2026 14th International Conference on Information and Communication Networks (ICICN) | 979-8-3195-3689-1/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICICN71760.2026.11745078

HiPrompt-JSCC: Hierarchical Prompt-Aided Joint Source-Channel Coding for Task-Adaptive Semantic Communication

Shuai Zhang[1], Bo Chen[1], Dongxuan He[1,4], Xiang Xie[1,4], Huachun Tan[2,4,*], Yanan Zhao[3,4,*]
[1] *School of Information and Electronics, Beijing Institute of Technology, Beijing, China;*
[2] *School of Interdisciplinary Science, Beijing Institute of Technology, Beijing, China;*
[3] *School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China;*
[4] *Beijing Institute of Technology, Zhuhai, China*
{shuaizhang, bochen, dongxuan_he, xiexiang, tanhc, zyn}@bit.edu.cn
* Corresponding authors: Huachun Tan and Yanan Zhao

***Abstract*—Deep joint source-channel coding (JSCC) for images is commonly trained to optimize reconstruction quality, whereas receiver-side applications are evaluated by classification, detection, or segmentation accuracy. This paper presents HiPrompt-JSCC, a task-adaptation framework that retains a frozen Swin-based JSCC backbone and optimizes a task-specific prompt adapter. The adapter fuses visual features with a learned task embedding and supplies stage-dependent key/value prompts to window attention. At 1 dB SNR under AWGN, HiPrompt-JSCC improves ImageNet100 top-1 accuracy by 7.82 percentage points over the pretrained reconstruction-oriented SwinJSCC baseline. Across the evaluated COCO operating points, it improves detection AP by 0.48–0.75 and instance-segmentation mask AP by 0.92–1.38. Task adaptation is performed exclusively through transmitter-side encoder prompts; neither the communication backbone nor the JSCC receiver is modified.**

***Index Terms*—deep joint source-channel coding, semantic communications, prompt learning, task-oriented transmission, wireless image transmission**

I. Introduction

Semantic communication aims to preserve the information required by a receiver task rather than reproduce every source bit [1]–[4]. In wireless image transmission, the reconstructed image may serve as an intermediate input to a classifier, detector, or segmentation model. Consequently, reconstruction quality alone is insufficient to characterize task performance.

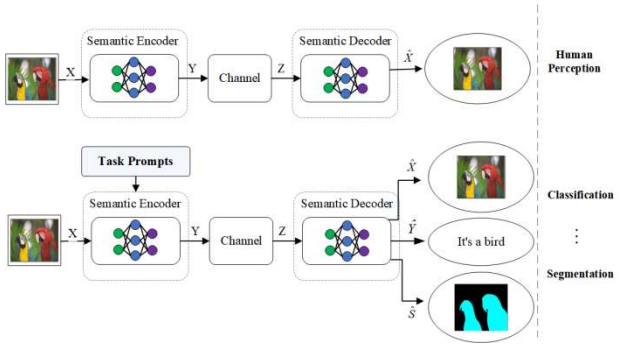

Fig. 1. Task adaptation in semantic image transmission. A conventional reconstruction-oriented system (top) is compared with a shared backbone conditioned by task-specific prompts (bottom).

Related Work and Positioning. Task-driven visual compression preserves features needed by downstream models. Deep feature compression, for example, reduces device-server traffic for collaborative object detection [5], whereas learned image codecs improve rate-distortion performance through hyperpriors, autoregressive priors, and attention [9]–[11]. These digital schemes generally target fixed tasks and do not learn robustness to noisy wireless channels.

Task-oriented semantic communication [1]–[3] and neural codecs such as DeepJSCC [6], WITT [7], and SwinJSCC [8] optimize end-to-end transmission, but their image objectives remain predominantly reconstruction oriented. Drawing on parameter-efficient prompt tuning [12]–[14], HiPrompt-JSCC freezes the channel-robust encoder and decoder, injects hierarchical prompts only at the transmitter, and preserves the transmitted-symbol interface and receiver. Each task therefore requires only a lightweight adapter; Fig. 1 illustrates this distinction.

The main contributions are as follows:

- We develop a transmitter-side adaptation scheme that equips a frozen SwinJSCC backbone with task-specific prompt modules.
- We design a hierarchical generator that combines visual features with a task embedding and injects stage-specific key/value prompts into W-MSA and SW-MSA.
- We evaluate ImageNet100 classification under AWGN and Rayleigh fading and COCO detection and instance segmentation under AWGN, including prompt-stage ablation and complexity analysis.

II. System Model

Consider task-oriented wireless image transmission [1], [15]. Let $\mathbf{x}$ and $\mathbf{t}$ denote the image and task, and let CBR denote the channel bandwidth ratio. The transmitter computes

$$\mathbf{y} = f_\theta(\mathbf{x}, \mathrm{CBR}; \{\mathbf{p}_i\}_{i=1}^{4}),$$

where f_θ is the JSCC encoder and $\{\mathbf{p}_i\}_{i=1}^{4}$ are task-specific encoder prompts. The backbone is conditioned on CBR, while

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 25,2026 at 14:36:45 UTC from IEEE Xplore. Restrictions apply.

$\mathbf{p}_i = \mathcal{P}_i(\mathbf{f}_v^{(i)}, \mathbf{e}_t)$ depends only on image features and the selected task embedding.

For either evaluated channel model, AWGN or Rayleigh fading, let $\mathcal{W}$ denote the channel at SNR γ. The received latent is

$$\hat{\mathbf{y}} = \mathcal{W}_\gamma(\mathbf{y}),$$
$$\hat{\mathbf{x}} = g_\phi(\hat{\mathbf{y}}, \mathrm{CBR}),$$

and the task network produces

$$\hat{\mathbf{z}} = h_{\psi_\mathbf{t}}(\hat{\mathbf{x}}),$$

where $h_{\psi_\mathbf{t}}$ is the selected classifier, detector, or segmenter and $\hat{\mathbf{z}}$ is its prediction.

During adaptation, the pretrained encoder, decoder, and channel model remain fixed. Trainable parameters are limited to the task embeddings, prompt generator, stage-wise projectors, and scalar gates. Classification is optimized using cross-entropy and reconstruction regularization, whereas detection and instance segmentation use feature matching from frozen R-50-FPN networks. COCO AP is used only for evaluation. The complete architecture is shown in Fig. 2.

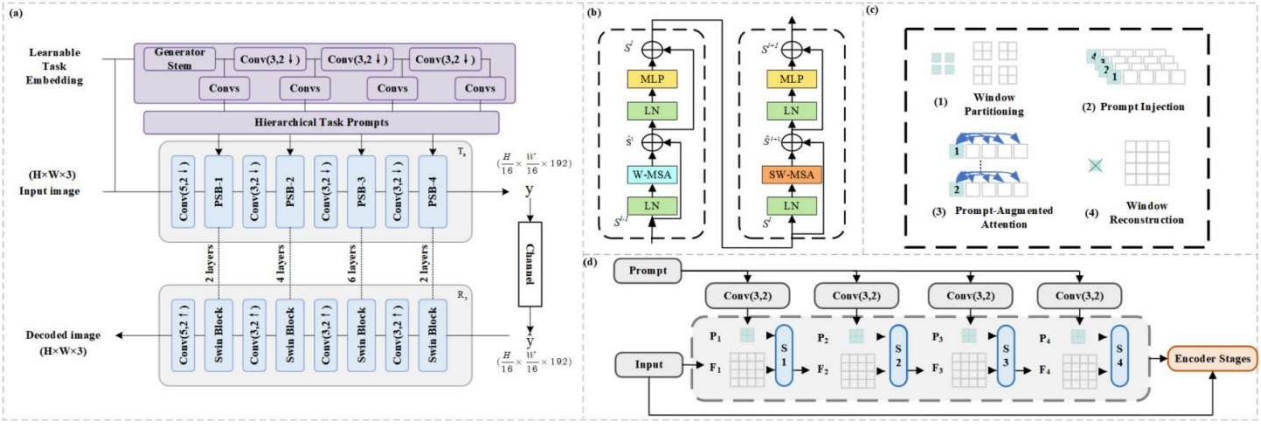

Fig. 2. HiPrompt-JSCC architecture. (a) Task embeddings and visual features generate hierarchical prompts for the JSCC encoder. (b) Prompted Swin blocks retain the W-MSA/SW-MSA sequence. (c) Prompt tokens enter window attention as key/value context. (d) Stage-specific prompts are generated at four encoder resolutions.

III. Proposed HiPrompt-JSCC Framework

HiPrompt-JSCC retains the pretrained Swin-based JSCC backbone and introduces trainable task embeddings, a visual prompt generator, and stage-wise prompt projectors at the transmitter.

A. Task Embedding and Prompt Conditioning

In Fig. 2(a), classification, detection, and instance segmentation are represented by three 64-dimensional embeddings, of which only the embedding associated with the current task is active. Four learned scalar gates control the prompt contribution at the encoder stages. The embeddings and gates contain 196 parameters, as listed in Table I. For the selected embedding $\mathbf{e}_t$, the task code is

$$\mathbf{e} = \mathcal{F}(\mathbf{e}_t),$$

where $\mathcal{F}(\cdot)$ is the task-conditioning transform. The conditioned representation is combined with visual features to produce encoder prompts at each stage.

B. Hierarchical Prompt Generation

The prompt-generation path is shown in Fig. 2(d). A three-layer visual stem with convolution strides (2,2,1) extracts $\mathbf{f}_v$, which is subsequently downsampled to the four encoder resolutions. Let $\mathbf{f}_v^{(i)}$ denote the feature map for stage i. The corresponding prompt is

$$\mathbf{p}_i = \mathcal{P}_i(\mathbf{f}_v^{(i)}, \mathbf{e}),$$

where $\mathcal{P}_i(\cdot)$ is the prompt head for stage i. At stage i, $\mathbf{p}_i$ is projected into 16 block-specific tokens that are shared across local windows.

Figures 2(b) and 2(c) specify the injection point. The encoder retains the standard W-MSA/SW-MSA sequence: W-MSA operates on non-overlapping windows, and SW-MSA applies a cyclic shift before partitioning and reverses it after attention. Prompt tokens are inserted after window partitioning. For one window and one attention head at stage i, let $\mathbf{F}_i \in \mathbb{R}^{N\times d_h}$ contain the image tokens and let $\mathbf{P}_i \in \mathbb{R}^{M_i\times d_h}$ contain the prompt tokens. Here, N is the number of image tokens and d_h is the head dimension. In our experiments, each 8×8 window contains $N = 64$ image tokens, and the projectors produce $M_i = 16$ prompt tokens (i.e., $N/4$). The joint token sequence is

$$\mathbf{X}_i = [\mathbf{P}_i; \mathbf{F}_i] \in \mathbb{R}^{(N+M_i)\times d_h}.$$

Queries are computed from image tokens, whereas keys and values are computed from the concatenated image and prompt tokens:

$$\mathbf{Q}_i = \mathbf{F}_i\mathbf{W}_Q, \quad \mathbf{K}_i = \mathbf{X}_i\mathbf{W}_K, \quad \mathbf{V}_i = \mathbf{X}_i\mathbf{W}_V,$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_h\times d_h}$. The resulting prompt-augmented attention is

$$\mathbf{O}_i = \mathrm{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}} + \widetilde{\mathbf{B}}_i\right)\mathbf{V}_i,$$

where $\widetilde{\mathbf{B}}_i \in \mathbb{R}^{N\times(N+M_i)}$ extends the Swin relative-position bias with zero-valued columns for the prompt tokens. The attention score matrix therefore has size $N\times(N+M_i)$. After multi-head aggregation, the prompt outputs are discarded and only the N updated image tokens are propagated.

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 25,2026 at 14:36:45 UTC from IEEE Xplore. Restrictions apply.

The four prompt scales coincide with the spatial hierarchy of the encoder and condition both high-resolution and semantic feature stages.

C. Swin-Based JSCC Encoder and Transmission Pipeline

The communication backbone is a hierarchical Swin-style JSCC encoder comprising patch embedding, transformer stages, and progressive spatial downsampling [8], [16], [17]. Prompt injection precedes the unchanged normalization and channel-transmission operations in all four encoder stages.

All prompts are injected exclusively into the transmitter-side encoder, while the receiver remains unchanged. A separate encoder-prompt adapter is trained for each task, while the communication backbone remains shared and frozen. Joint multitask adapter training is outside the scope of this study.

For a 256×256 input, four-stage prompting introduces 6.65M parameters and 2.12 G MACs per image. The additional computation corresponds to 12.13% of the 17.50 G-MAC encoder cost and 6.15% of the 34.52 G-MAC encoder-decoder cost. No computation is added at the receiver.

TABLE I. PROMPT OVERHEAD (ENCODER-SIDE ONLY, 256 X 256 INPUT).

Component	Extra Params	Extra MACs/image	Side
Task emb. +gates	196	< 0.001 G	Tx
Prompt generator	0.833 M	0.906 G	Tx
Stage projectors	5.812 M	0.006 G	Tx
Prompt attention	0	1.211 G	Tx
Total extra	6.646 M	2.123 G	Tx
Receiver extra	0	0	Rx

D. Training Objective

The active task determines the loss terms:

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{task}\mathcal{L}_{task}(\mathbf{z}, \hat{\mathbf{z}}),$$

where $(\lambda_{rec}, \lambda_{task})$ is (0.1,0.1) for classification, (0,0.2) for detection, and (0.1,0.2) for instance segmentation. Classification uses cross-entropy and L1 reconstruction without perceptual loss. The structured tasks use frozen-FPN feature matching, with instance segmentation additionally using L1 reconstruction. Their feature loss is

$$\mathcal{L}_{task}^{det/seg} = \frac{1}{|\mathcal{S}|}\sum_{\ell \in \mathcal{S}} \|\Phi_\ell(\hat{\mathbf{x}}) - \Phi_\ell(\mathbf{x})\|_2^2,$$

where $\Phi_\ell(\cdot)$ is the frozen R-50-FPN feature map at level ℓ and $\mathcal{S} = \{\text{p2}, \text{p3}, \text{p4}, \text{p5}, \text{p6}\}$. The aggregate feature loss is weighted by 0.2. Detection AP and mask AP follow the COCO definition averaged over IoU thresholds from 0.50 to 0.95 in increments of 0.05; neither metric is directly optimized.

Algorithm 1 HiPrompt-JSCC Prompt Adaptation

Require: Pretrained JSCC encoder/decoder (f_θ, g_ϕ), target task $\mathbf{t}$
1: Initialize the target-task embedding, prompt generator, and projectors
2: for each training iteration do
3: Sample mini-batch $(\mathbf{x}, \mathbf{z})$, CBR, and SNR γ for task $\mathbf{t}$
4: Compute task code $\mathbf{e} = \mathcal{F}(\mathbf{e}_t)$ and features $\{\mathbf{f}_v^{(i)}\}_{i=1}^4$
5: for i = 1, ..., 4 do
6: Generate prompt $\mathbf{p}_i = \mathcal{P}_i(\mathbf{f}_v^{(i)}, \mathbf{e})$
7: Inject $\mathbf{p}_i$ into W-MSA/SW-MSA as key/value context
8: end for
9: Encode and transmit $\mathbf{y} = f_\theta(\mathbf{x}, \text{CBR}; \{\mathbf{p}_i\})$, $\hat{\mathbf{y}} = \mathcal{W}_\gamma(\mathbf{y})$
10: Decode $\hat{\mathbf{x}} = g_\phi(\hat{\mathbf{y}}, \text{CBR})$, compute $\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{task}\mathcal{L}_{task}$
11: Update prompt-related parameters
12: end for

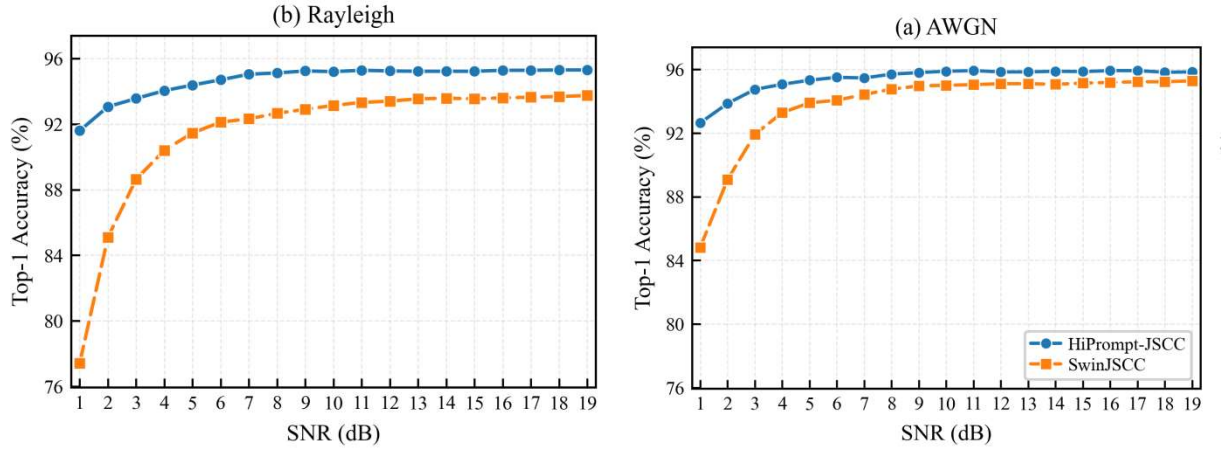

Fig. 3. ImageNet100 top-1 accuracy as a function of SNR under AWGN and Rayleigh fading.

IV. EXPERIMENTS

A. Experimental Setup

All experiments use 256×256 inputs, random seed 42, the Adam optimizer, a batch size of 8, gradient-norm clipping at 1.0, and one NVIDIA RTX 3090. Separate task-specific adapters are trained for ImageNet100 classification, COCO object detection, and COCO instance segmentation. The classification adapter is trained for 100 epochs, whereas the detection and instance-segmentation adapters are each trained for 50 epochs; SNR is sampled uniformly from 0–20 dB. The prompt configuration uses 64-dimensional task embeddings, hidden dimension 128, four encoder stages, and 16 prompt tokens per block.

For ImageNet100 classification [19], the learning rate is 3×10^{-4} and CBR is sampled from 0.1 to 0.9. The training set is partitioned once into disjoint subsets of at most 5000 images;

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 25,2026 at 14:36:45 UTC from IEEE Xplore. Restrictions apply.

each epoch cycles to one subset and is not a full-set pass. The validation subset is sampled once with seed 42 and contains at most 5000 images. Reported classification results use a single seed-42 run. Detection and segmentation adapters are trained on COCO train2017 and evaluated on val2017 [20]. The learning rate is 1×10^{-4}, and the R-50-FPN recognition networks remain frozen [18], [21]–[23].

For structured perception, CBR is sampled from $1/48$ to $1/8$ and evaluated at CBR $1/12$. The no-prompt baseline is the original reconstruction-oriented SwinJSCC model trained on ImageNet images using PSNR/MS-SSIM objectives. It is evaluated without task adaptation on the same data and with the same channel model, CBR, and downstream evaluators as HiPrompt-JSCC. The two-stage ablation activates prompts only in encoder stages 1–2. Reported metrics are RGB PSNR, top-1 accuracy, detection AP, and mask AP.

Classification is evaluated at SNR {1,4,7,10,13,16,19} dB under AWGN and Rayleigh fading. Structured perception uses the same SNR points under AWGN with CBR $1/12$. The BPG quality factors are 44, 41, 37, 35, 32, 31, and 29. BPG+LDPC uses 4QAM with code rates $1/3$, $1/2$, and $2/3$; 16QAM with rates $1/2$ and $2/3$; and 64QAM with rates $1/2$ and $2/3$. Its target BPG rates are $1/18$, $1/12$, $1/9$, $1/6$, $2/9$, $1/4$, and $1/3$ bpp (not JSCC CBR).

B. Comparison Methods

The comparison includes DeepJSCC [6], SwinJSCC [8], JPEG [24], fixed-quality BPG, and BPG+LDPC [25], [26].

C. Main Results

TABLE II. IMAGENET100 CLASSIFICATION AND PROMPT-STAGE ABLATION AT 1 DB AND CBR 1/2.

Method	Channel	Top-1 Acc. (%)
SwinJSCC baseline	AWGN	84.82
HiPrompt-JSCC (2-stage abl.)	AWGN	91.76
HiPrompt-JSCC (4-stage)	AWGN	92.64
SwinJSCC baseline	Rayleigh	77.42
HiPrompt-JSCC (4-stage)	Rayleigh	91.60

At 1 dB under AWGN, the four-stage model attains 92.64% top-1 accuracy, compared with 84.82% for the pretrained SwinJSCC baseline (Table II). Restricting prompts to stages 1–2 yields 91.76%, which is 0.88 percentage point below the four-stage result. Under Rayleigh fading, the corresponding values are 91.60% and 77.42%. At 10 and 19 dB, the two-stage model reaches 94.84% and 95.12%, whereas four-stage prompting reaches 95.88% and 95.84%, respectively. As shown in Fig. 3, the gain under AWGN is largest at low SNR.

Figure 4 and Table III summarize the structured-perception results. At 10 dB, HiPrompt-JSCC reaches 39.48 detection AP and 36.82 mask AP, exceeding the pretrained SwinJSCC baselines by 0.68 and 1.06 points, respectively. Over the evaluated SNR range, the detection difference is 0.48–0.75 AP and the mask difference is 0.92–1.38 AP.

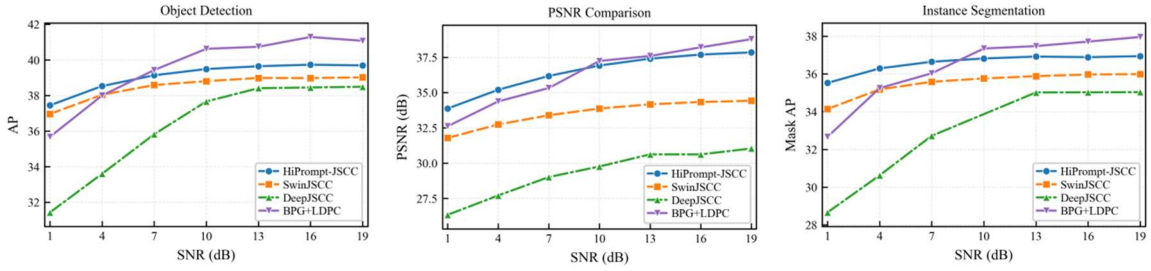

Fig. 4. Detection AP, reconstruction PSNR, and instance-segmentation mask AP under AWGN. Markers denote measured operating points. The DeepJSCC task-AP value at 16 dB is linearly interpolated from 13 and 19 dB and is excluded from the prompt-versus-SwinJSCC comparisons.

All points are measured except the DeepJSCC task-AP result at 16 dB, which is obtained by linear interpolation between 13 and 19 dB and is not included in the reported differences. DeepJSCC instance-segmentation results use COCO category-ID remapping.

HiPrompt-JSCC and the unadapted pretrained SwinJSCC baseline use the same downstream evaluators, channel types, SNR points, and CBR settings; their differences therefore isolate the effect of encoder-side prompt adaptation.

In Fig. 5, the classification response is concentrated on the foreground subject, whereas the detection response is concentrated around the rider and horse. These examples are illustrative and are not used to support the quantitative comparisons.

TABLE III. STRUCTURED-PERCEPTION AP UNDER AWGN

SNR	Detection AP		Mask AP	
(dB)	Swin	Ours	Swin	Ours
1	36.95	37.44	34.14	35.52
4	38.04	38.52	35.18	36.29
7	38.58	39.13	35.58	36.64
10	38.80	39.48	35.76	36.82
13	38.98	39.64	35.88	36.92
16	38.98	39.73	35.97	36.89
19	39.02	39.69	35.98	36.94

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 25,2026 at 14:36:45 UTC from IEEE Xplore. Restrictions apply.

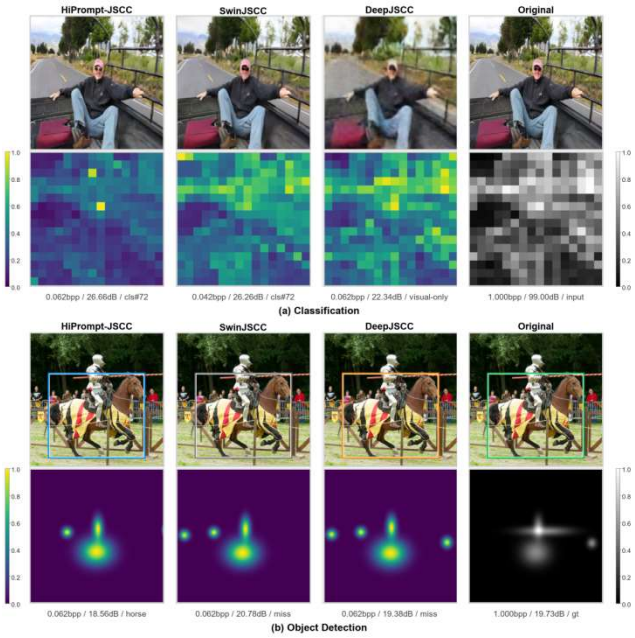

Fig. 5. Qualitative examples at 5 dB for classification and 10 dB for COCO detection.

V. Conclusion

HiPrompt-JSCC adds hierarchical task prompts to the transmitter of a frozen Swin-based JSCC system. Task-specific adapters improve ImageNet100 classification and COCO detection and instance segmentation over the corresponding pretrained SwinJSCC baselines while leaving the receiver unchanged. Evaluation uses one seed: classification covers AWGN and Rayleigh fading, whereas structured perception covers AWGN. Future work will study black-box adaptation without backbone gradients, over-the-air robustness to estimation errors and hardware impairments, multi-task joint prompting with shared and task-specific tokens, and multi-seed adapter-accuracy trade-offs.

Acknowledgment

This work was supported by the National Key Research and Development Program of China under Grant 2023YFB2504704-02, the Key Research and Development Program of Shandong Province under Grant 2023CXPT032, and the Beijing Science and Technology Planning Project under Grant Z241100003524011.

References

[1] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[2] W. Yang *et al.*, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, First Quarter 2023.

[3] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[4] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, Art. no. 107930, May 2021.

[5] H. Choi and I. V. Bajić, "Deep feature compression for collaborative object detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 3743–3747.

[6] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[7] K. Yang *et al.*, "WITT: A wireless image transmission transformer for semantic communications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

[8] K. Yang *et al.*, "SwinJSCC: Taming Swin Transformer for deep joint source-channel coding," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 1, pp. 90–104, Feb. 2025, doi: 10.1109/TCCN.2024.3424842.

[9] J. Ballé *et al.*, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[10] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 10771–10780.

[11] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7936–7945.

[12] M. Jia *et al.*, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 709–727.

[13] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[14] N. Houlsby *et al.*, "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2790–2799.

[15] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: University of Illinois Press, 1949.

[16] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[17] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[19] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[20] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

[22] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[24] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992.

[25] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[26] R. G. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, Jan. 1962.

Authorized licensed use limited to: BEIJING INSTITUTE OF TECHNOLOGY. Downloaded on November 25,2026 at 14:36:45 UTC from IEEE Xplore. Restrictions apply.