbf{v}}_i^{\mathrm{adv}}$ is selected as:

$$y_i^v = \arg\min_{j\neq i}\left(\tilde{\mathbf{v}}_i^{\mathrm{adv}\top}\mathbf{v}_j + \tilde{\mathbf{v}}_i^{\mathrm{adv}\top}\mathbf{u}_j\right), \tag{11}$$

where lower similarity indicates a larger semantic distance in the shared embedding space. The adversarial targets for $\tilde{\mathbf{v}}_i^{\mathrm{adv}}$ are then defined as $\mathbf{p}_i = \{\mathbf{v}_{y_i^v}, \mathbf{u}_{y_i^v}\}$, where $\mathbf{v}_{y_i^v}$ and $\mathbf{u}_{y_i^v}$ are the clean image and text embeddings with the index y_i^v.

2) Contrastive Objective: We optimize a margin-based contrastive objective that encourages adversarial embeddings to align with the farthest cross-modal targets while separating them from their clean counterparts. For adversarial target $\mathbf{p}_i$, the per-embedding image-branch loss is:

$$\ell_i^{(\mathbf{v})}(\mathbf{p}_i) = -\log \frac{\mathrm{sim}_m(\tilde{\mathbf{v}}_i^{\mathrm{adv}}, \mathbf{p}_i)}{\Omega_i^{(\mathbf{v})}(\mathbf{p}_i)}, \tag{12}$$

where $\mathrm{sim}_m(\mathbf{a}, \mathbf{b}) = \exp((\mathbf{a}^\top \mathbf{b} - m)/\tau)$ is a margin-based similarity function with margin $m > 0$ and temperature τ; and $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \exp(\mathbf{a}^\top \mathbf{b}/\tau)$. The partition function $\Omega_i^{(\mathbf{v})}(\mathbf{p}_i) = \mathrm{sim}(\tilde{\mathbf{v}}_i^{\mathrm{adv}}, \mathbf{p}_i) + \mathrm{sim}(\tilde{\mathbf{v}}_i^{\mathrm{adv}}, \mathbf{v}_i) + \sum_{j \neq i,\, j \neq y_i^v} \mathrm{sim}(\tilde{\mathbf{v}}_i^{\mathrm{adv}}, \mathbf{v}_j)$ aggregates: (i) the identified target $\mathbf{p}_i$, (ii) the sample's own clean feature $\mathbf{v}_i$, and (iii) the remaining clean image features within the batch as negatives. The image branch uses image-only negatives; the text branch uses text-only negatives symmetrically.

The final image-branch loss averages over both image and text targets for $\{\tilde{\mathbf{v}}_i^{\mathrm{adv}}\}_{i=1}^{|\mathcal{B}_x|}$:

$$\mathcal{L}_{\mathrm{img}} = \frac{1}{2|\mathcal{B}_x|} \sum_{i=1}^{|\mathcal{B}_x|} \Big(\ell_i^{(\mathbf{v})}\big(\mathbf{v}_{y_i^v}\big) + \ell_i^{(\mathbf{v})}\big(\mathbf{u}_{y_i^v}\big) \Big). \tag{13}$$

The text-branch loss $\mathcal{L}_{\mathrm{txt}}$ is defined symmetrically by replacing $\tilde{\mathbf{v}}_i^{\mathrm{adv}}$ with $\tilde{\mathbf{u}}_i^{\mathrm{adv}}$ throughout, with text-only negatives in the partition function.

$$\mathcal{L}_{\mathrm{txt}} = \frac{1}{2|\mathcal{B}_t|} \sum_{i=1}^{|\mathcal{B}_t|} \Big(\ell_i^{(\mathbf{u})}\big(\mathbf{v}_{y_i^u}\big) + \ell_i^{(\mathbf{u})}\big(\mathbf{u}_{y_i^u}\big) \Big), \tag{14}$$

where $\mathbf{v}_{y_i^u}$ and $\mathbf{u}_{y_i^u}$ are the clean image and text embeddings with the selected target index y_i^u for $\tilde{\mathbf{u}}_i^{\mathrm{adv}}$.

E. Adaptive Surrogate Selection

To improve transferability beyond the primary surrogate under the proposed grey-box threat model (Section III), we expose the generators to diverse surrogate gradient fields by adaptively selecting auxiliary surrogate models during training. Given the surrogate set $\mathcal{F} = \{f^{(1)}, \ldots, f^{(S)}\}$, the first model $f^{(1)}$ serves as the primary surrogate, while an auxiliary surrogate $f^{(s)}$ $(s \neq 1)$, is dynamically selected at each training step to provide additional optimization signals. For each surrogate model $f^{(s)}$, the adversarial cross-modal contrastive losses introduced in Section IV-D are computed, yielding $\mathcal{L}_{\mathrm{img}}^{(s)}$ and $\mathcal{L}_{\mathrm{txt}}^{(s)}$.

1) Surrogate Taxonomy: Motivated by multi-surrogate transfer attacks and recent surrogate-scaling strategies for CLIP UAPs [23], [24], we formulate auxiliary surrogate selection as an online model-selection problem and use an Upper Confidence Bound (UCB)-based criterion to balance exploration and exploitation [25]. We categorize auxiliary surrogates according to their compatibility with the text perturbation mechanism. (1) *Same-system* surrogates belong to the same model family as the primary surrogate $f^{(1)}$ (e.g., BERT-based models such as TCL, BLIP, and XVLM when ALBEF is used as the primary model). These models share compatible token embedding interfaces, allowing supervision for both image and text branches. (2) *Cross-system* surrogates belong to a different model family from $f^{(1)}$ (e.g., CLIP-based ViT-B/32 and ViT-B/16 models when ALBEF is used as the primary model). Since their text encoders operate in a different embedding space, they supervise only the image branch. Accordingly, the surrogate loss for model $f^{(s)}$ $(s \neq 1)$ is defined as:

$$\mathcal{L}^{(s)} = \begin{cases} \mathcal{L}_{\mathrm{img}}^{(s)} + \mathcal{L}_{\mathrm{txt}}^{(s)} & \text{if } f^{(s)} \text{ is same-system,} \\ \mathcal{L}_{\mathrm{img}}^{(s)} & \text{if } f^{(s)} \text{ is cross-system.} \end{cases} \tag{15}$$

2) UCB-based Surrogate Selection: We model surrogate selection as a multi-armed bandit and maintain a running reward $\bar{R}(s)$ for each surrogate $f^{(s)}$. Surrogates yielding higher adversarial loss expose harder vulnerable directions and should be selected more often. We therefore treat the normalized loss as the bandit reward. To make rewards comparable across heterogeneous model families (e.g., BERT vs. CLIP), we apply the z-score normalization for each surrogate loss:

$$\hat{\mathcal{L}}(s) = \frac{\mathcal{L}^{(s)} - \mu_s}{\sigma_s + \epsilon_0}, \tag{16}$$

where μ_s and σ_s are exponential moving averages of the mean and standard deviation of $\mathcal{L}^{(s)}$, and $\epsilon_0 > 0$ is a small constant for numerical stability. All reward estimates are initialized to zero, $\bar{R}(s) \leftarrow 0$, and selection counts to $n_s \leftarrow 0$ for all $s \neq 1$. The reward is then updated with momentum η:

$$\bar{R}(s) \leftarrow (1 - \eta)\, \bar{R}(s) + \eta\, \hat{\mathcal{L}}(s). \tag{17}$$

We select the surrogate using the UCB criterion:

$$\mathrm{UCB}(s) = \bar{R}(s) + c_{\mathrm{ucb}} \sqrt{\frac{\log(T+1)}{n_s + \epsilon_0}}, \tag{18}$$

where T is the total selection count, n_s is the number of times surrogate s has been selected, and $c_{\mathrm{ucb}} > 0$ is the exploration coefficient. At each step, the auxiliary surrogate with the highest $\mathrm{UCB}(s)$ score is selected $(s \neq 1)$, balancing exploitation of high-reward surrogates and exploration of less frequently selected ones. To prevent under-utilization, each surrogate is guaranteed to be selected at least $\lfloor n_{\mathrm{bat}} \cdot \rho_{\min} \rfloor$ times per epoch, where n_{bat} is the number of batches per epoch, preventing starvation of cross-system surrogates.

3) Adaptive Loss Weighting: The selected surrogate's loss is scaled by a sigmoid gate:

$$\mathcal{L}_{\mathrm{sur}}^{(s^\star)} = \lambda_{\mathrm{sur}} \cdot \sigma\big(\bar{R}(s^\star)\big) \cdot \mathcal{L}^{(s^\star)}, \tag{19}$$

where $\lambda_{\mathrm{sur}} > 0$ is a global scale and $\sigma(\cdot)$ is the sigmoid function, dynamically up-weighting surrogates that expose harder vulnerable directions.

F. Training Procedure

1) Stratified SNR Sampling: Optimizing under a fixed SNR may lead to overfitting to a narrow channel regime. To improve robustness across varying channel conditions, we adopt stratified SNR sampling. Specifically, the SNR range $[\gamma_{\min}, \gamma_{\max}]$

is divided into M equal intervals, and one SNR value is uniformly sampled from each interval:

$$\gamma_m \sim \mathcal{U}\left(\gamma_{\min} + \frac{(m-1)\Delta}{M},\ \gamma_{\min} + \frac{m\Delta}{M}\right), \tag{20}$$

where $m = 1, \ldots, M$ and $\Delta = \gamma_{\max} - \gamma_{\min}$. This design guarantees coverage across the entire SNR range during training.

For each sampled SNR γ_m, the corresponding condition vectors are constructed and used to generate channel-aware perturbations. The generated adversarial embeddings are then corrupted by the embedding-space AWGN channel, and used to compute the contrastive losses for both the primary and auxiliary surrogates. The step-wise primary loss is obtained by averaging across all SNR strata:

$$\mathcal{L}_{\text{step}}^{(1)} = \frac{1}{M} \sum_{m=1}^{M} \left(\mathcal{L}_{\text{img},m}^{(1)} + \mathcal{L}_{\text{txt},m}^{(1)} \right), \tag{21}$$

where $\mathcal{L}_{\text{img},m}^{(1)}$ and $\mathcal{L}_{\text{txt},m}^{(1)}$ denote the image- and text-branch contrastive losses computed using the adversarial embeddings generated under the sampled SNR γ_m. The surrogate loss $\mathcal{L}_{\text{sur}}^{(s^\star)}$ is computed similarly. We use $M = 3$ in all experiments.

2) Overall Objective: The full per-iteration objective is:

$$\mathcal{L} = \mathcal{L}_{\text{step}}^{(1)} + \mathcal{L}_{\text{sur}}^{(s^\star)}, \tag{22}$$

where $\mathcal{L}_{\text{sur}}^{(s^\star)} = 0$ during the warm-up phase ($\text{epoch} < E_{\text{warm}}$). After which, the adaptive surrogate selection is enabled. The generators G_x, G_t, and the SNR encoder f_{snr} are optimized with a primary Adam optimizer at a learning rate η_g. The latent codes $\mathbf{z}_x$ and $\mathbf{z}_t$ are optimized separately using Adam with scaled learning rate $\eta_z = \alpha\eta_g$, where $\alpha < 1$ prevents the latent codes from dominating generator optimization.

3) Computational Considerations: Training is performed *entirely offline* prior to any attack. Once trained, the generators produce a fixed universal perturbation pair (δ_x, δ_t) in a single forward pass by averaging over the SNR strata, so that no instantaneous channel estimate is required at deployment. The image UAP is added directly to any query at zero cost, while the text UAP only requires locating the substitution position on the surrogate, without access to the victim model. This asymmetry—expensive offline training, near-instant online deployment—matches the UAP threat model in Section III: an adversary investing one-time compute obtains perturbations reusable across arbitrary queries and the full SNR range.

V. Experiments

A. Experimental Setup

1) Datasets and Metrics: We evaluate on the Flickr30K [26] dataset, which contains 31,783 images each paired with five human-annotated captions, using the standard 1,000-image test split. We report results across six evaluation VLP models: ALBEF [27], TCL [28], BLIP [29], X-VLM [30], and ViT-B/16 and ViT-B/32 [3].

a) Retrieval tasks: We evaluate two standard cross-modal retrieval directions: I2T and T2I. For I2T, adversarial image queries (perturbed with UAP and AWGN) are matched against a *clean* text gallery. For T2I, adversarial text queries (with word-level perturbation and AWGN) are matched against an *adversarial* image gallery (also corrupted with UAP and AWGN), reflecting a realistic SemCom scenario where both the transmitted query and the indexed images pass through the wireless channel. We report Recall at Rank 1 (R@1) for both directions.

b) Pure UAP ASR: Attack success rate (ASR) may still include cases where the UAP and channel noise fail the query *jointly* but the UAP alone would not suffice. To further isolate the UAP's standalone contribution, we define *Pure UAP ASR* at a given SNR γ:

$$\text{ASR}_{\text{pure}}(\gamma) = \frac{\left|\mathcal{S}_{\text{clean}}(\gamma) \cap \mathcal{Q}_{\text{adv}}^{\text{no-noise}}\right|}{\left|\mathcal{S}_{\text{clean}}(\gamma)\right|}, \tag{23}$$

where $\mathcal{S}_{\text{clean}}(\gamma)$ is the set of queries that succeed under clean inputs at SNR γ, and $\mathcal{Q}_{\text{adv}}^{\text{no-noise}}$ is the set of queries that fail under adversarial inputs transmitted over a noiseless channel ($\sigma_n = 0$). The denominator conditions on queries that the clean system handles correctly at the operating SNR, while the numerator counts those defeated by the UAP alone—without any channel noise assistance. This provides a conservative lower bound on attack effectiveness: a UAP achieving high $\text{ASR}_{\text{pure}}(\gamma)$ is genuinely effective through semantic perturbation alone, independent of channel conditions.

2) Implementation Details: All experiments are conducted on a single NVIDIA H100 GPU per run, with six independent runs corresponding to the six primary encoder configurations. The image perturbation budget is $\epsilon_x = 12/255$ under the ℓ_∞ norm, and the number of substituted words is $\epsilon_t = 1$. Generators are trained for 100 epochs with batch size 64 using the Adam optimizer at learning rate $\eta_g = 2 \times 10^{-4}$; the latent codes $\mathbf{z}_x$ and $\mathbf{z}_t$ are updated by separate Adam optimizers at scaled learning rate $\eta_z = 0.05 \cdot \eta_g$. The SNR range is $[\gamma_{\min}, \gamma_{\max}] = [0, 20]$ dB with $M = 3$ stratified samples per step. The contrastive temperature is $\tau = 0.07$ and margin $m = 0.02$ for both branches. For UCB surrogate selection, the exploration coefficient is $c_{\text{ucb}} = 1.0$, EMA momentum $\eta = 0.1$, auxiliary loss scale $\lambda_{\text{sur}} = 0.65$, floor ratio $\rho_{\min} = 0.02$, and warm-up phase $E_{\text{warm}} = 40$ epochs.

B. Pure UAP ASR Analysis

UAP intrinsic strength. Table I reports Pure UAP ASR at 20 dB and 0 dB across all six primary encoder configurations. SemCom-UAP consistently and substantially outperforms C-PGC across all source–target pairs and both retrieval directions. With ALBEF as primary surrogate, our method achieves 81.32%/88.98% I2T/T2I Pure UAP ASR at 20 dB, compared to 19.01%/52.99% for C-PGC—a gap exceeding 60 percentage points. The uniformly deep-red heatmap in Figure 2 further confirms that this advantage holds consistently across all source–target combinations at both 5 dB and 15 dB. These

TABLE I

PURE UAP ASR@R1 (%) ON FLICKR30K AT 20 DB AND 0 DB SNR. THE FAILURE SET IS FIXED AT $\gamma{=}\infty$ ACROSS ALL SNR CONDITIONS, ISOLATING THE UAP'S STANDALONE CONTRIBUTION (SEE EQ. (23)). EACH PAIR OF ROWS: CPGC (TOP) / **SEMCOM-UAP (OURS, BOTTOM)**. **BOLD** = HIGHER PURE UAP ASR.

Source	SNR	Method	ALBEF		TCL		BLIP		XVLM		ViT-B/16		ViT-B/32	
			I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
ALBEF	20 dB	CPGC	19.01	52.99	20.98	52.47	13.77	52.70	6.77	45.10	7.28	51.99	12.40	54.24
		Ours	**81.32**	**88.98**	**62.02**	**75.90**	**42.99**	**70.61**	**71.18**	**75.83**	**12.01**	**60.94**	**13.05**	**62.09**
	0 dB	CPGC	17.22	42.94	18.33	48.87	11.98	46.82	5.14	40.16	7.90	46.02	11.06	50.18
		Ours	**80.21**	**84.92**	**59.10**	**73.53**	**35.33**	**65.34**	**67.22**	**73.03**	**9.98**	**55.26**	**11.50**	**57.73**
BLIP	20 dB	CPGC	3.03	6.27	3.06	5.69	5.29	54.67	6.54	33.23	6.84	50.28	12.60	50.88
		Ours	**51.73**	**50.13**	**57.89**	**55.08**	**75.23**	**88.02**	**70.01**	**83.56**	**12.41**	**56.87**	**14.70**	**59.51**
	0 dB	CPGC	2.85	5.64	4.31	8.44	7.23	49.53	6.02	29.16	6.54	43.89	12.08	46.40
		Ours	**49.50**	**43.22**	**55.56**	**51.72**	**67.47**	**86.27**	**67.74**	**81.88**	**11.45**	**52.46**	**15.68**	**54.87**
TCL	20 dB	CPGC	16.09	18.69	41.27	60.98	15.32	56.69	7.75	42.62	8.05	55.47	12.02	54.10
		Ours	**48.81**	**48.62**	**91.38**	**93.62**	**40.03**	**70.74**	**58.84**	**79.15**	**13.15**	**61.60**	**15.50**	**61.38**
	0 dB	CPGC	14.91	15.08	38.11	58.38	12.28	50.00	7.36	37.55	8.52	49.17	9.73	48.65
		Ours	**47.30**	**43.13**	**90.57**	**93.08**	**34.43**	**65.34**	**57.08**	**77.09**	**13.10**	**56.54**	**12.61**	**56.87**
XVLM	20 dB	CPGC	6.61	9.46	6.86	10.29	9.61	12.28	11.29	50.53	8.67	54.77	10.34	52.78
		Ours	**20.15**	**28.40**	**28.95**	**32.20**	**23.38**	**23.37**	**91.53**	**90.06**	**12.24**	**58.74**	**16.36**	**59.42**
	0 dB	CPGC	5.76	8.02	8.18	11.73	10.09	15.78	9.29	46.04	8.25	50.43	7.06	49.10
		Ours	**17.54**	**22.84**	**27.04**	**29.72**	**19.58**	**21.02**	**90.57**	**89.52**	**11.27**	**53.53**	**13.02**	**54.96**
ViT-B/16	20 dB	CPGC	9.74	12.50	15.99	18.08	13.72	17.84	8.84	10.59	19.90	60.71	15.44	55.02
		Ours	**14.14**	**17.66**	**24.32**	**26.34**	**17.50**	**20.34**	**11.63**	**14.66**	**79.46**	**92.41**	**28.99**	**63.66**
	0 dB	CPGC	8.67	12.36	13.80	17.41	15.88	17.70	8.27	10.10	18.54	57.75	13.18	50.41
		Ours	**12.47**	**13.69**	**19.97**	**24.60**	**14.92**	**17.99**	**9.47**	**12.66**	**77.63**	**91.57**	**25.68**	**59.62**
ViT-B/32	20 dB	CPGC	10.30	13.92	17.25	19.54	13.47	18.55	9.17	10.72	15.06	22.74	28.39	64.17
		Ours	**19.80**	**23.00**	**32.43**	**31.54**	**23.68**	**26.21**	**16.54**	**18.13**	**31.04**	**44.41**	**79.44**	**90.30**
	0 dB	CPGC	9.13	12.50	15.24	18.28	11.79	16.70	8.58	9.64	13.75	22.03	25.31	61.40
		Ours	**17.98**	**18.49**	**29.04**	**29.88**	**17.21**	**23.14**	**14.48**	**15.55**	**26.72**	**40.71**	**78.75**	**89.84**

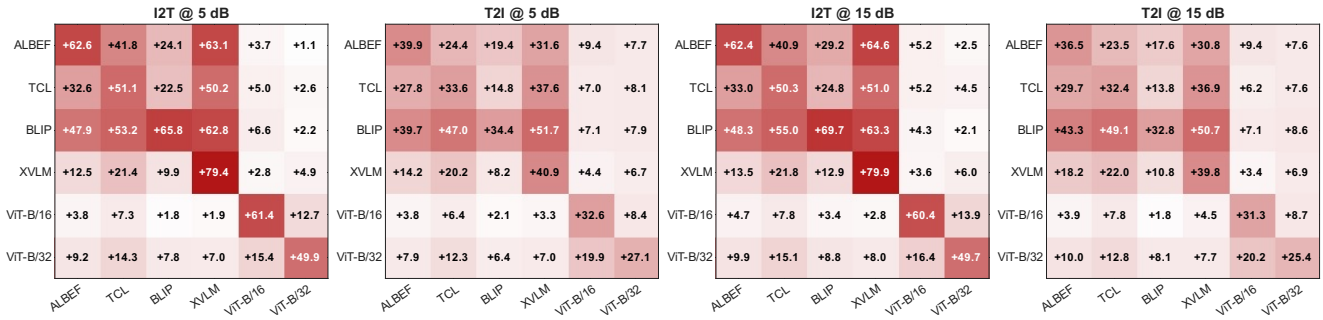

Fig. 2. Pure UAP ASR gain (%) of SemCom-UAP over C-PGC on Flickr30K at SNR = 5 dB and 15 dB (I2T and T2I). Rows = source model; columns = target model. Deeper red indicates larger gain.

results demonstrate that SemCom-UAP's perturbations are genuinely effective through semantic disruption alone, whereas C-PGC's overall ASR is largely sustained by the *joint* effect of the UAP and AWGN rather than the perturbation itself.

Channel robustness. The difference between 20 dB and 0 dB Pure UAP ASR remains within 8 percentage points for our method across all configurations (e.g., ALBEF I2T: 81.32%→80.21%, a relative drop of 1.4%), compared with a 19.0% relative drop for C-PGC on ALBEF T2I, confirming that the learned perturbations retain their intrinsic attack strength across a wide range of channel conditions. This robustness is a direct consequence of stratified SNR sampling and SNR-aware conditioning during training (Section IV-F), which exposes the generators to diverse channel conditions rather than overfitting to a single SNR regime.

Cross-model transferability. For BERT-family primary surrogates, SemCom-UAP achieves strong cross-model transfer: with BLIP as primary, our method reaches 57.89%/55.08% on TCL and 70.01%/83.56% on XVLM at 20 dB, versus 3.06%/5.69% and 6.54%/33.23% for C-PGC. Transfer gains are more modest for CLIP-to-BERT pairs due to the architectural gap between encoder families, yet SemCom-UAP still

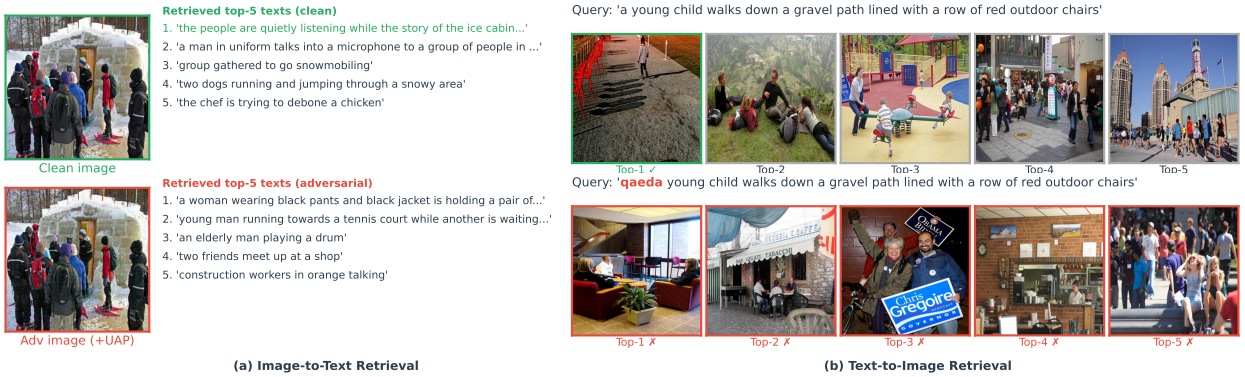

Fig. 3. Qualitative examples of SemCom-UAP on Flickr30K (ALBEF as the primary surrogate, SNR = 10 dB). In T2I, the adversarially substituted word is highlighted in red.

TABLE II

PURE UAP ASR@R1 (%) UNDER VARIOUS INPUT-SPACE DEFENSES ON FLICKR30K. SOURCE MODEL: ALBEF. THE FAILURE SET IS FIXED AT $\gamma=\infty$ TO ISOLATE THE UAP'S STANDALONE CONTRIBUTION (SEE EQ. (23)). DEFENSES ARE APPLIED PRIOR TO SEMANTIC ENCODING.

Defense	SNR	ALBEF		TCL		BLIP		XVLM		ViT-B/16		ViT-B/32	
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
No Defense	20 dB	81.32	88.98	62.02	75.90	42.99	70.61	71.18	75.83	12.01	60.94	13.05	62.09
	0 dB	80.21	84.92	59.10	73.53	35.33	65.34	67.22	73.03	9.98	55.26	11.50	57.73
Gaussian	20 dB	49.94	76.62	27.58	56.42	20.91	61.02	9.88	40.75	8.57	59.67	13.87	59.98
	0 dB	46.81	70.90	25.31	53.26	18.57	55.71	8.35	37.07	8.23	53.72	12.03	56.03
JPEG	20 dB	47.17	71.50	39.29	62.22	34.01	66.45	19.54	48.47	13.94	60.89	12.83	59.58
	0 dB	45.80	67.99	36.54	59.80	27.30	61.63	16.26	43.98	12.96	54.98	10.24	55.55
Feature Squeezing	20 dB	78.26	88.20	58.06	72.60	37.97	69.05	52.41	65.28	9.46	59.29	13.48	61.98
	0 dB	76.79	84.58	56.43	70.74	30.89	64.67	48.54	61.35	11.32	52.15	11.58	57.78
Rand. Smoothing	20 dB	75.90	87.65	54.76	71.67	38.43	68.81	47.19	63.14	9.59	60.39	13.22	61.81
	0 dB	73.45	85.95	53.82	69.62	31.33	65.26	44.96	59.12	10.49	53.89	12.03	57.30

outperforms C-PGC on all such pairs, confirming the benefit of UCB-based surrogate selection in capturing cross-family vulnerable directions.

C. Qualitative Analysis

Figure 3 presents representative retrieval examples using ALBEF as the primary surrogate. In I2T, the clean image correctly retrieves semantically matched captions, while the adversarially perturbed image retrieves entirely unrelated texts across all top-5 results. In T2I, substituting a single word suffices to completely redirect retrieval—all top-5 retrieved images under the adversarial query are unrelated to the ground-truth image, whereas the clean query correctly retrieves the matching image at rank 1. These examples demonstrate that SemCom-UAP achieves effective semantic disruption with minimal and imperceptible modifications to the input.

D. Robustness Against Input-Space Defenses

We evaluate SemCom-UAP against four standard input-space defenses applied to adversarial inputs at the receiver: Gaussian noise augmentation, JPEG compression, feature squeezing, and randomized smoothing. Table II reports Pure UAP ASR@R1 with ALBEF as the source model at 20 dB and 0 dB SNR.

Resistance to preprocessing defenses. Gaussian noise and JPEG compression reduce Pure UAP ASR on the primary-surrogate model (ALBEF) from 81.32% to 49.94% and 47.17% (I2T, 20 dB), respectively, indicating that pixel-level perturbations are partially disrupted by these transformations. Substantial attack success remains across all target models, particularly in T2I where Pure UAP ASR stays above 50% in nearly all configurations even under these defenses.

Ineffectiveness of feature-level defenses. Feature squeezing and randomized smoothing provide negligible reduction in attack strength: Pure UAP ASR on ALBEF drops by only 3.1 and 5.4 percentage points (I2T, 20 dB) relative to the undefended baseline, respectively. This suggests that SemCom-UAP produces perturbations that operate primarily in the semantic embedding space rather than relying on high-frequency pixel artifacts, making them inherently resistant to smoothing-based defenses.

Channel condition independence. Across defenses, the gap between 20 dB and 0 dB Pure UAP ASR remains small (within 8 percentage points in all cases), consistent with the channel robustness observed in Table I and confirming that defense effectiveness does not vary substantially with channel quality.

VI. Conclusion

This paper investigated the vulnerability of VLP-based SemCom systems to UAPs under realistic wireless transmission conditions. Unlike existing channel-blind attacks, SemCom-UAP explicitly incorporates channel characteristics into the perturbation generation and optimization process, enabling a single perturbation to remain effective across different SNR regimes. SemCom-UAP combines channel-aware perturbation generation with stratified SNR training and adaptive surrogate coordination to improve both attack transferability and robustness against channel-induced embedding corruption. Extensive experiments demonstrated that channel effects significantly influence adversarial effectiveness, and that modeling such effects during optimization substantially improves attack performance across multiple VLP architectures and retrieval tasks. These findings expose critical security risks in cross-modal SemCom and motivate defenses that jointly address semantic alignment, wireless dynamics, and adversarial robustness.

VII. Acknowledgement

This work was supported in part by the National Science Foundation under Grants CNS-2153358, CNS-2245918, OAC-2413654, DoW CoE-AIML under Grant W911NF-20-2-0277, DoW Griffiss Institute under Grant SA10012022030485, and by the Commonwealth Cyber Initiative (CCI) and COVA CCI.

References

[1] J. Peng, H. Xing, L. Xu, S. Luo, L. Ale, Z. Zhu, and Z. Xiao, "Semaugcom: Semantic augmented communication for few-shot learning tasks," *IEEE Transactions on Network Science and Engineering*, vol. 13, pp. 10 083–10 100, 2026.

[2] F. Jiang, C. Tang, L. Dong, K. Wang, K. Yang, and C. Pan, "Visual language model based cross-modal semantic communication systems," *IEEE Transactions on Wireless Communications*, 2025.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[5] V. Khrulkov and I. Oseledets, "Art of singular vectors and universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8562–8570.

[6] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4422–4431.

[7] M. M. Naseer, S. H. Khan, M. H. Khan, F. Shahbaz Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] J. Zhang, Q. Yi, and J. Sang, "Towards adversarial attack on vision-language pre-training models," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5005–5013.

[9] D. Lu, Z. Wang, T. Wang, W. Guan, H. Gao, and F. Zheng, "Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 102–111.

[10] H. Wang, K. Dong, Z. Zhu, H. Qin, A. Liu, X. Fang, J. Wang, and X. Liu, "Transferable multimodal attack on vision-language pre-training models," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 1722–1740.

[11] P.-F. Zhang, Z. Huang, and G. Bai, "Universal adversarial perturbations for vision-language pre-trained models," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 862–871.

[12] H. Fang, J. Kong, W. Yu, B. Chen, J. Li, H. Wu, S.-T. Xia, and K. Xu, "One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4090–4100.

[13] Y. E. Sagduyu, T. Erpek, S. Ulukus, and A. Yener, "Is semantic communication secure? a tale of multi-domain adversarial attacks," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 50–55, 2023.

[14] J. Peng, H. Xing, X. Chen, Y. Li, Y. Cui, D. Zheng, L. Ale, and L. Feng, "Security enhanced computation offloading for collaborative inference at semantic-communication-empowered edge," *IEEE Transactions on Mobile Computing*, vol. 24, no. 9, pp. 8071–8088, 2025.

[15] D. Kozlov, M. Mirmohseni, and R. Tafazolli, "Secure semantic communication over wiretap channel," in *2024 IEEE Information Theory Workshop (ITW)*. IEEE, 2024, pp. 37–42.

[16] Y. Li, Z. Shi, H. Hu, Y. Fu, H. Wang, and H. Lei, "Secure semantic communications: From perspective of physical layer security," *IEEE Communications Letters*, vol. 28, no. 10, pp. 2243–2247, 2024.

[17] W. Chen, Q. Yang, S. Shao, Z. Shi, J. Chen, and X. Shen, "Can knowledge improve security? a coding-enhanced jamming approach for semantic communication," *IEEE Journal on Selected Areas in Communications*, 2025.

[18] Z. Yang, M. Chen, G. Li, Y. Yang, and Z. Zhang, "Secure semantic communications: Fundamentals and challenges," *IEEE network*, vol. 38, no. 6, pp. 513–520, 2024.

[19] M. Isakov, V. Gadepally, K. M. Gettings, and M. A. Kinsy, "Survey of attacks and defenses on edge-deployed neural networks," in *2019 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2019, pp. 1–8.

[20] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[21] A. Anjum, A. Mitra, P. Agbaje, M. A. Alam, D. Roy, M. S. Parwez, and H. Olufowobi, "Semperge: Unveiling text-based adversarial attacks on semantic communication," in *2025 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2025, pp. 1–9.

[22] Z. Wang, Y. Du, R. Ning, D. Takabi, and L. Li, "Ecsc: Energy-calibrated semantic communication," in *MILCOM 2025*. IEEE, 2025, pp. 1359–1364.

[23] H. Lu, Y. Yu, S. Xia, Y. Yang, D. Rajan, B. P. Ng, A. Kot, and X. Jiang, "From pretrain to pain: Adversarial vulnerability of video foundation models without task knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 9, 2026, pp. 7548–7556.

[24] H. Huang, S. M. Erfani, Y. Li, X. Ma, and J. Bailey, "X-transfer attacks: Towards super transferable adversarial attacks on CLIP," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 13–19 Jul 2025, pp. 25 204–25 234.

[25] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[26] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014.

[27] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[28] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 671–15 680.

[29] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, 2022, pp. 12 888–12 900.

[30] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022, pp. 25 994–26 009.