TarGSC: Text-Parameterized Autoregressive Generative Semantic Communication

Anonymous ICME submission

***Abstract*—Current Generative Semantic Communication (GSC) typically relies on conditional diffusion models for generation, facing two key challenges: vulnerability of text semantics under channel noise and low inference efficiency. Therefore, we propose TarGSC, an end-to-end semantic multimodal generative communication framework with a Text2Para encoder and an efficient Autoregressive (AR) model. At the transmitter, Text2Para maps text into six separate distribution parameters; then a feature encoder extracts compact semantic representations from images—both transmitted over an Additive White Gaussian Noise (AWGN) channel. At the receiver, the AR model conditions on image features and text parameters via cross-modal conditioning to enable fast, high-fidelity image generation. Experiments on common datasets demonstrate that TarGSC is superior across a wide range of SNRs and channel bandwidth ratios. Notably, at SNR = -13dB, it achieves a PSNR of 23.2 with an average generation time of 0.537 s, delivering an order-of-magnitude efficiency improvement over diffusion-based GSC methods.**

***Index Terms*—Generative Semantic Communication, Multimodal Semantic Transmission, Visual Autoregressive Model, Image Generation**

I. Introduction

With the rapid advancement of diffusion models, Generative Semantic Communication (GSC) has emerged as a prominent research direction in semantic communication [1], [2]. Unlike methods focusing on pixel-level information [3], [4], a diffusion-driven semantic communication framework with bandwidth constraints leverage VAEs to compress high-dimensional features into a latent space prior to diffusion-based generation [5]; SGD-JSCC [6], which guides diffusion denoising via text and edge map conditioning. While these methods significantly reduce channel bandwidth demands, their reliance on diffusion-based generation results in low inference efficiency [7].

Meanwhile, generative methods are highly dependent on prompt description quality. Both Gao et al. and Zhang et al. require text as one of the conditions to guide diffusion denoising, but text semantics are susceptible to channel perturbations and sensitive to numerical precision [6], [8]. Thus, how to represent text semantics in a compact and robust manner while ensuring generation quality remains a pressing open problem in GSC [9], [10].

This paper proposes **TarJSCC**, a GSC method based on semantic multimodal encoding, relying on two core insights: First, the multi-scale visual autoregressive model by Tian et al. achieves significantly faster image generation than diffusion models with comparable quality—an observation that inspires our work [11]. Our TarGSC model leverages depth features as structural conditional inputs to achieve fast, high-quality semantic generation with low computational overhead. Second, to address the limitations of existing methods, which suffer from redundant text semantic representations, high communication overhead, and vulnerability to channel noise, we design Text2Para, a module for discretized text distribution representation. Taking text vectors as input, Text2Para uses a multi-layer perceptron to discretize continuous semantics into distribution parameters, reducing transmission bandwidth while enhancing the robustness of text conditions against channel noise. TarJSCC conditions the SCVAR model on deep features and Text2Para-mapped text parameters to achieve efficient, robust image generation.

In summary, our contributions are as follows:

1. We propose the TarGSC architecture, integrating compressed depth representations with Text2Para-output hyperparameters to enable high-quality, efficient image generation via the SCVAR module.

2. Propose the Text2Para encoder, which maps textual information to hyperparameter representations, ensuring strong robustness over noisy channels.

II. Related work

Generative semantic communication has garnered widespread attention as a novel paradigm in recent years, with its core tenet shifting the communication objective from traditional signal reconstruction to semantics-aware content generation. Qin et al. proposed a systematic modeling framework for diffusion models in semantic communication and summarized their performance merits under low-SNR conditions [12]. Furthermore, Wang et al. integrated stochastic differential equation theory with practical 6G system design by leveraging diffusion models [13]. These studies demonstrate that diffusion models empower semantic communication to overcome the performance bottleneck of conventional discriminative JSCC frameworks.

Tian et al. proposed a novel generation paradigm—the Visual Autoregressive (VAR) model—which achieves a deterministic and efficient generation process via hierarchical autoregressive prediction [11]. It ensures high generation quality while achieving substantial reduction in inference complexity compared to diffusion models. Thanks to its deterministic decoding mechanism and low-latency properties, VAR is naturally suited for generative semantic communication scenarios, where semantic consistency and real-time performance are core priorities.

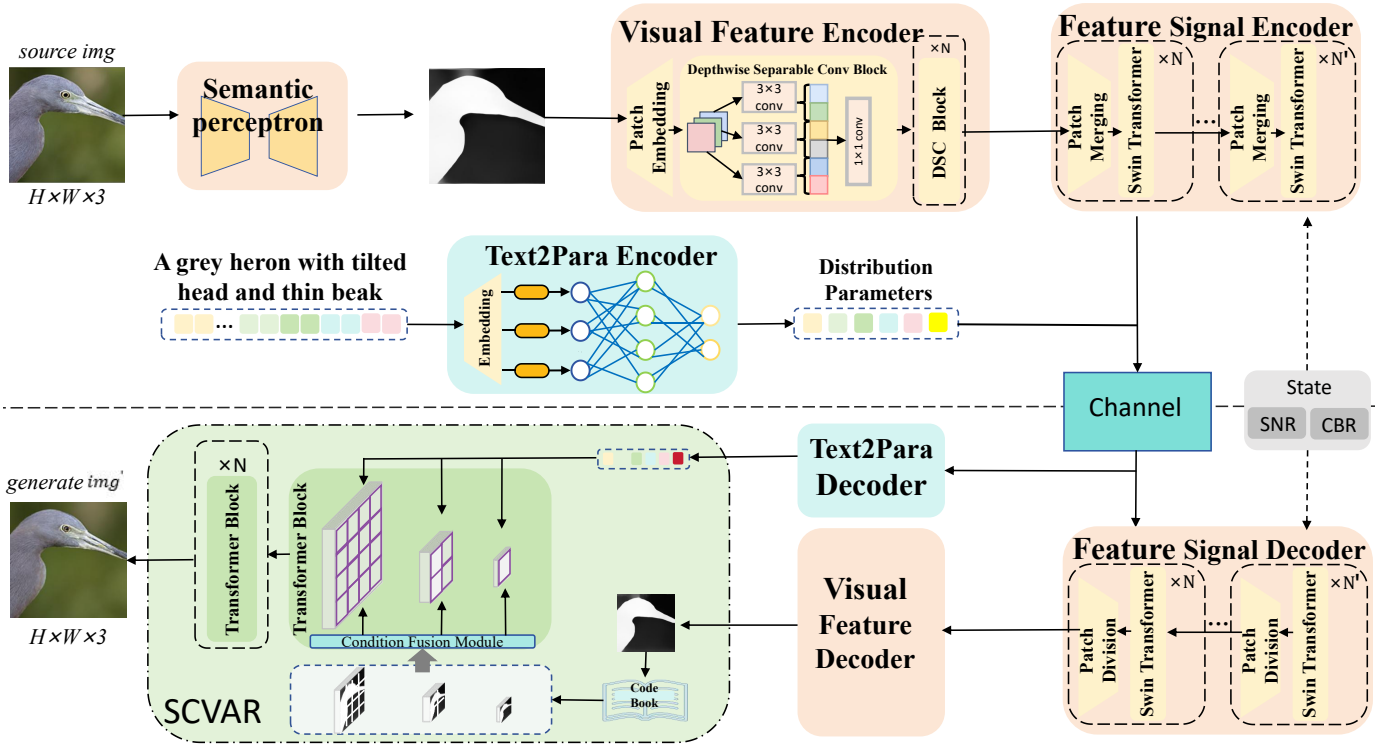

Fig. 1: Schematic of the TarGSC framework. Text is encoded into low-dimensional parameters via the Text2Para encoder, while images have semantic features extracted by the visual feature encoder—both are transmitted over the channel. At the receiver, the SCVAR module is employed for conditional decoding to generate high-quality images.

In generative semantic communication, text is typically mapped to semantic vectors via pre-trained encoders (e.g., CLIP [14], BLIP [15]. Alternatively, recent works have designed task-specific encoders to extract semantic information more efficiently in communication scenarios, enhancing semantic representation quality [16]. However, the spatial compactness of semantic vectors renders them highly susceptible to noise. Drawing on AdaLN, text vectors can be further mapped to modulation parameters of the generative network, dynamically adjusting feature distributions for more robust semantic expression [17].

III. Method

This paper proposes **TarGSC**, an end-to-end semantic multimodal transmission and generation framework, as shown in Fig. 1. At the transmitter, the original image first undergoes a semantic-aware module to extract depth information, followed by a visual feature extraction module for further semantic feature extraction; text is mapped into a discrete parameterized representation via the Text2Para module. Image features and text parameters are jointly transmitted over the channel. At the receiver, the received semantic features and text parameters are decoded separately: image features are reconstructed into depth information via a visual decoding module, which—together with text parameters—serves as conditional inputs to the SCVAR network for fast, high-quality image generation.

A. Sender

Let $f_{\text{sem}}(\cdot)$, $f_v(\cdot)$, $f_c(\cdot)$, and $f_{\text{T2P}}(\cdot)$ denote the semantic-aware module, visual feature extraction module, feature encoding module, and text parameter mapping module, respectively.

Given the original image $\text{img} \in \mathbb{R}^{H\times W\times 3}$ at the transmitter.First, it is fed into the semantic-aware module $f_{\text{sem}}(\cdot)$ to extract a structural representation, yielding a depth map $d \in \mathbb{R}^{H\times W\times 1}$ that explicitly characterizes the scene's geometric semantics. Next, d is mapped via the visual feature extraction module $f_v(\cdot)$ to a semantic feature $F = f_v(d)$ for high-level representation learning. F is then encoded by the feature encoding module $f_c(\cdot)$ into the transmit feature signal $X_d = f_c(F)$. Meanwhile, textual information t—one of the generation conditions—is encoded via Text2Para into a low-dimensional parameter representation $X_t = f_{\text{T2P}}(t)$, where X_t serves as the transmit text semantic signal. This design ensures high robustness of text conditions over noisy channels.

B. Channel Status

To simulate unavoidable noise and fading in real-world wireless environments, we adopt the Additive White Gaussian Noise (AWGN) channel as the transmission model. Let X_d and X_t denote the transmitter's output feature signal and text parameters, respectively. The received signal over the channel is given by:

$$Y = X + n, \tag{1}$$

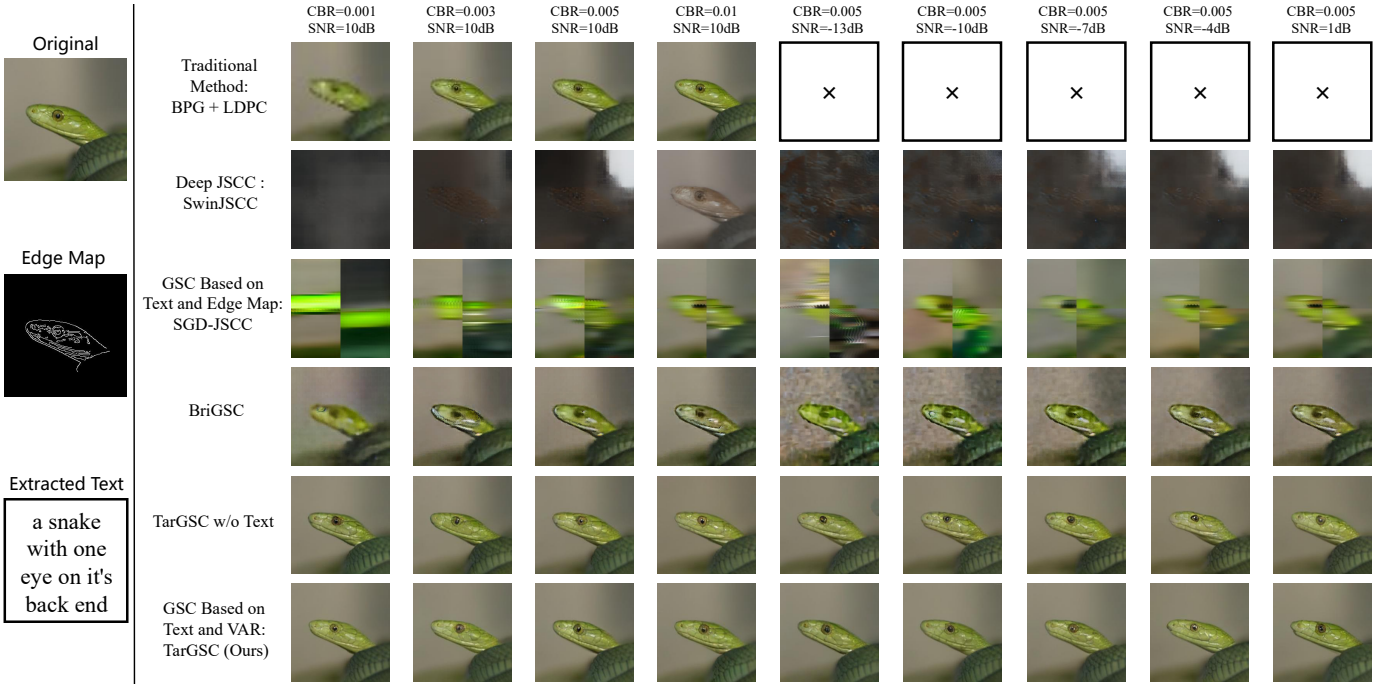

Fig. 2: On ImageNet, we compare semantic image transmission results under varying SNR and channel bandwidth ratio (CBR) conditions. All images are uniformly cropped or resampled to 256×256. Notably, at low SNR, LDPC coding may suffer decoding failure—unsuccessfully decoded images or texts are marked with "×" in the figures.

where $X \in \{X_d, X_t\}$ and n follows a Gaussian distribution with zero mean and variance σ^2. Variations in σ^2 correspond to different levels of channel degradation. The noise variance is determined by the specified signal-to-noise ratio (SNR):

$$\sigma^2 = \frac{P_X}{10^{\text{SNR}/10}}, \tag{2}$$

where P_X is the transmitted signal power. Adjusting SNR simulates different channel qualities, enabling evaluation of the system's robustness under bandwidth-constrained and noisy conditions.

C. Receiver

At the receiver, we adopt SCVAR as the core generative model, which enables fast, high-quality image reconstruction via its scale decoupling and alignment mechanism. As a multi-scale visual autoregressive model, SCVAR's key insight is to decouple image representations into multiple scale sequences while preserving the image's 2D structure, performing autoregressive prediction at the scale level to enable fast and stable conditional generation [18]. Its basic workflow is summarized as follows:

1. SCVAR uses a multi-scale VQVAE to quantize the input image into multi-scale discrete representations

$$r = \{r_1, r_2, \ldots, r_n\}, \tag{3}$$

aligning with visual conditions at each scale [19].

2. The aligned conditional embeddings are combined via linear interpolation to form the initial control signal:

$$f_r = \text{LN}(\alpha r + (1-\alpha) r_{\text{cond}}), \tag{4}$$

where α denotes a control coefficient.

3. SCVAR then performs autoregressive prediction on the multi-scale sequences, updating features via two parallel branches:

$$f_{\text{CA}}(r, r_{\text{cond}}) = r + \text{FFN}(\text{Cross-Attn}(r, r_{\text{cond}})), \tag{5}$$

$$f_{\text{SA}}(r) = r + \text{FFN}(\text{Self-Attn}(r)), \tag{6}$$

where Cross-Attn enables conditional alignment and Self-Attn models 2D context.

4. Finally, the model predicts the next scale as (conditioned on generated low-scale representations)

$$p(r_i \mid r_{<i}, r_{\text{cond},<i}), \tag{7}$$

and reconstructs the full image scale-by-scale, enabling fast, stable, and accurate conditional generation.

Let $f_v^{-1}(\cdot)$, $f_c^{-1}(\cdot)$, and $f_{\text{T2P}}^{-1}(\cdot)$ denote the feature restoration module, feature decoding module, and Text2Para decoder, respectively.

At the receiver, the feature signal $\hat{X}_d$ and text hyperparameter signal $\hat{X}_t$ are received, and the feature signal and text hyperparameter signal are sent to their corresponding decoding modules for semantic recovery. Among them, $\hat{X}_d$ passes through the feature decoding module $f_c^{-1}(\cdot)$ to obtain the semantic feature $\hat{F} = f_c^{-1}(\hat{X}_d)$; $\hat{F}$ then passes through the feature restoration module $f_v^{-1}(\cdot)$ to get the reconstructed depth map d'; $\hat{X}_t$ passes through the Text2Para decoder $f_{\text{T2P}}^{-1}(\cdot)$ to obtain the text hyperparameter $X_t' = f_{\text{T2P}}^{-1}(\hat{X}_t)$. We input d' and X_t' into the SCVAR network to provide structural

constraints and semantic conditions, thereby generating the target image

$$\text{img}' = \text{SCVAR}(d', X_t') \tag{8}$$

with consistent content and accurate semantics.

IV. Experimental Settings And Results

A. Experimental Settings

Datasets: Our work employs the ImageNet-1K [20], COCO [21], and Kodak [22] datasets: ImageNet-1K consists of 1,000 classes of natural images; COCO comprises complex scenes with multiple objects; Kodak contains 24 high-quality natural images.

Model Details: The depth-aware module employs a pre-trained U-Net with frozen parameters. The visual feature encoder comprises 4 Depthwise Separable Convolution blocks with sequential channel dimensions 32, 64, 128, and 256, while the corresponding decoder features a symmetric up-sampling architecture with inverse channel dimensions 256, 128, 64, and 1 [23]. The Text2Para encoder is a three-layer MLP (256, 128, 64) outputting a 6-dimensional parameter representation, and its decoder is a two-layer MLP with residual connections (16, 32) for parameter denoising. For the feature signal encoder-decoder, we adopt SwinJSCC [24], and the channel model is Additive White Gaussian Noise (AWGN) [13], [25], [26].

Comparison Algorithms: We compare against BriGSC [8], SwinJSCC [24], SGD-JSCC [6], and BPG+LDPC [27]–[29]. For fair comparison under similar compression ratios, SwinJSCC [24] and SGD-JSCC [6] perform downsampling before transmission and upsampling after reception. The provided pre-trained weights are fine-tuned following the authors' recommended parameters.

Evaluation Metrics: We adopt FID [30], LPIPS [31], and MLLMP to evaluate semantic image similarity from multiple perspectives. Specifically, FID measures feature distribution consistency between generated and real images, LPIPS focuses on perceptual visual similarity, and MLLMP assesses the global semantic alignment between reconstructed and original images. For MLLMP evaluation, we use the multimodal large language model GPT-5 [32] to determine which method generates images closest to the original.

B. Result Analysis

As shown in Fig. 2, Visual comparison of semantic image transmission results on ImageNet. Channel Bandwidth Ratio (CBR) denotes the dimension ratio between encoded data and original image data. All images are uniformly resized to 256×256 for comparison.

At a fixed SNR of 10 dB, BPG+LDPC, SGD-JSCC, and BriGSC exhibit improved generation quality with increasing channel bandwidth ratio (CBR). However, when SNR drops to negative ranges: BPG+LDPC fails to decode images; SGD-JSCC generates images with prominent noise artifacts; while BriGSC outperforms SGD-JSCC under negative SNR, it still

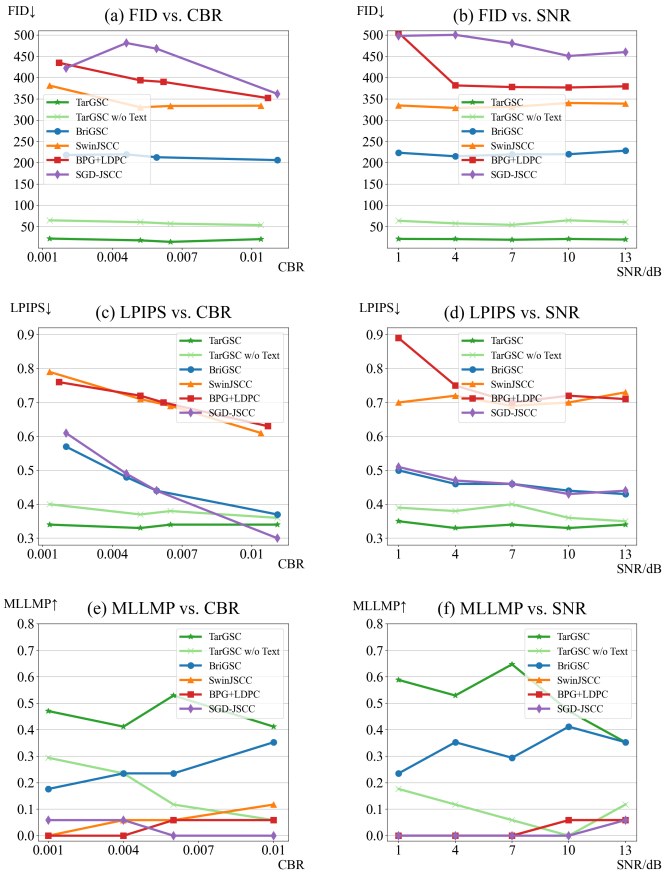

Fig. 3: On the ImageNet dataset, we show the trend curves of FID, LPIPS, and MLLMP metrics as they change with the signal-to-noise ratio (SNR) or channel bandwidth ratio (CBR). Among them, under the condition of a fixed CBR of 0.006, the SNR is adjusted as the bandwidth condition changes; while under the condition of a fixed SNR of 10 dB, the CBR changes with the signal-to-noise ratio.

suffers from noticeable noise. For SwinJSCC, generation performance is poor at extremely low bandwidths, only presenting basic outlines when bandwidth improves—indicating it struggles to fully transmit high-dimensional visual features in bandwidth-constrained scenarios. In contrast, our proposed TarGSC demonstrates significant robustness under negative SNR. Even TarGSC w/o Text (lacking text parameter modulation) generates images with low noise, with only occasional minor noise artifacts or slight distortions at extremely low SNR. The full TarGSC leverages text parameters and depth map conditional modulation, maintaining high semantic fidelity and stable visual quality across varying SNR and CBR conditions. This highlights the method's advantage in reliable transmission and generation of semantic information in complex channel environments.

Fig. 3 presents the combined quantitative results of TarGSC, its ablated variant (TarGSC w/o Text, which removes textual parameter modulation), and four baseline methods (BPG+LDPC, SwinJSCC, BriGSC, and SGD-JSCC) on Ima-

geNet, evaluated via FID, LPIPS, and MLLMP across diverse CBR and SNR settings. Overall, TarGSC exhibits consistently robust performance under all tested conditions, with coherent trends observed across the integrated subpanels. For the key metrics FID and LPIPS, TarGSC maintains substantially lower scores than all baselines across a wide range of CBR and SNR values. Even under extremely constrained operating conditions (e.g., CBR = 0.001 and SNR = 1 dB), TarGSC achieves an FID below 30 and an LPIPS below 0.35.

Its ablated variant, TarGSC w/o Text, while also outperforming all baseline methods (FID $\approx$ 50 and LPIPS $\approx$ 0.38–0.40), exhibits slightly degraded performance relative to the full TarGSC model. This direct comparison highlights the performance gains introduced by textual parameter modulation. By contrast, baseline methods lag significantly behind, including BPG+LDPC (FID $\approx$ 430–500 and LPIPS $\approx$ 0.8–1.0), SwinJSCC (FID $\approx$ 330–400 and LPIPS $\approx$ 0.7–0.9), and BriGSC (FID $\approx$ 210–250 and LPIPS $\approx$ 0.5–0.6). A minor exception is observed at higher CBR values (e.g., CBR $\geq$ 0.01), where TarGSC yields a slightly higher LPIPS score than SGD-JSCC ($\approx$ 0.30 vs. $\approx$ 0.28). This behavior indicates a mild trade-off between semantic consistency and fine-grained perceptual fidelity when sufficient bandwidth is available.

For the semantic consistency metric MLLMP, TarGSC consistently outperforms all baselines across all evaluated settings, achieving an overall improvement of approximately 15–30%. Meanwhile, TarGSC w/o Text exhibits an approximately 10–15% reduction in MLLMP relative to TarGSC, although it still outperforms most baseline methods. This result further validates the critical role of textual modulation in preserving semantic consistency. These advantages stem from two core design choices of TarGSC: leveraging depth maps as stable structural priors to preserve scene layouts and geometric relationships under noisy channel conditions, and compressing textual information into compact low-dimensional parameters followed by conditional modulation via the SCVAR module. The ablation results of TarGSC w/o Text directly confirm that textual modulation is a key contributor to the observed gains in both perceptual quality and semantic consistency.

TABLE I: Quantitative Analysis of PSNR for Image Generation Across Different Methods and Text Channel Noise Intensities.

Method \ SNR	10 dB	4 dB	1 dB	-4 dB	-7 dB	-13 dB
SGD-JSCC	23.6	17.2	13.1	12.3	11.0	10.4
BriGSC	24.4	18.3	13.5	12.1	9.8	9.5
TarGSC(Ours)	**32.3**	**28.5**	**26.8**	**23.8**	**24.1**	**23.2**

Generative semantic communication leverages powerful generative models to mitigate noise impact on visual features to some extent, yet its generation process relies heavily on text conditional distributions and is highly sensitive to text perturbations. As shown in Table I, TarGSC significantly enhances text robustness during channel transmission by mapping text to low-dimensional parameters and integrating them into the VAR framework for conditional modulation. In contrast, traditional methods typically require higher transmission redundancy or more robust error correction mechanisms to ensure accurate text recovery, incurring additional bandwidth and computational overhead.

TABLE II: Comparison of average transmission time per image on ImageNet for different generative models.

Method	TarGSC	TarGSC w/o Text	BriGSC	SGD-JSCC
Time (s)	0.537	**0.494**	15.224	10.1

We further assess the efficiency of all methods. As shown in Table II, TarGSC leverages a VAR-based autoregressive generation mechanism that exhibits causality and determinism, requiring only a single forward pass for image generation. This eliminates the computational overhead associated with extensive iterative denoising steps in diffusion models. Additionally, the integrated SCVAR module enables conditional modulation of the generation process via low-dimensional text parameters, enhancing semantic controllability and generation quality without introducing extra sampling complexity. Thus, TarGSC achieves significantly higher generation efficiency than diffusion-based GSC methods while preserving high-quality generation.

V. Conclusion

Our proposed TarGSC framework exhibits remarkable advantages in semantic multimodal transmission and generation. Across varying channel bandwidth ratios (CBR) and signal-to-noise ratios (SNR), TarGSC stably generates high-quality images, outperforming diffusion-based GSC methods in semantic fidelity while achieving substantially improved generation efficiency. Its strengths stem from the SCVAR module: integrating depth map guidance to enhance structural preservation and leveraging text parameter modulation for fine-grained generation control. However, we do not conduct systematic ablation studies on text parameter dimensionality (e.g., 4 or 5 dimensions), primarily due to the high cost of large-scale training. Future work will investigate the relationship between parameter dimensionality and generation performance to further optimize textual conditional representations.

References

[1] A. D. Raha, M. S. Munir, A. Adhikary, Y. Qiao, and C. S. Hong, "Generative ai-driven semantic communication framework for nextg wireless network,"

[2] X. Liu, M. B. Mashhadi, L. Qiao, Y. Ma, R. Tafazolli, and M. Bennis, "Diffusion-based generative multicasting with intent-aware semantic decomposition," *arXiv preprint arXiv:2411.02334*, 2024.

[3] M. A. Jarrahi, E. Bourtsoulatze, and V. Abolghasemi, "Joint source-channel coding for wireless image transmission: A deep compressed-sensing based method," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2024.

[4] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[5] L. Guo, W. Chen, Y. Sun, B. Ai, N. Pappas, and T. Quek, "Diffusion-driven semantic communication for generative models with bandwidth constraints," *IEEE Transactions on Wireless Communications*, 2025.

[6] M. Zhang, H. Wu, G. Zhu, R. Jin, X. Chen, and D. Gündüz, "Semantics-guided diffusion for deep joint source-channel coding in wireless image transmission," *arXiv preprint arXiv:2501.01138*, 2025.

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[8] D. Gao, Y. Yi, M. Yang, J. Li, D. Liu, and W. Xu, "Bridging semantic scale gaps in image transmission through multi-scale joint perception and generation," *IEEE Wireless Communications Letters*, vol. 14, no. 10, pp. 3314–3318, 2025.

[9] L. Jiang, N. Hassanpour, M. Salameh, M. S. Singamsetti, F. Sun, W. Lu, and D. Niu, "Frap: Faithful and realistic text-to-image generation with adaptive prompt weighting," *Transactions on Machine Learning Research*.

[10] R. Wang, Z. Chen, C. Chen, J. Ma, H. Lu, and X. Lin, "Compositional text-to-image synthesis with attention map control of diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 5544–5552, 2024.

[11] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *Advances in neural information processing systems*, vol. 37, pp. 84839–84865, 2024.

[12] H.-L. Qin, J. Dai, G. Lu, S. Shao, S. Wang, T. Xu, W. Zhang, P. Zhang, and K. B. Letaief, "Generative ai meets 6g and beyond: Diffusion models for semantic communications," *arXiv preprint arXiv:2511.08416*, 2025.

[13] X. Wang, H. Jia, and N. Cheng, "Latent diffusion model based denoising receiver for 6g semantic communication: From stochastic differential theory to application," *arXiv preprint arXiv:2506.05710*, 2025.

[14] 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*, pp. 8748–8763, PmLR, 2021.

[15] 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*, pp. 12888–12900, PMLR, 2022.

[16] Y. Liu, S. Jiang, Y. Zhang, K. Cao, L. Zhou, B.-C. Seet, H. Zhao, and J. Wei, "Extended context-based semantic communication system for text transmission," *Digital Communications and Networks*, vol. 10, no. 3, pp. 568–576, 2024.

[17] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

[18] F. Xie, D. Gao, R. Liu, M. Yang, Y. Zhang, and W. Wang, "Stable control visual autoregressive model: Precise and efficient image generation via scale alignment," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[19] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[20] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[22] R. Franzen, "Kodak lossless true color image suite." http://r0k.us/graphics/kodak/, 2010.

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications,"

[24] K. Yang, S. Wang, J. Dai, X. Qin, K. Niu, and P. Zhang, "Swinjscc: Taming swin transformer for deep joint source-channel coding," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[25] H. Rezaei, T. Sivalingam, and N. Rajatheva, "Automatic and flexible transmission of semantic map images using polar codes for end-to-end semantic-based communication systems," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, 2023.

[26] K. Zhang, L. Li, W. Lin, Y. Yan, R. Li, W. Cheng, and Z. Han, "Semantic successive refinement: a generative ai-aided semantic communication framework," *IEEE Transactions on Cognitive Communications and Networking*, 2025.

[27] F. Bellard, "Bpg image format." https://bellard.org/bpg/, 2014. Accessed: 2025-12.

[28] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.

[29] Y. Liu, Y. Wang, and X. Zhang, "Wireless image transmission with bpg and ldpc codes," in *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2016.

[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6629–6640, 2017.

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[32] OpenAI, "Gpt-5 system card." https://openai.com/blog/gpt-5-system-card/, 2025. Accessed: 2025-12.