Restoration Flow Matching-Based Channel Refinement and Equalization Correction for MIMO Semantic Communications

Wenkai Liu, Nan Ma, *Member, IEEE,* Jianqiao Chen, Xiaodong Xu, *Senior Member, IEEE,* Meixia Tao, *Fellow, IEEE,* and Ping Zhang, *Fellow, IEEE*

***Abstract*—In multiple-input multiple-output (MIMO) semantic communication, imperfect channel state information (CSI) and equalization mismatch can seriously degrade semantic reconstruction quality. To address this issue, we propose a unified restoration flow matching (RFM)-based framework for channel refinement and equalization correction. Specifically, the channel RFM (CRFM) module is developed to refine the coarse channel, thereby improving channel estimation accuracy. Based on the refined channel, the developed semantic RFM (SRFM) module is employed to correct the residual distortions in the post-equalization latent space. The key idea is to formulate the two cascaded inverse problems of channel estimation and equalization as the unified conditional restoration task, in which the learned conditional velocity field guides the perturbed distribution towards the target distribution. To enhance the robustness of these two modules under various distortion conditions, we develop a dual-anchor perturbation training strategy that jointly learns near-manifold refinement and large-error correction, and implement inference through a few-step deterministic ordinary differential equation (ODE) solver. Extensive experiments on MIMO channels and visual semantic transmission tasks demonstrate that the proposed scheme improves key metrics for channel estimation and semantic reconstruction quality. Moreover, compared with representative diffusion-based generative baselines, the proposed method requires fewer sampling steps.**

***Index Terms*—Semantic communication, flow matching, channel estimation and equalization, MIMO communications.**

I. Introduction

Semantic communication has emerged as a highly promising paradigm for next-generation wireless networks by shifting the design objective from bit-level fidelity to the transmission of task-relevant information and semantic preservation [1], [2], [3]. This paradigm is particularly appealing for 6G visual transmission services, as the massive scale of source data and complex channel environments make traditional approaches that separate source coding from channel coding increasingly inefficient. In this context, deep joint source-channel coding (DeepJSCC) has become one of the most representative implementation frameworks for semantic communication, as it enables the end-to-end optimization of semantic representation, channel adaptation, and reconstruction quality within a neural network architecture [4], [5].

Wenkai Liu, Nan Ma, Xiaodong Xu and Ping Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: liuwenkai@bupt.edu.cn; manan@bupt.edu.cn; xuxiaodong@bupt.edu.cn; pzhang@bupt.edu.cn).(Corresponding author: Nan Ma.)

Jianqiao Chen is with the ZGC Institute of Ubiquitous-X Innovation and Applications, Beijing 100876, China (e-mail: jqchen1988@163.com).

Meixia Tao is with the Cooperative Medianet Innovation Center (CMIC) and the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (mxtao@sjtu.edu.cn).

Following this line of thought, a deep learning-based JSCC framework was proposed in [6], in which image compression and error correction were jointly realized through a convolutional neural network (CNN)-based autoencoder architecture. In addition, Transformer-based architectures have demonstrated superior capabilities in modeling high-dimensional visual semantics [7]. In particular, the hierarchical shifted-window design of the Swin Transformer achieves an ideal balance among multi-scale feature extraction, long-range dependency modeling, and computational efficiency, making it highly suitable for semantic image transmission [8]. The performance of these models degrades sharply when channel noise is severe or the testing statistics deviate from the training distribution. To address this issue, the authors in [9] proposed an adaptive JSCC with an attention module, exhibiting enhanced robustness under additive white Gaussian noise (AWGN) channel mismatch conditions. Subsequently, DeepJSCC explicitly established signal-to-noise ratio (SNR) adaptation as a design objective, emphasizing the robustness of a single model when operating under varying SNR conditions [10]. In [11], the authors used the Swin transformer as the backbone of the JSCC, and the SNR and rate adaptation techniques were able to adapt to different channel conditions and transmission rates. These end-to-end scheme studies enable semantic transceivers to adapt to channel variations, but they rely on direct decoding from degraded semantic features, which limits the high-fidelity recovery of source information in the presence of severe noise or channel distribution mismatches.

With the emergence of diffusion models, which excel at generating realistic textures in image synthesis [12], various diffusion models are now being applied to improve the quality of reconstructed images, even under high compression ratios [13], [14]. This trend has further propelled the development of generative denoising and recovery methods tailored for semantic communication systems [15], [16], [17]. In wireless semantic communication, the channel denoising diffusion models (CDDM) is deployed as a denoising module following channel equalization to achieve channel adaptivity in [15]. The authors in [16] proposed an unsupervised latent diffusion posterior sampling method that achieves joint semantic equalization and denoising. Further, the authors in [17] proposed an interference cancellation diffusion model (ICDM), which formulates the interference cancellation problem as a maximum a posteriori problem for the joint posterior probability of the signal and interference. Therefore, under the influence of multiplicative channel distortion, explicit equalization becomes a critical component of the semantic receiver. Its role is to suppress amplitude-phase distortion and feature coupling caused by the

channel, thereby providing a latent representation that is easier to recover for subsequent semantic reconstruction.

However, the majority of existing works primarily focus on single-input single-output (SISO) channels [18]. In view of the leading role that MIMO technology plays in enhancing channel capacity and transmission reliability, extending semantic communication from single-antenna links to MIMO systems is critically important [19]. In MIMO configurations, the received semantic information exhibits the nonlinear mixture and multi-stream coupling induced by the channel matrix, which poses significant challenges to robust semantic transmission. To address this issue, several representative works have explored robust transmission mechanisms from various perspectives in the context of MIMO semantic communications [20], [21], [22], [23]. In [20], orthogonal space-time block coding (OSTBC) was utilized to improve the robustness of image transmission against MIMO channel variations, and an equalizer was employed to estimate the codewords. In [21], the authors performed diffusion-based denoising on the equalized subchannels. Furthermore, the authors developed the DeepJSCC framework, which is specifically tailored for MIMO communications. This framework employed a vision transformer-based architecture that leverages both contextual semantic features and channel conditions through self-attention mechanisms [22]. Another study proposed a communication framework for MIMO-OFDM systems based on semantic importance, which utilizes semantic importance to jointly optimize quantization, subcarrier mapping, and power allocation [23].

Although the aforementioned studies have improved the recovery quality of corrupted semantic representations and advanced research on semantic communication, they largely rely on the assumption of perfect CSI at the receiver. This assumption is often difficult to satisfy in practical wireless environments, thereby limiting the real-world applicability of these methods. Therefore, it is crucial to design high-precision, low-complexity channel estimation methods under pilot-constrained conditions. In fact, channel estimation itself is a classic inverse problem with a wealth of existing literature [24], [25]. The least squares (LS) was widely used due to its simplicity, but its estimation accuracy degrades significantly under low SNR conditions [24]. Sparse recovery methods, such as orthogonal matching pursuit (OMP) [25], can leverage channel sparsity to achieve superior performance. However, their performance depends on the quality of the sparse representation dictionary and the validity of the sparsity assumption. In recent years, deep learning (DL) methods have provided new insights into channel estimation by exploiting statistical correlations in wireless channels [26], [27], [28], [29], [30]. The authors in [26] proposed a denoising CNN based on deep residual learning for channel denoising. In [27], the authors further proposed a score-based method with posterior sampling. In [28], a generative prior-based MIMO channel estimation method was proposed, which can handle quantized observations even when using low-resolution analog-to-digital converters (ADCs). In the field of semantic communications, a few studies have also begun to explicitly consider the channel estimation module. For example, [29] developed an end-to-end digital semantic communication system and introduced a ResNet-based channel estimator for OFDM transmission over frequency-selective fading channels. Additionally, [30] proposed a semantic-aware system tailored for speech-to-text transmission and extended it with a neural network channel estimation module to mitigate the system's reliance on precise CSI.

Nevertheless, existing research still lacks systematic studies on semantic recovery under imperfect-CSI conditions. In MIMO semantic communication, the receiver essentially needs to solve two cascaded and coupled inverse problems. First, the channel is estimated based on pilot observations, and then equalization is performed using imperfect-CSI to recover the semantic latent state. Since channel estimation errors directly lead to equalization mismatch [31] and propagate to subsequent semantic reconstruction modules, the overall semantic recovery performance is significantly degraded.

Therefore, to solve these cascaded inverse problems of channel estimation and equalization, we transform both into generative model-based restoration tasks. However, for latency-sensitive communication receivers, a key limitation of diffusion-based generative models is that the sampling process is typically iterative and relatively slow [32], [33], [34], [17], [35]. Compared with standard diffusion sampling, accelerated variants such as DDIM reduce the reverse sampling step [32], [33]. Some studies truncated the denoising chain according to channel conditions or receiver-side SNR information [17], [34]. In addition, the variational inference method based on score matching was introduced to directly approximate the posterior distribution of the MIMO channel, thereby achieving efficient channel recovery by reducing sampling steps [35]. Nevertheless, the inference process generally still requires multiple successive denoising evaluations. Recently, the flow matching (FM) framework [36], based on continuous normalizing flows, represents the latest advancement in generative modeling. Instead of explicitly simulating step-by-step reverse diffusion on stochastic differential equation (SDE) like traditional diffusion models, FM directly learns a continuous, deterministic transport process from a simple distribution to the target distribution by regressing a time-dependent vector field along a predefined probability path. More importantly, FM supports deterministic inference based on ordinary differential equation (ODE) and enables fast sampling utilizing standard numerical ODE solvers. Recent work [37], [38] has introduced FM to wireless communications, achieving faster inference while maintaining recovery accuracy by learning the vector field of the signal distribution.

Motivated by these observations, we develop a unified RFM-based framework for robust semantic recovery in practical MIMO receivers subject to cascaded distortions from pilot-based CSI acquisition and CSI-dependent equalization. The framework formulates channel refinement and equalization correction as two domain-specific yet structurally aligned conditional restoration tasks. The main contributions of this work are summarized as follows.

- We formulate the problem of practical MIMO semantic image receiver under imperfect CSI conditions as a cascaded inverse recovery problem. This modeling approach elucidates how channel estimation errors caused by pilot

signals propagate through the CSI-dependent equalization process and ultimately translate into structured distortion in the semantic latent domain. Furthermore, we reveal that channel refinement and equalization correction share a common restoration structure, both recovering the clean target from a disturbed state.

- To counteract these two distinct types of interference that affect semantic decoding, we unify channel refinement and equalization correction into a cascaded RFM task. We construct optimal transport-based conditional restoration paths in the channel and semantic-latent domains, where controlled corrupted source states are transported toward paired clean target states through velocity fields learned under constant-velocity supervision. Specifically, we develop a CRFM module to mitigate channel estimation errors, and a SRFM module to suppress residual latent distortion after equalization. We further show that the learned velocity field provides a local restoration direction of the Gaussian-smoothed clean-state distribution.
- We propose a dual-anchor perturbation training strategy for learning robust restoration flows under various receiver distortion conditions. By learning a shared velocity field through a mixture of low-perturbation and high-perturbation paths, this approach provides an implicit multiscale restoration mechanism that enables the restoration model to handle varying degrees of distortion.
- Based on the learned restoration velocity field, we propose an efficient cascaded ODE inference scheme. Compared with representative diffusion-based generative baselines, the proposed method requires fewer sampling steps. Experiments on MIMO channels and visual semantic transmission tasks demonstrate that the proposed scheme achieves superior performance in both channel estimation and semantic reconstruction quality.

A. Organization

The remainder of this paper is organized as follows. Section II presents the system model and problem formulation. Section III develops the cascaded restoration flow matching within the SwinT-JSCC framework. Section IV provides numerical results and ablation studies. Section V concludes the paper.

Notation: Scalars, vectors, and matrices are denoted by italic letters, boldface lower-case letters, and boldface upper-case letters, respectively. The superscripts $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$, $(\cdot)^{-1}$, and $(\cdot)^\dagger$ denote transpose, conjugate, Hermitian transpose, inverse, and pseudo-inverse, respectively. The operators $\otimes$, $\odot$, $\mathrm{vec}(\cdot)$, and $\mathrm{tr}(\cdot)$ denote Kronecker product, Hadamard product, vectorization, and trace. The Euclidean and Frobenius norms are denoted by $\|\cdot\|_2$ and $\|\cdot\|_F$, respectively. Moreover, $\lceil\cdot\rceil$ denotes the ceiling function, and $\mathbb{E}[\cdot]$ denotes expectation. Finally, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote real Gaussian and circularly symmetric complex Gaussian distributions, respectively.

II. System Model and Problem Formulation

A. Semantic Transmitter Design

As illustrated in Fig. 1, we consider a visual semantic communication system over a MIMO channel $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ with N_t transmit antennas and N_r receive antennas. The transmitter sends N_s spatial streams, where $N_s \le \min\{N_t, N_r\}$. The transmitted frame is composed of a pilot block followed by a semantic data block. The pilot block is represented by $\mathbf{P} \in \mathbb{C}^{N_s \times T_p}$, where T_p denotes the pilot length. Let $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$ denote the source image, where C, H, and W represent the number of channels, height, and width, respectively. At the transmitter, a pretrained Swin-Transformer-based JSCC (SwinT-JSCC) encoder is employed to extract compact semantic features from the source image. Specifically, the encoder maps $\mathbf{I}$ into a low-dimensional latent representation

$$\mathbf{Z} = f_{\theta_e^\star}(\mathbf{I}), \qquad \mathbf{Z} \in \mathbb{R}^{L \times D}, \tag{1}$$

where $f_{\theta_e^\star}(\cdot)$ denotes the pretrained semantic encoder with fixed parameters $\theta_e^\star$, L is the latent sequence length, and D is the feature dimension. The latent representation is then quantized by a scalar quantizer $\mathcal{Q}(\cdot)$, yielding $\mathbf{U} = \mathcal{Q}(\mathbf{Z}), \mathbf{U} \in \mathbb{R}^{L \times D}$. To facilitate physical-layer transmission, the quantized latent $\mathbf{U}$ is mapped into a complex baseband signal through a deterministic layer mapping operator

$$\mathbf{X} = \mathcal{T}(\mathbf{U}), \qquad \mathcal{T} : \mathbb{R}^{L \times D} \to \mathbb{C}^{N_s \times T_d}, \tag{2}$$

where $\mathbf{X}$ denotes the MIMO baseband signal and T_d is the number of channel uses. For clarity, the mapping $\mathcal{T}(\cdot)$ is detailed as follows.

First, $\mathbf{U}$ is vectorized into a one-dimensional sequence

$$\mathbf{u} = \mathrm{vec}(\mathbf{U}) = [u_1, u_2, \ldots, u_{LD}]^\mathsf{T} \in \mathbb{R}^{LD}. \tag{3}$$

To construct complex symbols, two adjacent real-valued entries are grouped into the in-phase and quadrature components. If LD is odd, one zero is appended to the tail of $\mathbf{u}$ for pairing. Let $\widetilde{\mathbf{u}} = [\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_{2N_c}]^\mathsf{T} \in \mathbb{R}^{2N_c}$ denote the zero-padded sequence, where $N_c = \lceil LD/2 \rceil$ is the total number of complex channel symbols. The resulting complex symbol sequence is given by

$$r_m = \widetilde{u}_{2m-1} + j\widetilde{u}_{2m}, \qquad m = 1, 2, \ldots, N_c. \tag{4}$$

To match the spatial degrees of freedom of the MIMO channel, we reconstruct the one-dimensional symbol stream $\mathbf{r} = [r_1, r_2, \ldots, r_{N_c}]^\mathsf{T}$ into N_s parallel streams. Let $T_d = \lceil N_c/N_s \rceil$ denote the required number of channel uses. After zero-padding, the symbol stream is rearranged into $\mathbf{R} = \mathrm{reshape}(\mathbf{r}, N_s, T_d) \in \mathbb{C}^{N_s \times T_d}$. To satisfy the average transmit power constraint, $\mathbf{R}$ is normalized as

$$\mathbf{X} = \frac{\mathbf{R}}{\sqrt{\|\mathbf{R}\|_\mathrm{F}^2/(N_s T_d)}}, \tag{5}$$

where the matrix $\mathbf{X} \in \mathbb{C}^{N_s \times T_d}$ is the complex baseband signal transmitted over the MIMO channel. Before MIMO transmission, both pilot and semantic data blocks are precoded by the same QR-orthonormalized random matrix $\mathbf{F} \in \mathbb{C}^{N_t \times N_s}$ satisfying $\mathbf{F}^H\mathbf{F} = \mathbf{I}_{N_s}$. Accordingly, the equivalent channel after precoding is defined as

$$\mathbf{H}_\mathrm{e} = \mathbf{H}\mathbf{F} \in \mathbb{C}^{N_r \times N_s}. \tag{6}$$

This equivalent channel is shared by the pilot and semantic data blocks and will be estimated at the receiver based on the pilot observation.

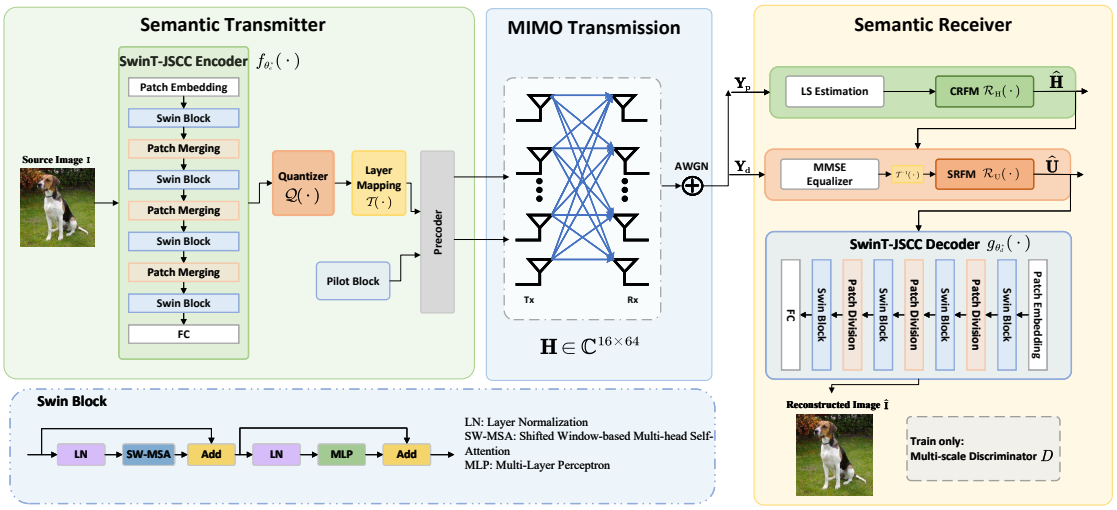

Fig. 1. Overall architecture of the cascaded RFM-based MIMO semantic communication system.

B. Semantic Receiver Design

At the receiver, the received frame is first divided into a pilot observation block and a semantic data observation block. The pilot block is used to acquire the MIMO channel, whereas the semantic data block is used to recover the transmitted semantic representation. The corresponding received signals are given by

$$\mathbf{Y}_{\mathrm{p}} = \mathbf{H}_{\mathrm{e}}\mathbf{P} + \mathbf{N}_{\mathrm{p}}, \tag{7}$$

$$\mathbf{Y}_{\mathrm{d}} = \mathbf{H}_{\mathrm{e}}\mathbf{X} + \mathbf{N}_{\mathrm{d}}, \tag{8}$$

where $\mathbf{N}_{\mathrm{p}}$ and $\mathbf{N}_{\mathrm{d}}$ denote AWGN matrices whose entries independently follow $\mathcal{CN}(0, \sigma_n^2)$. The receiver first obtains a coarse estimate of the channel from the pilot observation. The LS estimate is given by

$$\mathbf{H}_{\mathrm{LS}} = \mathbf{Y}_{\mathrm{p}}\mathbf{P}^H \left(\mathbf{P}\mathbf{P}^H\right)^{\dagger}. \tag{9}$$

This estimate provides an informative but noise-contaminated channel state. To mitigate the LS estimation error before semantic equalization, we introduce a channel restoration flow matching (CRFM) module, which refines the coarse estimate as

$$\hat{\mathbf{H}} = \mathcal{R}_{\mathrm{H}}\left(\mathbf{H}_{\mathrm{LS}}\right), \tag{10}$$

where $\mathcal{R}_{\mathrm{H}}(\cdot)$ denotes the CRFM operator[1]. Based on the refined channel, the receiver constructs the minimum mean square error (MMSE) equalizer as

$$\mathbf{W}\left(\hat{\mathbf{H}}\right) = \left(\hat{\mathbf{H}}^H\hat{\mathbf{H}} + \sigma_n^2\mathbf{I}_{N_s}\right)^{-1}\hat{\mathbf{H}}^H. \tag{11}$$

The equalized semantic signal is then obtained by

$$\hat{\mathbf{X}}_{\mathrm{MMSE}} = \mathbf{W}\left(\hat{\mathbf{H}}\right)\mathbf{Y}_{\mathrm{d}}. \tag{12}$$

Through the deterministic receive-side demapping operator $\mathcal{T}^{-1}(\cdot)$, the corresponding coarse semantic latent representation is obtained as $\hat{\mathbf{U}}_{\mathrm{MMSE}} = \mathcal{T}^{-1}\left(\hat{\mathbf{X}}_{\mathrm{MMSE}}\right)$. Although channel restoration and MMSE equalization suppress part of the physical-layer distortion, $\hat{\mathbf{U}}_{\mathrm{MMSE}}$ may still contain residual perturbations caused by imperfect-CSI, equalization mismatch, and data-domain noise. Therefore, a semantic restoration flow matching (SRFM) module is further employed to correct the coarse semantic latent representation

$$\hat{\mathbf{U}} = \mathcal{R}_{\mathrm{U}}\left(\hat{\mathbf{U}}_{\mathrm{MMSE}}\right), \tag{13}$$

where $\mathcal{R}_{\mathrm{U}}(\cdot)$ denotes the SRFM operator. Finally, the restored semantic latent representation is fed into the pretrained SwinT-JSCC decoder to reconstruct the source image

$$\hat{\mathbf{I}} = g_{\theta_{\mathrm{d}}^\star}\left(\hat{\mathbf{U}}\right), \tag{14}$$

where $g_{\theta_{\mathrm{d}}^\star}(\cdot)$ denotes the pretrained semantic decoder with fixed parameters $\theta_{\mathrm{d}}^\star$.

C. Problem Formulation

In imperfect-CSI-MIMO semantic communication, semantic recovery at the receiver is essentially performed in a cascaded manner, since the semantic equalizer is constructed based on channel estimates rather than the actual channel. Consequently, channel estimation errors are not confined to the channel domain but are converted by the equalizer into structured distortions in the transmitted semantic signal, thereby affecting the latent semantic representation.

Substituting (7) into (9), we obtain

$$\mathbf{H}_{\mathrm{LS}} = \mathbf{H}_{\mathrm{e}} + \underbrace{\mathbf{N}_{\mathrm{p}}\mathbf{P}^{\mathrm{H}}\left(\mathbf{P}\mathbf{P}^{\mathrm{H}}\right)^{\dagger}}_{\Delta\mathbf{H}}, \tag{15}$$

where $\Delta\mathbf{H}$ denotes the channel estimation error. Based on this imperfect-CSI, the MMSE equalizer given by (11) is

$$\mathbf{W}(\mathbf{H}_{\mathrm{LS}}) = \left[\left(\mathbf{H}_{\mathrm{LS}}\right)^H\mathbf{H}_{\mathrm{LS}} + \sigma_n^2\mathbf{I}_{N_s}\right]^{-1}\left(\mathbf{H}_{\mathrm{LS}}\right)^H. \tag{16}$$

[1]Here, all channel states fed into the CRFM module are first mapped into real-valued tensors by stacking their real and imaginary parts, i.e., $\mathbb{C}^{N_r\times N_s} \rightarrow \mathbb{R}^{2\times N_r\times N_s}$.

For the semantic data block, substituting (11) into the MMSE equalization gives

$$\begin{aligned}\mathbf{X}_{\mathrm{MMSE}} &= \mathbf{W}(\mathbf{H}_{\mathrm{LS}})\,\mathbf{Y}_{\mathrm{d}} \\ &= \mathbf{W}(\mathbf{H}_{\mathrm{e}}+\Delta\mathbf{H})\,(\mathbf{H}_{\mathrm{e}}\mathbf{X}+\mathbf{N}_{\mathrm{d}}) \\ &= \mathbf{W}(\mathbf{H}_{\mathrm{e}}+\Delta\mathbf{H})\,\mathbf{H}_{\mathrm{e}}\mathbf{X}+\mathbf{W}(\mathbf{H}_{\mathrm{e}}+\Delta\mathbf{H})\,\mathbf{N}_{\mathrm{d}}.\end{aligned} \tag{17}$$

Accordingly, the equalization error can be decomposed as

$$\begin{aligned}\mathbf{X}_{\mathrm{MMSE}}-\mathbf{X} = [\mathbf{W}(\mathbf{H}_{\mathrm{e}}+\Delta\mathbf{H})\,\mathbf{H}_{\mathrm{e}}-\mathbf{I}_{N_s}]\,\mathbf{X} \\ +\mathbf{W}(\mathbf{H}_{\mathrm{e}}+\Delta\mathbf{H})\,\mathbf{N}_{\mathrm{d}}.\end{aligned} \tag{18}$$

Therefore, the channel estimation error $|\Delta\mathbf{H}|$ typically amplifies the residual equalization mismatch in $\mathbf{X}_{\mathrm{MMSE}}$ and further propagates it into the semantic latent representations [31]. The above error propagation relationship indicates that it is necessary to improve the accuracy of channel estimation before performing semantic equalization. Therefore, we first introduce the CRFM operator $\mathcal{R}_{\mathrm{H}}(\cdot)$

$$\mathcal{R}_{\mathrm{H}}^{\star}=\arg\min_{\mathcal{R}_{\mathrm{H}}}\mathbb{E}\left[\left\|\mathcal{R}_{\mathrm{H}}\left(\mathbf{H}_{\mathrm{LS}}\right)-\mathbf{H}_{\mathrm{e}}\right\|_{\mathrm{F}}^{2}\right]. \tag{19}$$

Although CRFM improves the accuracy of channel estimation, as derived in (12), the equalized semantic latent remain a disturbed representation [15], [16], as follows

$$\widehat{\mathbf{U}}_{\mathrm{MMSE}}=\mathcal{T}^{-1}\left(\widehat{\mathbf{X}}_{\mathrm{MMSE}}\right)=\mathbf{U}+\Delta\mathbf{U}, \tag{20}$$

where $\Delta\mathbf{U}$ denotes the effective semantic latent perturbation. This motivates the introduction of the SRFM operator $\mathcal{R}_{\mathrm{U}}(\cdot)$

$$\mathcal{R}_{\mathrm{U}}^{\star}=\arg\min_{\mathcal{R}_{\mathrm{U}}}\mathbb{E}\left[\left\|\mathcal{R}_{\mathrm{U}}\left(\widehat{\mathbf{U}}_{\mathrm{MMSE}}\right)-\mathbf{U}\right\|_{\mathrm{F}}^{2}\right]. \tag{21}$$

In short, (15)–(18) show that channel estimation errors are converted by a CSI-based equalizer into structured equalization mismatches, which in turn induce the semantic latent disturbances described in (20). Although the two inverse problems of channel estimation and equalization differ in terms of signal-domain distortion and distortion characteristics, making it difficult to address them adequately with a single model, they share the same restoration structure. Whether for channel refinement or equalization correction, their objective is to recover a clean target from observation-dependent corrupted states. Section III addresses this issue using a cascaded RFM framework.

III. Cascaded Restoration Flow Matching within the SwinT-JSCC Framework

Tailored for the SwinT-JSCC-based MIMO semantic communication framework, this section develops a cascaded restoration flow matching receiver for the two successive inverse recovery tasks at the receiver, namely channel restoration flow matching (CRFM) and semantic restoration flow matching (SRFM).

A. Semantic Representation Learning Based on SwinT-JSCC

As shown in Fig. 1, before developing the restoration flow modules, we first pretrain the SwinT-JSCC encoder-decoder to establish a stable semantic latent representation space [8]. To improve perceptual realism, an adversarial loss is incorporated through a multi-scale discriminator D, $\widetilde{\mathbf{I}}$ denotes the interpolated sample between the real image and the reconstructed image. We design the loss function as follows

$$\mathcal{L}_{\mathrm{JSCC}}=\lambda_{\mathrm{mse}}\mathcal{L}_{\mathrm{mse}}+\lambda_{\mathrm{ssim}}\mathcal{L}_{\mathrm{ssim}}+\lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}}+\lambda_{\mathrm{gp}}\mathcal{L}_{\mathrm{gp}}, \tag{22}$$

where

$$\mathcal{L}_{\mathrm{mse}}=\mathbb{E}\left[\|\widehat{\mathbf{I}}-\mathbf{I}\|_2^2\right], \tag{23}$$

$$\mathcal{L}_{\mathrm{ssim}}=1-\mathrm{SSIM}(\widehat{\mathbf{I}},\mathbf{I}), \tag{24}$$

$$\mathcal{L}_{\mathrm{adv}}=-\mathbb{E}\left[D(\widehat{\mathbf{I}})\right], \tag{25}$$

$$\mathcal{L}_{\mathrm{gp}}=\mathbb{E}_{\widetilde{\mathbf{I}}}\left[\left(\|\nabla_{\widetilde{\mathbf{I}}}D(\widetilde{\mathbf{I}})\|_2-1\right)^2\right], \tag{26}$$

where λ_{mse}, λ_{ssim}, λ_{adv}, and λ_{gp} are nonnegative weighting coefficients. By minimizing $\mathcal{L}_{\mathrm{JSCC}}$, the pretrained semantic encoder and decoder are obtained as

$$(\theta_{\mathrm{e}}^{\star},\theta_{\mathrm{d}}^{\star})=\arg\min_{\theta_{\mathrm{e}},\theta_{\mathrm{d}}}\mathcal{L}_{\mathrm{JSCC}}, \tag{27}$$

which yield the semantic encoding and decoding mappings $f_{\theta_{\mathrm{e}}^{\star}}(\cdot)$ and $g_{\theta_{\mathrm{d}}^{\star}}(\cdot)$, respectively. After pretraining, these two mappings are fixed to provide a stable latent representation space.

B. Unified Flow Matching Framework

In this subsection, we first introduce a unified flow matching framework for the two recovery tasks described in (19) and (21). To simplify the derivation, we use $(\mathbf{S}^0,\mathbf{S}^1)$ to denote a generic source-target pair along a restoration path. Specifically,

$$(\mathbf{S}^0,\mathbf{S}^1)=\begin{cases}(\mathbf{H}^0,\mathbf{H}^1), & \text{for the CRFM operator,}\\ (\mathbf{U}^0,\mathbf{U}^1), & \text{for the SRFM operator.}\end{cases} \tag{28}$$

where $\mathbf{H}^0,\mathbf{H}^1\in\mathbb{R}^{2\times N_r\times N_s}$ denote the source and target endpoints in the real-valued channel tensor space, respectively, and $\mathbf{U}^0,\mathbf{U}^1\in\mathbb{R}^{L\times D}$ denote the source and target endpoints in the semantic latent space, respectively. FM learns a time-dependent vector field $\mathbf{v}_t^{\star}(\mathbf{S})$ that transports a source distribution p_0 to a target distribution p_1 [36]. Let $\{p_t\}_{t\in[0,1]}$ be a probability path connecting p_0 and p_1. The corresponding marginal velocity field $\mathbf{v}_t^{\star}(\mathbf{S})$ is defined through the continuity equation

$$\frac{\partial p_t(\mathbf{S})}{\partial t}+\nabla_{\mathbf{S}}\cdot\left[p_t(\mathbf{S})\,\mathbf{v}_t^{\star}(\mathbf{S})\right]=0. \tag{29}$$

Ideally, one would learn a neural field $\mathbf{v}_\theta(\mathbf{S}_t,t)$ by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta)=\int_0^1\mathbb{E}_{\mathbf{S}_t\sim p_t}\left[\|\mathbf{v}_\theta(\mathbf{S}_t,t)-\mathbf{v}_t^{\star}(\mathbf{S})\|_2^2\right]\mathrm{d}t. \tag{30}$$

However, directly minimizing (30) is generally intractable since the marginal velocity field is unavailable. According to conditional flow matching (CFM), this objective can be

replaced by a tractable conditional velocity regression with the same optimal marginal vector field [36], i.e.,

$$\mathcal{L}_{\mathrm{CFM}}(\theta)=\mathbb{E}_{t,\,p_t(\mathbf{S}_t|\mathbf{S}^1)}\left[\left\|\mathbf{v}_\theta(\mathbf{S}_t,t)-\mathbf{v}_t^\star(\mathbf{S}_t|\mathbf{S}^1)\right\|_2^2\right]. \quad (31)$$

For the optimal transport path [39]

$$\mathbf{S}_t=t\mathbf{S}^1+(1-t)\mathbf{S}^0, \quad (32)$$

In standard FM, the two endpoints of the probability path are given by $\mathbf{S}^0\sim p_0$ and $\mathbf{S}^1\sim p_1$, where $\mathbf{S}^0\sim p_0$ denotes the random source state, which is usually chosen as a standard Gaussian noise, i.e., $\mathbf{p}_0\sim\mathcal{N}(0,1)$. The resulting conditional velocity field is as follows

$$\mathbf{v}_t^\star(\mathbf{S}_t|\mathbf{S}^1)=\frac{\mathrm{d}\mathbf{S}_t}{\mathrm{d}t}=\mathbf{S}^1-\mathbf{S}^0. \quad (33)$$

Therefore, the CFM loss is rewritten as

$$\mathcal{L}_{\mathrm{CFM}}(\theta)=\mathbb{E}_{t,\,(\mathbf{S}^0,\mathbf{S}^1)}\left[\left\|\mathbf{v}_\theta(\mathbf{S}_t,t)-(\mathbf{S}^1-\mathbf{S}^0)\right\|_2^2\right]. \quad (34)$$

Based on the probabilistic path model described above, the generation process can be viewed as a transition from Gaussian noise to the target data distribution. However, for communication recovery tasks, this noise-to-data paradigm is not always the most natural choice. Unlike unconditional generation, the receiver's goal is to recover the specific channel or semantic latent state corresponding to the currently received observations. Therefore, the recovered results must not only be distribution-wise reasonable but also consistent with the received pilot signals or semantic observations. To address this issue, we propose training and inference schemes for the RFM model in Sections III-C and III-D, respectively.

C. Training of Restoration Flow Matching Models

In this subsection, we train the CRFM and SRFM modules as coarse-to-clean restoration flow models. For each clean target, we construct a controlled perturbed source state and learn a constant-velocity path toward the paired clean state. The training of the CRFM velocity field is parameterized using a UNet-based network [12], while the training of the SRFM velocity field is parameterized using a DiT-based network [13].

Specifically, we inject controlled perturbation into the clean channel state and the clean semantic latent variables as the source states, while using the corresponding clean states as the recovery targets, as follows

$$\mathbf{S}_\tau^0=\mathbf{S}^1+\tau\boldsymbol{\xi},\qquad \boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), \quad (35)$$

where $\tau>0$ controls the perturbation strength around the clean-state manifold. Here, $\mathbf{S}_\tau^0$ represents the perturbed channel state in CRFM, or the perturbed semantic latent state in SRFM. In this way, the source states are locally perturbed states around the clean state manifold. This provides a manageable training path from coarse to clean states for learning the restoration vector field used for channel refinement and equalization correction.

For $t\in[0,1]$, the associated interpolation path is defined as

$$\mathbf{S}_\tau^t=t\mathbf{S}^1+(1-t)\mathbf{S}_\tau^0=\mathbf{S}^1+(1-t)\tau\boldsymbol{\xi}. \quad (36)$$

Taking the derivative of (36) with respect to t gives the conditional velocity

$$\mathbf{u}_\tau^t\triangleq\frac{\mathrm{d}\mathbf{S}_\tau^t}{\mathrm{d}t}=-\tau\boldsymbol{\xi}. \quad (37)$$

This path starts from the perturbed source state $\mathbf{S}_\tau^0$ at $t=0$ and reaches the clean target state $\mathbf{S}^1$ at $t=1$. Therefore, the restoration flow field can be trained through the following single-anchor velocity regression objective

$$\mathcal{L}_{\mathrm{RFM}}(\theta;\tau)=\mathbb{E}_{\mathbf{S}^1,\boldsymbol{\xi},t}\left[\left\|\mathbf{v}_\theta(\mathbf{S}_\tau^t,t)-\mathbf{u}_\tau^t\right\|_2^2\right]. \quad (38)$$

Accordingly, the generic restoration path in (36) is instantiated for CRFM and SRFM as

$$\mathbf{H}_\tau^t=\mathbf{H}^1+(1-t)\tau\boldsymbol{\xi}, \quad (39)$$

$$\mathbf{U}_\tau^t=\mathbf{U}^1+(1-t)\tau\boldsymbol{\xi}, \quad (40)$$

where $\mathbf{H}_\tau^t$ and $\mathbf{U}_\tau^t$ denote the intermediate restoration states at time t in the channel and semantic latent domains, respectively. For notational simplicity, $\boldsymbol{\xi}$ denotes an independently sampled Gaussian perturbation in each domain, with dimensions conformable to $\mathbf{H}^1$ or $\mathbf{U}^1$.

Since the two restoration paths are linear in t, their target velocities are both given by $-\tau\boldsymbol{\xi}$. Accordingly, the CRFM and SRFM velocity fields are learned in their respective domains by minimizing the following task-specific velocity-regression objectives

$$\mathcal{L}_{\mathrm{H}}(\theta_{\mathrm{H}};\tau)=\mathbb{E}_{\mathbf{H}^1,\boldsymbol{\xi},t}\left[\left|\mathbf{v}_{\theta_{\mathrm{H}}}(\mathbf{H}_\tau^t,t)+\tau\boldsymbol{\xi}\right|_2^2\right], \quad (41)$$

$$\mathcal{L}_{\mathrm{U}}(\theta_{\mathrm{U}};\tau)=\mathbb{E}_{\mathbf{U}^1,\boldsymbol{\xi},t}\left[\left|\mathbf{v}_{\theta_{\mathrm{U}}}(\mathbf{U}_\tau^t,t)+\tau\boldsymbol{\xi}\right|_2^2\right], \quad (42)$$

where θ_{H} and θ_{U} are the trainable parameters of the CRFM and SRFM velocity networks, respectively. These objectives force the predicted velocity fields to match the constant restoration directions along the channel-domain and semantic-domain paths. Therefore, the learned networks are encouraged to progressively transport perturbed states toward their corresponding clean targets. The following proposition shows that the resulting optimal restoration velocity is proportional to the score of the Gaussian-smoothed clean-state distribution.

Proposition 1: Under the fixed-perturbation path in (36) with the standard Gaussian perturbation in (35), let $\mathbf{v}_\tau^\star(\mathbf{s},t)$ be the population-risk minimizer of (38) over square-integrable velocity fields, where $\tau>0$ and $t\in[0,1)$. Assume that the marginal density $p_\tau^t(\mathbf{s})$ of $\mathbf{S}_\tau^t$ is differentiable and positive. Then,

$$\mathbf{v}_\tau^\star(\mathbf{s},t)=(1-t)\tau^2\nabla_{\mathbf{s}}\log p_\tau^t(\mathbf{s}). \quad (43)$$

Proof: For a fixed $t\in[0,1)$, (38) reduces to a squared-error regression problem that maps the intermediate state $\mathbf{S}_\tau^t$ to the target velocity $\mathbf{u}_\tau^t$ [36]. Therefore, the population-risk minimizer is given by the conditional mean estimator

$$\mathbf{v}_\tau^\star(\mathbf{s},t)=\mathbb{E}\left[\mathbf{u}_\tau^t\,\middle|\,\mathbf{S}_\tau^t=\mathbf{s}\right]. \quad (44)$$

From the restoration path in (36) and the target velocity in (37), we have

$$\mathbf{u}_\tau^t=-\tau\boldsymbol{\xi}=\frac{\mathbf{S}^1-\mathbf{S}_\tau^t}{1-t}. \quad (45)$$

Substituting (45) into (44) gives

$$\mathbf{v}_\tau^\star(\mathbf{s},t)=\frac{\mathbb{E}\left[\mathbf{S}^1 \mid \mathbf{S}_\tau^t=\mathbf{s}\right]-\mathbf{s}}{1-t}. \tag{46}$$

Under the Gaussian perturbation model in (35), $\mathbf{S}_\tau^t \mid \mathbf{S}^1 \sim \mathcal{N}(\mathbf{S}^1,\sigma_{\tau,t}^2\mathbf{I})$, where $\sigma_{\tau,t}^2=(1-t)^2\tau^2$. Therefore, $p_\tau^t(\mathbf{s})$ is the Gaussian-smoothed density of the clean state. By Tweedie's identity [40], we have

$$\mathbb{E}\left[\mathbf{S}^1 \mid \mathbf{S}_\tau^t=\mathbf{s}\right]-\mathbf{s}=\sigma_{\tau,t}^2\nabla_{\mathbf{s}}\log p_\tau^t(\mathbf{s}). \tag{47}$$

Substituting (47) into (46) yields

$$\mathbf{v}_\tau^\star(\mathbf{s},t)=\frac{\sigma_{\tau,t}^2}{1-t}\nabla_{\mathbf{s}}\log p_\tau^t(\mathbf{s})=(1-t)\tau^2\nabla_{\mathbf{s}}\log p_\tau^t(\mathbf{s}). \tag{48}$$

This completes the proof. ■

Remark 1: Proposition 1 provides a score-based interpretation of the fixed-perturbation RFM objective. Specifically, the population-optimal velocity field does not merely fit the artificially injected Gaussian perturbation of each training sample. Rather, it captures a local restoration direction characterized by the score of the Gaussian-smoothed clean-state density, which guides perturbed states toward the clean-state manifold. These directions point toward regions of high probability on the "clean channel" or the "clean semantic latent". Hence, the online initial state need not exactly follow the artificial Gaussian perturbation model used in training, as long as the receiver-side coarse estimate lies within the clean-state neighborhood.

However, a single perturbation anchor is insufficient to characterize the heterogeneous distortion conditions encountered by practical receivers [41], [42]. Specifically, a small perturbation anchor generates training states close to the clean-state manifold and is suitable for fine-grained refinement, but it provides limited coverage for severely corrupted coarse receiver states. In contrast, a large perturbation anchor enlarges the coverage of heavily distorted states, but may weaken the local restoration accuracy around the clean-state manifold.

To accommodate different distortion levels, we introduce a dual-anchor perturbation training strategy with two perturbation anchors $0<\tau_\ell<\tau_h$. The low perturbation anchor τ_ℓ focuses on local refinement around the clean-state manifold, whereas the high perturbation anchor τ_h improves the coverage of strongly corrupted receiver states. The sensitivity to the perturbation-anchor selection is empirically evaluated in Section IV-B.

Specifically, the low- and high-anchor restoration paths are defined as

$$\mathbf{S}_{\tau_\ell}^t=\mathbf{S}^1+(1-t)\tau_\ell\boldsymbol{\xi}, \quad \mathbf{S}_{\tau_h}^t=\mathbf{S}^1+(1-t)\tau_h\boldsymbol{\xi}, \tag{49}$$

The corresponding conditional restoration velocities are defined as

$$\mathbf{u}_{\tau_\ell}^t\triangleq\frac{\mathrm{d}\mathbf{S}_{\tau_\ell}^t}{\mathrm{d}t}=-\tau_\ell\boldsymbol{\xi}, \quad \mathbf{u}_{\tau_h}^t\triangleq\frac{\mathrm{d}\mathbf{S}_{\tau_h}^t}{\mathrm{d}t}=-\tau_h\boldsymbol{\xi}. \tag{50}$$

Algorithm 1 Training of CRFM

Require: Clean channel tensor dataset $\mathcal{D}_{\mathrm{H}}$, perturbation anchors $\{\tau_\ell,\tau_h\}$, batch size B, iterations N_{it}.
Ensure: Trained channel velocity field $\mathbf{v}_{\theta_{\mathrm{H}}^\star}(\cdot,t)$.
1: **for** $n=1,\ldots,N_{\mathrm{it}}$ **do**
2: Sample clean channel targets $\{\mathbf{H}^{1,(b)}\}_{b=1}^B$ from $\mathcal{D}_{\mathrm{H}}$.
3: Sample $t^{(b)}\sim\mathcal{U}(0,1)$ and $\tau^{(b)}\in\{\tau_\ell,\tau_h\}$.
4: Construct the channel restoration state and target velocity according to (49) and (50).
5: Update θ_{H} by gradient step on $\mathcal{L}_{\mathrm{H}}(\theta_{\mathrm{H}};\tau_\ell,\tau_h)$ in (52).
6: **end for**

Algorithm 2 Training of SRFM

Require: Image dataset $\mathcal{D}_{\mathrm{I}}$, pretrained semantic encoder $f_{\theta_{\mathrm{e}}}(\cdot)$, quantizer $\mathcal{Q}(\cdot)$, perturbation anchors $\{\tau_\ell,\tau_h\}$, batch size B, iterations N_{it}.
Ensure: Trained semantic velocity field $\mathbf{v}_{\theta_{\mathrm{U}}^\star}(\cdot,t)$.
1: **for** $n=1,\ldots,N_{\mathrm{it}}$ **do**
2: Sample images $\{\mathbf{I}^{(b)}\}_{b=1}^B$ from $\mathcal{D}_{\mathrm{I}}$ and obtain clean semantic targets $\mathbf{U}^{1,(b)}=\mathcal{Q}(f_{\theta_{\mathrm{e}}}(\mathbf{I}^{(b)}))$.
3: Sample $t^{(b)}\sim\mathcal{U}(0,1)$ and $\tau^{(b)}\in\{\tau_\ell,\tau_h\}$.
4: Construct the semantic restoration state and target velocity according to (49) and (50).
5: Update θ_{U} by gradient step on $\mathcal{L}_{\mathrm{U}}(\theta_{\mathrm{U}};\tau_\ell,\tau_h)$ in (53).
6: **end for**

Dual-anchor Perturbation Training Objective: Based on (49) and (50), the dual-perturbation restoration flow matching objective is formulated as

$$\begin{aligned}\mathcal{L}_{\mathrm{RFM}}(\theta;\tau_\ell,\tau_h)=&\,\omega_\ell\mathbb{E}_{\mathbf{S}^1,\boldsymbol{\xi},t}\left[\left\|\mathbf{v}_\theta\left(\mathbf{S}_{\tau_\ell}^t,t\right)-\mathbf{u}_{\tau_\ell}^t\right\|_2^2\right]\\&+\omega_h\mathbb{E}_{\mathbf{S}^1,\boldsymbol{\xi},t}\left[\left\|\mathbf{v}_\theta\left(\mathbf{S}_{\tau_h}^t,t\right)-\mathbf{u}_{\tau_h}^t\right\|_2^2\right],\end{aligned} \tag{51}$$

In this work, we set $\omega_\ell=\omega_h=1/2$ to balance near-manifold refinement and large-error correction. Based on the single-anchor restoration objectives in (41) and (42), the proposed dual-anchor perturbation training objectives for CRFM and SRFM are obtained by weighted aggregation over the low- and high-perturbation anchors

$$\theta_{\mathrm{H}}^\star=\arg\min_{\theta_{\mathrm{H}}}\underbrace{\{\omega_\ell\mathcal{L}_{\mathrm{H}}(\theta_{\mathrm{H}};\tau_\ell)+\omega_h\mathcal{L}_{\mathrm{H}}(\theta_{\mathrm{H}};\tau_h)\}}_{\triangleq\,\mathcal{L}_{\mathrm{H}}(\theta_{\mathrm{H}};\tau_\ell,\tau_h)}, \tag{52}$$

$$\theta_{\mathrm{U}}^\star=\arg\min_{\theta_{\mathrm{U}}}\underbrace{\{\omega_\ell\mathcal{L}_{\mathrm{U}}(\theta_{\mathrm{U}};\tau_\ell)+\omega_h\mathcal{L}_{\mathrm{U}}(\theta_{\mathrm{U}};\tau_h)\}}_{\triangleq\,\mathcal{L}_{\mathrm{U}}(\theta_{\mathrm{U}};\tau_\ell,\tau_h)}, \tag{53}$$

where $\theta_{\mathrm{H}}^\star$ and $\theta_{\mathrm{U}}^\star$ are the optimized parameters for CRFM and SRFM, respectively. It is worth noting that each of these two modules learns a mixed restoration vector field on the composite trajectory distribution induced by these two perturbation anchors.

The following analysis shows that the dual-perturbation objectives in (52) and (53) can be interpreted as a velocity regression problem over a mixture of restoration-path distributions. This interpretation explains how a single shared velocity field can adapt to different receiver-side corruption levels without requiring a densely sampled perturbation schedule.

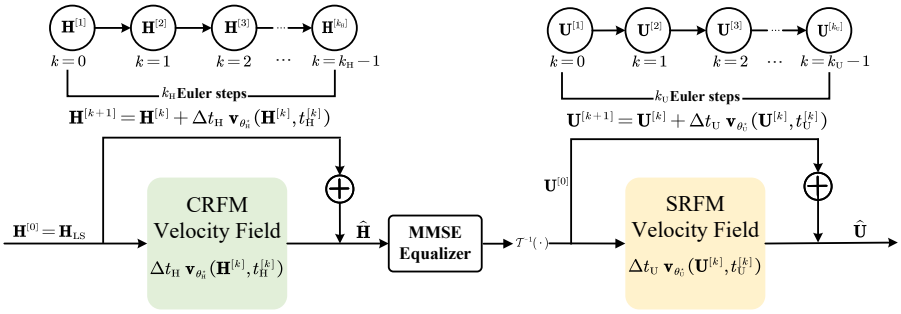

Fig. 2. Overall workflow of the proposed cascaded RFM inference.

Let $\mu^t_{\tau_\ell}$ and $\mu^t_{\tau_h}$ denote the joint distributions of $(\mathbf{S}^t_{\tau_\ell}, \mathbf{u}^t_{\tau_\ell})$ and $(\mathbf{S}^t_{\tau_h}, \mathbf{u}^t_{\tau_h})$, respectively. The mixture training law is given by

$$\mu^t_{\tau_\ell,\tau_h} = \omega_\ell \mu^t_{\tau_\ell} + \omega_h \mu^t_{\tau_h}. \tag{54}$$

Let $p^t_\tau(\mathbf{s})$ denote the marginal density of $\mathbf{S}^t_\tau$ for $\tau \in \{\tau_\ell, \tau_h\}$. Since $\mathbf{S}^t_\tau = \mathbf{S}^1 + (1-t)\tau\boldsymbol{\xi}$, we have

$$p^t_\tau(\mathbf{s}) = \mathbb{E}_{\mathbf{S}^1\sim p_1}\left[\mathcal{N}\left(\mathbf{s}; \mathbf{S}^1, (1-t)^2\tau^2\mathbf{I}\right)\right], \quad \tau \in \{\tau_\ell, \tau_h\}. \tag{55}$$

Accordingly,

$$p^t_{\tau_\ell,\tau_h}(\mathbf{s}) = \omega_\ell p^t_{\tau_\ell}(\mathbf{s}) + \omega_h p^t_{\tau_h}(\mathbf{s}). \tag{56}$$

For each anchor, define the corresponding population optimal velocity field as

$$\mathbf{v}^\star_\tau(\mathbf{s}, t) = \mathbb{E}_{\mu^t_\tau}\left[\mathbf{u} \,|\, \mathbf{S} = \mathbf{s}\right], \quad \tau \in \{\tau_\ell, \tau_h\}. \tag{57}$$

Proposition 2: For a fixed $t \in [0, 1)$, consider the population squared-risk minimizer under the mixture law $\mu^t_{\tau_\ell,\tau_h}$,

$$\mathbf{v}^\star_{\tau_\ell,\tau_h}(\cdot, t) = \arg\min_{\mathbf{v}} \mathbb{E}_{\mu^t_{\tau_\ell,\tau_h}}\left[\|\mathbf{v}(\mathbf{S}, t) - \mathbf{u}\|_2^2\right]. \tag{58}$$

For any $\mathbf{s}$ with $p^t_{\tau_\ell,\tau_h}(\mathbf{s}) > 0$, the optimal velocity field satisfies

$$\mathbf{v}^\star_{\tau_\ell,\tau_h}(\mathbf{s}, t) = \gamma_{\tau_\ell}(\mathbf{s}, t)\mathbf{v}^\star_{\tau_\ell}(\mathbf{s}, t) + \gamma_{\tau_h}(\mathbf{s}, t)\mathbf{v}^\star_{\tau_h}(\mathbf{s}, t), \tag{59}$$

where

$$\gamma_{\tau_\ell}(\mathbf{s}, t) = \frac{\omega_\ell p^t_{\tau_\ell}(\mathbf{s})}{\omega_\ell p^t_{\tau_\ell}(\mathbf{s}) + \omega_h p^t_{\tau_h}(\mathbf{s})}, \tag{60}$$

$$\gamma_{\tau_h}(\mathbf{s}, t) = \frac{\omega_h p^t_{\tau_h}(\mathbf{s})}{\omega_\ell p^t_{\tau_\ell}(\mathbf{s}) + \omega_h p^t_{\tau_h}(\mathbf{s})}. \tag{61}$$

Proof. See Appendix A. ■

Remark 2: Proposition 2 shows that dual-anchor perturbation training yields a shared population-optimal velocity field that adaptively combines the anchor-specific velocities through posterior responsibilities. These responsibilities act as implicit soft gates over different corruption scales: the low-perturbation anchor dominates near the clean-state manifold, whereas the high-perturbation anchor becomes more influential for severely distorted states. Therefore, the proposed strategy provides a compact multi-scale restoration mechanism for imperfect-CSI MIMO semantic reception, without requiring a dense perturbation schedule.

D. Cascaded ODE-Based Inference for Restoration Flow Matching

Based on the learned restoration vector fields, we develop a cascaded RFM inference scheme by solving two successive recovery ODEs for channel refinement and equalization correction, as illustrated in Fig. 2. An important feature of the proposed inference is that the restoration trajectories are coupled with the physical measurements through observation-dependent initialization. In the channel domain, the ODE starts from the LS estimate, which is directly obtained from the pilot observation and therefore already encodes the pilot measurement constraint. Hence, CRFM does not generate a channel realization from an uninformative prior, but performs prior-guided residual correction around a pilot-consistent coarse CSI. Similarly, in the semantic domain, the initial state is the MMSE-equalized latent representation obtained from the actual data observation using the restored channel, followed by the deterministic receive-side demapping. This state inherits the regularized data-fitting property of the MIMO receiver chain and serves as a coarse semantic representation. This design reduces the risk of the generated distribution being reasonable but inconsistent with measurement results, while retaining the ability to perform rapid ODE inference within a few steps. Accordingly, fast online inference is achieved with only $k_H = 5$ CRFM Euler steps and $k_U = 2$ SRFM Euler steps, as determined by the sensitivity analysis in Section IV-B.

Specifically, the CRFM and SRFM operators are implemented as terminal-state mappings of two deterministic restoration ODEs. For the channel-domain trajectory, the initial and terminal states are given by $\mathbf{H}^{[0]} = \mathbf{H}_{LS}$ and $\widehat{\mathbf{H}} = \mathbf{H}^{[k_H]}$, respectively. Using a uniform time grid $t_H^{[k]} = k/k_H$, where k_H denotes the number of CRFM sampling steps, and the time step is $\Delta t_H = 1/k_H$, the explicit Euler discretization can be expressed as

$$\mathbf{H}^{[k+1]} = \mathbf{H}^{[k]} + \Delta t_H\, \mathbf{v}_{\theta_H^\star}\left(\mathbf{H}^{[k]}, t_H^{[k]}\right), \quad k = 0, \ldots, k_H - 1. \tag{62}$$

After channel restoration, the refined channel $\widehat{\mathbf{H}}$ is used for MMSE equalization according to (12), yielding the coarse semantic latent representation $\widehat{\mathbf{U}}_{MMSE}$.

For the semantic-domain trajectory, the initial and terminal states are $\mathbf{U}^{[0]} = \widehat{\mathbf{U}}_{MMSE}$ and $\widehat{\mathbf{U}} = \mathbf{U}^{[k_U]}$, respectively. With

$t_{\mathrm{U}}^{[k]} = k/k_{\mathrm{U}}$ and $\Delta t_{\mathrm{U}} = 1/k_{\mathrm{U}}$, the explicit Euler discretization of the SRFM ODE is given by

$$\mathbf{U}^{[k+1]} = \mathbf{U}^{[k]} + \Delta t_{\mathrm{U}}\, \mathbf{v}_{\theta_{\mathrm{U}}^{\star}}\left(\mathbf{U}^{[k]}, t_{\mathrm{U}}^{[k]}\right), \quad k = 0, \ldots, k_{\mathrm{U}} - 1. \tag{63}$$

As a result, CRFM first calibrates the pilot-derived LS channel estimate, thereby reducing the CSI mismatch before equalization. SRFM then compensates for the residual latent distortion after MMSE equalization. The restored semantic latent representation $\widehat{\mathbf{U}}$ is finally fed into the semantic decoder according to (14).

IV. Numerical Results

In this section, we conduct a series of experiments to evaluate the performance of the proposed scheme, providing a comprehensive demonstration of its effectiveness across various scenarios.

A. Experimental Setup

1) Dataset and System Configurations: The wireless propagation channels are generated according to the CDL-B and CDL-C models specified in 3GPP TR 38.901 [43]. Unless otherwise stated, 10,000 independent channel realizations are used for training and 1,000 independent channel realizations are used for testing. The carrier frequency is set to 40 GHz, and the normalized antenna spacing is fixed at 0.5. Throughout the experiments, we consider a MIMO configuration with $N_t = 64$ transmit antennas and $N_r = 16$ receive antennas. The number of spatial streams is set to $N_s = 4$ by default. For visual semantic transmission, we use two image datasets constructed from ImageNet-mini [44] and CelebA [45]. All images are center-cropped and resized to 512×512 before semantic encoding. The ImageNet-mini split contains 34,745 training images and 3,923 testing images, while the CelebA split contains 2,693 training images and 300 testing images.

2) Training and Inference Configurations: All our experiments are performed on a Linux server with RTX 4090 GPU. The overall framework consists of three learnable components, namely the SwinT-JSCC semantic transceiver, the CRFM module and the SRFM module. The SwinT-JSCC encoder-decoder is first trained over an AWGN channel at $\mathrm{SNR} = 10\,\mathrm{dB}$ with a learning rate of 1×10^{-4}. After pretraining, the semantic transceiver is frozen, providing fixed latent representations for the SRFM training. The CRFM and SRFM modules are then trained separately according to the restoration flow matching objectives in (52) and (53), respectively. For both modules, the learning rate is set to 1×10^{-5}. The dual-anchor perturbation strategy is adopted with $(\tau_\ell, \tau_h) = (0.1, 1.0)$. During inference, $k_{\mathrm{H}} = 5, k_{\mathrm{U}} = 2$ Euler sampling steps are used for CRFM and SRFM, respectively.

3) Evaluation Metrics: For channel estimation, we use the normalized mean square error (NMSE) [27], defined as

$$\mathrm{NMSE}\,[\mathrm{dB}] = 10 \log_{10}\left(\frac{\|\widehat{\mathbf{H}} - \mathbf{H}_{\mathrm{e}}\|_F^2}{\|\mathbf{H}_{\mathrm{e}}\|_F^2}\right). \tag{64}$$

For semantic reconstruction, the channel bandwidth ratio (CBR) is defined as

$$\rho = \frac{N_c}{C \times H \times W}. \tag{65}$$

The CBR was set to $\rho = 1/96$ throughout the experiment. The reconstruction quality is evaluated using two widely adopted image-quality metrics, namely PSNR and MS-SSIM [46].

4) Baselines: We consider two benchmark groups for channel estimation and end-to-end semantic reconstruction.

For channel estimation, the proposed CRFM is benchmarked against the representative estimators covering classical model-based estimation, sparse recovery, supervised denoising, and generative recovery. Specifically, we consider **LS** estimation [24] and **Approximate MMSE** [27], **orthogonal matching pursuit (OMP)** [25] as a sparsity-driven recovery baseline, and **Score-based method** [27] and **DDIM-based method** [32] as diffusion based generative baselines.

For semantic reconstruction, we evaluate several ablation configurations of the receiver to quantify the individual and joint effects of CRFM and SRFM. The AWGN schemes, including **SwinT-JSCC (train** $10\,\mathrm{dB}$**)** and **SwinT-JSCC with SRFM**, are used as non-MIMO reference cases that only involve AWGN. For practical imperfect-CSI MIMO transmission, we consider **LS-MMSE**, **LS-MMSE with SRFM**, **LS-MMSE with CRFM**, and **LS-MMSE with CRFM and SRFM**. **LS-MMSE** denotes the baseline receiver that uses LS-estimated CSI to construct the MMSE equalizer. In addition, **Perfect-CSI MMSE** and **Perfect-CSI MMSE with SRFM** are included as upper-bound references to quantify the performance loss caused by pilot-based CSI acquisition. A conventional **JPEG-1/2LDPC-4QAM** system is also evaluated over the same MIMO channel as a separated source-channel coding baseline [6].

B. Sensitivity Analysis of RFM Design Parameters

This subsection examines two key factors of the proposed cascaded RFM receiver: the number of ODE sampling steps and the perturbation anchors used for RFM training. Fig. 3(a) shows the effect of the number of Euler steps on CRFM-based channel restoration. A single step is insufficient, especially in the low- and medium-SNR regions, since the learned restoration field cannot be fully integrated. As the number of steps increases from 1 to 2, the NMSE decreases rapidly, while the gain becomes marginal beyond 10 steps. This indicates that CRFM can reach a near-converged channel estimate with only a few deterministic ODE evaluations. Accordingly, we use $k_{\mathrm{H}} = 5$ in the following experiments. Fig. 3(b) presents the corresponding result for SRFM. The MS-SSIM improves noticeably with only a few semantic restoration steps, confirming that SRFM effectively compensates for residual latent distortion after equalization. Although more steps can further improve the reconstruction quality, the incremental gain gradually decreases. Considering the latency requirement of semantic receivers, we adopt $k_{\mathrm{U}} = 2$ as a practical operating point. Figs. 3(c) and 3(d) compare different perturbation anchors. The low-perturbation anchor $\tau_\ell = 0.1$ is more effective in the high-SNR region, where the coarse estimate is already

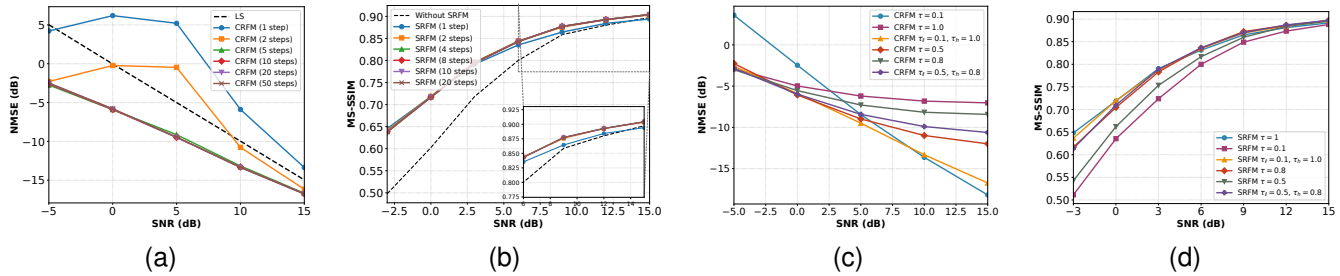

Fig. 3. Ablation study of the proposed cascaded RFM receiver under different ODE sampling steps and perturbation anchors. (a) Channel-estimation NMSE versus SNR over the CDL-C channel for different CRFM Euler steps, with $T_p = 4$. (b) Semantic reconstruction MS-SSIM versus SNR for different SRFM Euler steps under perfect CSI. (c) Channel-estimation NMSE versus SNR over the CDL-C channel for different CRFM perturbation anchors, with $T_p = 4$. (d) Semantic reconstruction MS-SSIM versus SNR for different SRFM perturbation anchors under perfect CSI.

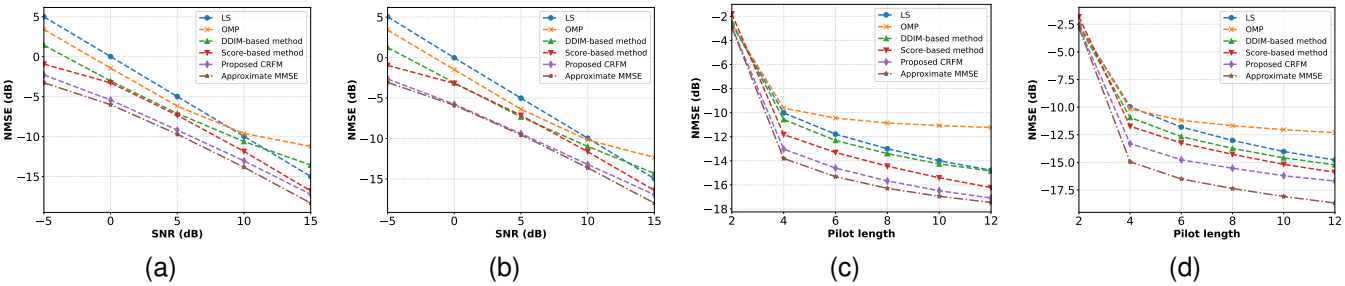

Fig. 4. Channel-estimation NMSE comparison over 3GPP CDL-B and CDL-C MIMO channels. (a) NMSE versus SNR over CDL-B with $T_p = 4$. (b) NMSE versus SNR over CDL-C with $T_p = 4$. (c) NMSE versus pilot length over CDL-B at $\mathrm{SNR} = 10\,\mathrm{dB}$. (d) NMSE versus pilot length over CDL-C at $\mathrm{SNR} = 10\,\mathrm{dB}$.

close to the clean manifold. By contrast, the high-perturbation anchor $\tau_h = 1.0$ provides stronger robustness in the low-SNR region with severe corruption. The dual-anchor model with $(\tau_\ell, \tau_h) = (0.1, 1.0)$ achieves the most balanced performance over the whole SNR range. These results verify that the proposed receiver is both sample-efficient and robust: few-step ODE inference is sufficient for online restoration, while dual-anchor training enables the learned flow to handle both coarse recovery and near-manifold refinement.

C. Evaluation of Channel Estimation Performance

We next evaluate the channel-estimation performance of the proposed CRFM under different SNRs and pilot lengths. The purpose is to verify whether CRFM can improve the coarse LS estimate under both noise-limited and pilot-limited conditions. Figs. 4(a) and 4(b) show the NMSE versus SNR over the CDL-B and CDL-C channels, respectively, with the pilot length fixed at $T_p = 4$. The proposed CRFM consistently achieves lower NMSE than LS, OMP, DDIM-based estimation, and score-based estimation over both channel models. The gain is especially evident in the low-SNR region, where the LS observation is severely corrupted and conventional sparse or generative baselines suffer from residual estimation errors. This confirms that CRFM can effectively learn a coarse-to-clean restoration field from noisy pilot-based channel estimates. Compared with the approximate MMSE estimator, CRFM also shows competitive performance without requiring explicit online covariance modeling. Figs. 4(c) and 4(d) further compare the NMSE performance versus pilot length at $\mathrm{SNR} = 10\,\mathrm{dB}$. As expected, increasing T_p improves the estimation accuracy of all methods, since more pilot observations provide a more reliable coarse channel estimate. However, CRFM maintains a clear advantage in the short-pilot regime, where channel recovery is more challenging and pilot overhead is more critical. The performance gap gradually decreases as the pilot length increases, but CRFM remains among the best practical schemes over the whole pilot range.

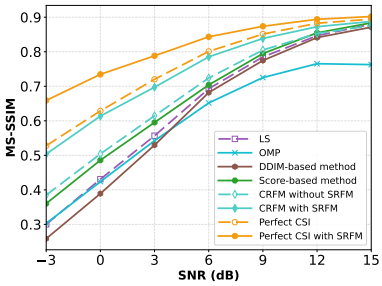

Fig. 5. Semantic reconstruction MS-SSIM versus SNR under different channel-estimation algorithms, with $T_p = 8$ and $\rho = 1/96$.

D. The Impact of Different Channel Estimation Algorithms on Semantic Reconstruction

Figs. 5 and 6 show the MS-SSIM and PSNR performance under different channel-estimation algorithms. The results reveal a clear correlation between CSI accuracy and semantic reconstruction quality. In this experiment, different channel estimators are used to construct the MMSE equalizer, while the

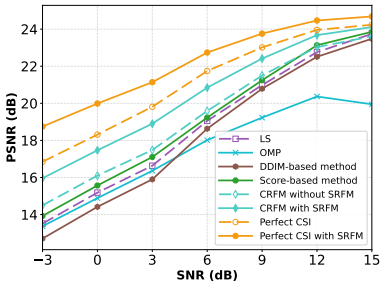

Fig. 6. Semantic reconstruction PSNR versus SNR under different channel-estimation algorithms, with $T_p = 8$ and $\rho = 1/96$.

semantic encoder, decoder, MIMO configuration, and restoration settings are kept unchanged. This comparison isolates the impact of CSI quality on the final image reconstruction performance. Less accurate channel estimates introduce stronger MMSE equalization mismatch, which appears as residual amplitude-phase distortion and inter-stream interference in the recovered semantic latent representation. Consequently, both perceptual similarity and pixel-level fidelity are degraded. With the proposed CRFM, the receiver achieves consistently better semantic reconstruction over the whole SNR range, indicating that the channel-estimation NMSE gain is effectively transferred to the semantic domain through more reliable equalization. Moreover, SRFM further improves the reconstruction quality by refining the residual latent distortion after equalization.

E. Ablation Study of Cascaded CRFM and SRFM for Image Reconstruction

Fig. 7 evaluates the image reconstruction performance of different receiver configurations on ImageNet-mini and CelebA, including AWGN reference schemes, imperfect-CSI MIMO receivers, perfect-CSI upper-bound receivers, and the JPEG-1/2LDPC-4QAM baseline. For the AWGN reference case, the degradation mainly comes from additive channel noise and SNR mismatch with the pretrained SwinT-JSCC model. The performance gain of AWGN+SRFM over the plain AWGN receiver shows that SRFM can effectively refine semantic latent perturbations caused by additive noise alone. In imperfect-CSI MIMO transmission, the LS-MMSE receiver suffers further degradation because pilot-induced channel-estimation errors lead to MMSE equalization mismatch, which introduces structured distortion into the recovered semantic latent representation. The ablation results show that CRFM and SRFM play complementary roles. Therefore, the cascaded CRFM-SRFM receiver achieves the best performance among all imperfect-CSI schemes and substantially approaches the perfect-CSI upper bound over a wide SNR range. The consistent improvements on both datasets demonstrate the effectiveness of the proposed cascaded restoration design in enhancing perceptual similarity and pixel-level fidelity. In contrast, the JPEG-1/2LDPC-4QAM baseline exhibits a typical low-SNR cliff effect, whereas the JSCC-based schemes degrade more gracefully. Fig. 8 provides representative visual examples. The LS-MMSE receiver produces visible artifacts and structural distortions, whereas CRFM alleviates channel-induced degradation and SRFM further improves semantic details and textures. The cascaded CRFM-SRFM receiver yields the most faithful reconstruction among the imperfect-CSI schemes, which is consistent with the quantitative results in Fig. 7.

TABLE I
COMPUTATIONAL COMPLEXITY AND INFERENCE LATENCY COMPARISON.

Method	MACs/Forward	Params	Latency
Score-based method	40.98M	2.68M	6.58s
DDIM-based method	41.05M	2.68M	831.19ms
CRFM	40.98M	2.68M	0.52ms
SRFM	21.8G	129.81M	35.78ms

F. Complexity and Latency Comparison

To evaluate the online receiver overhead from an implementation perspective, Table I reports the multiply–accumulate operations (MACs), the number of trainable parameters, and the inference latency of the considered receiver-side modules. MACs/Forward denotes the computational cost of one neural-network evaluation, whereas Latency denotes the average runtime of the complete inference procedure under the adopted sampling configuration. For a fair comparison,the score-based method, the DDIM-based method, and CRFM are implemented with the same UNet-based network backbone [12]. Therefore, these three methods have nearly identical single-forward computational costs and model sizes, with about 41M MACs and 2.68M trainable parameters. Their latency difference mainly comes from the number and type of sampling steps. Specifically, the latency of the score-based method includes the complete sampling chain with 50 noise levels and 3 Langevin correction steps at each level, resulting in 50×3 score-network evaluations. The latency of the DDIM-based method includes the full 20-step reverse sampling chain. In contrast, CRFM performs deterministic ODE inference with only $k_{\mathrm{H}} = 5$ Euler steps. Under the measured setting, CRFM achieves the lowest infercence latency among the compared generative channel estimators. The semantic restoration module SRFM has a larger model size than CRFM, with 21.8G MACs and 129.81M trainable parameters. This is expected because SRFM operates on the high-dimensional semantic latent representation rather than the low-dimensional channel tensor.

V. CONCLUSION

This paper investigated MIMO semantic communications under imperfect-CSI and formulated the receiver-side processing as two cascaded inverse recovery problems. To address the error propagation from channel estimation to semantic reconstruction, we proposed a cascaded restoration flow matching receiver consisting of CRFM and SRFM. CRFM refines the coarse LS channel estimate before MMSE equalization, while SRFM restores the residual distortion in the equalized semantic latent representation. Simulation results over MIMO channels and visual semantic transmission tasks showed that

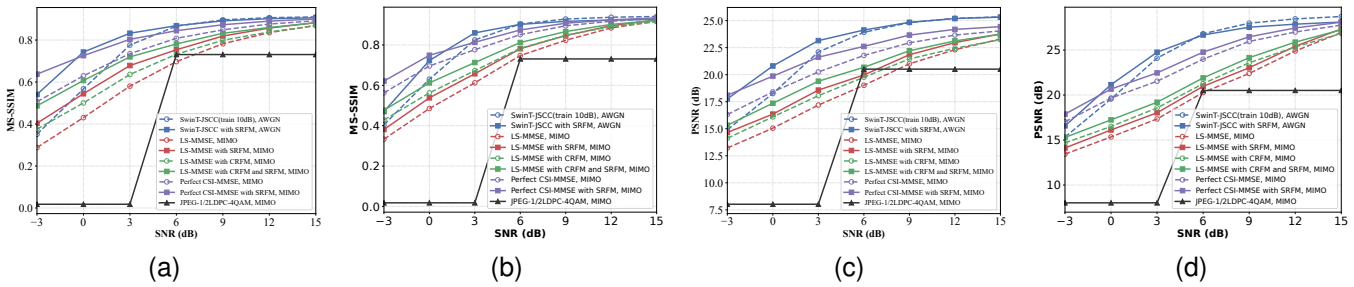

Fig. 7. Semantic reconstruction performance under different transmission and receiver schemes on the ImageNet-mini and CelebA datasets, with $T_p = 8$ and $\rho = 1/96$. (a) MS-SSIM versus SNR on ImageNet-mini. (b) MS-SSIM versus SNR on CelebA. (c) PSNR versus SNR on ImageNet-mini. (d) PSNR versus SNR on CelebA.

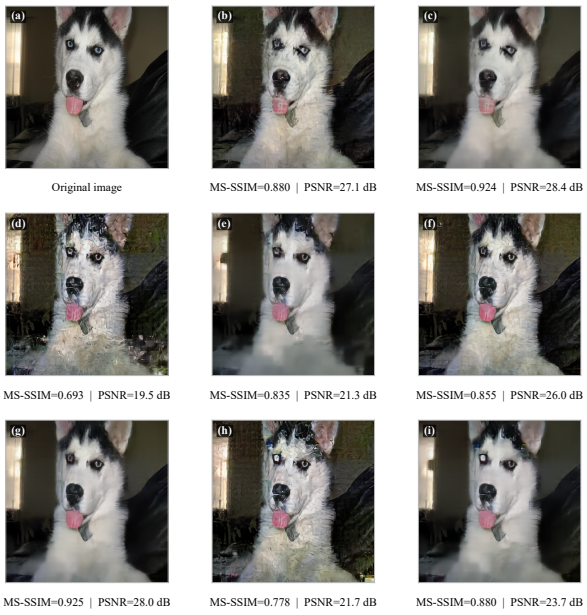

Fig. 8. Visual reconstruction examples at $\mathrm{SNR} = 3\,\mathrm{dB}$. (a) Original image; (b) SwinT-JSCC trained at $10\,\mathrm{dB}$ over AWGN; (c) SwinT-JSCC with SRFM over AWGN; (d) LS-MMSE over MIMO; (e) LS-MMSE with SRFM over MIMO; (f) perfect-CSI-MMSE over MIMO; (g) perfect-CSI-MMSE with SRFM over MIMO. (h) LS-MMSE with CRFM over MIMO; (i) LS-MMSE with CRFM and SRFM over MIMO; The MS-SSIM and PSNR values are reported below each reconstructed image.

the proposed method improves both physical-layer channel estimation and image reconstruction. The results confirm that channel restoration and semantic latent restoration are complementary for robust MIMO semantic reception. Future work will extend the proposed framework to time-varying and multi-user MIMO scenarios, and further explore adaptive sampling strategies and multimodal semantic transmission.

Appendix A

By the L^2 projection property of conditional expectation, the population minimizer of (58) is given by

$$\mathbf{v}^{\star}_{\tau_\ell,\tau_h}(\mathbf{s},t) = \mathbb{E}_{\mu^t_{\tau_\ell,\tau_h}}\left[\mathbf{u}\,|\,\mathbf{S}=\mathbf{s}\right]. \tag{66}$$

Using the mixture law in (54), the above conditional expectation can be decomposed as

$$\begin{aligned}&\mathbb{E}_{\mu^t_{\tau_\ell,\tau_h}}\left[\mathbf{u}\,|\,\mathbf{S}=\mathbf{s}\right]\\ &= \frac{\omega_\ell p^t_{\tau_\ell}(\mathbf{s})}{p^t_{\tau_\ell,\tau_h}(\mathbf{s})}\mathbb{E}_{\mu^t_{\tau_\ell}}\left[\mathbf{u}\,|\,\mathbf{S}=\mathbf{s}\right]\\ &\quad+\frac{\omega_h p^t_{\tau_h}(\mathbf{s})}{p^t_{\tau_\ell,\tau_h}(\mathbf{s})}\mathbb{E}_{\mu^t_{\tau_h}}\left[\mathbf{u}\,|\,\mathbf{S}=\mathbf{s}\right].\end{aligned} \tag{67}$$

From the marginal mixture density in (56), we have

$$p^t_{\tau_\ell,\tau_h}(\mathbf{s}) = \omega_\ell p^t_{\tau_\ell}(\mathbf{s}) + \omega_h p^t_{\tau_h}(\mathbf{s}). \tag{68}$$

Moreover, by the anchor-wise optimal velocity definition in (57),

$$\mathbb{E}_{\mu^t_{\tau_\ell}}\left[\mathbf{u}\,|\,\mathbf{S}=\mathbf{s}\right] = \mathbf{v}^{\star}_{\tau_\ell}(\mathbf{s},t), \qquad \mathbb{E}_{\mu^t_{\tau_h}}\left[\mathbf{u}\,|\,\mathbf{S}=\mathbf{s}\right] = \mathbf{v}^{\star}_{\tau_h}(\mathbf{s},t). \tag{69}$$

Substituting (68) and (69) into (67) yields

$$\begin{aligned}\mathbf{v}^{\star}_{\tau_\ell,\tau_h}(\mathbf{s},t) &= \frac{\omega_\ell p^t_{\tau_\ell}(\mathbf{s})}{\omega_\ell p^t_{\tau_\ell}(\mathbf{s}) + \omega_h p^t_{\tau_h}(\mathbf{s})}\mathbf{v}^{\star}_{\tau_\ell}(\mathbf{s},t)\\ &\quad+\frac{\omega_h p^t_{\tau_h}(\mathbf{s})}{\omega_\ell p^t_{\tau_\ell}(\mathbf{s}) + \omega_h p^t_{\tau_h}(\mathbf{s})}\mathbf{v}^{\star}_{\tau_h}(\mathbf{s},t).\end{aligned} \tag{70}$$

Recognizing the two coefficients in (70) as $\gamma_{\tau_\ell}(\mathbf{s},t)$ and $\gamma_{\tau_h}(\mathbf{s},t)$ defined in (60) and (61), respectively, we obtain (59). This completes the proof. ■

References

[1] P. Zhang et al., "Toward wisdom-evolutionary and primitive-concise 6G:A new paradigm of semantic communication networks," *Engineering.*, vol. 8, pp. 60–73, Jan. 2022.

[2] W. Yang et al., "Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges," *IEEE Commun. Surv. Tutorials.*, vol. 25, no. 1, pp. 213-250, Firstquarter, 2023.

[3] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. Vincent Poor, "Less Data, More Knowledge: Building Next-Generation Semantic Communication Networks," *IEEE Commun. Surv. Tutorials.*, vol. 27, no. 1, pp. 37-76, Feb. 2025.

[4] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, Jun. 2021.

[5] J. Dai et al., "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Aug. 2022.

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

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[9] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, Apr. 2022.

[10] M. Ding, J. Li, M. Ma, and X. Fan, "SNR-adaptive deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 1555–1559.

[11] K. Yang, S. Wang, J. Dai, X. Qin, K. Niu, and P. Zhang, "SwinJSCC: Taming Swin Transformer for Deep Joint Source-Channel Coding," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 1, pp. 90–104, Feb. 2025.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[13] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4195–4205.

[14] S. Welker, H. N. Chapman, and T. Gerkmann, "DriftRec: Adapting Diffusion Models to Blind JPEG Restoration," *IEEE Trans. Image Process.*, vol. 33, pp. 2795–2807, 2024.

[15] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, M. Tao, and W. Zhang, "CDDM: Channel denoising diffusion models for wireless semantic communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11168–11183, Sep. 2024.

[16] B. Xu, S. Han, X. Xu, W. Li, R. Meng, C. Dong, and P. Zhang, "Semantic prior aided channel-adaptive equalizing and de-noising semantic communication system with latent diffusion model," *IEEE Trans. Wireless Commun.*, vol. 24, no. 6, pp. 4614–4630, Jun. 2025.

[17] T. Wu, Z. Chen, D. He, F. Yang, M. Tao, X. Xu, W. Zhang, and P. Zhang, "ICDM: Interference cancellation diffusion models for wireless semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 2528–2543, 2026.

[18] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 42–48, Nov. 2023.

[19] R. Chataut and R. Akl, "Massive MIMO systems for 5G and beyond networks—Overview, recent trends, challenges, and future research direction," *Sensors.*, vol. 20, no. 10, Art. no. 2753, May 2020.

[20] C. Bian, Y. Shao, H. Wu, and D. Gündüz, "Space-time design for deep joint source-channel coding of images over MIMO channels," in *Proc. IEEE 24th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2023, pp. 616–620.

[21] Y. Duan, T. Wu, Z. Chen, and M. Tao, "DM-MIMO: Diffusion models for robust semantic communications over MIMO channels," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Hangzhou, China, 2024, pp. 1609–1614.

[22] H. Wu, Y. Shao, C. Bian, K. Mikolajczyk, and D. Gündüz, "Deep joint source-channel coding for adaptive image transmission over MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 15002–15017, Oct. 2024.

[23] J. Park, Y. Oh, J. Park, and Y.-S. Jeon, "Importance-aware semantic communication in MIMO-OFDM systems using vision transformer," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 13494–13510, Mar. 2026.

[24] M. K. Ozdemir and H. Arslan, "Channel estimation for wireless ofdm systems," *IEEE Commun. Surv. Tutorials.*, vol. 9, no. 2, pp. 18-48, Second Quarter 2007.

[25] C. Ruan, Z. Zhang, H. Jiang, Y. Qi, J. Dang, and L. Wu, "Low complexity orthogonal matching pursuit-based near-field channel estimation in XL-MIMO systems," *IEEE Commun. Lett.*, vol. 28, no. 12, pp. 2859–2863, Dec. 2024.

[26] C. Liu, X. Liu, D. W. K. Ng, and J. Yuan, "Deep residual learning for channel estimation in intelligent reflecting surface-assisted multi-user communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 898–912, Feb. 2022.

[27] M. Arvinte and J. I. Tamir, "MIMO Channel Estimation Using Score-Based Generative Models," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3698-3713, Jun. 2023.

[28] X. Zhou, L. Liang, J. Zhang, P. Jiang, Y. Li, and S. Jin, "Generative Diffusion Models for High Dimensional Channel Estimation," *IEEE Trans. Wireless Commun.*, vol. 24, no. 7, pp. 5840-5854, Jul. 2025.

[29] Y. Feng, Z. Shan, H. Shen, X. Shi, Q. Yang, and J. Chen, "Digital semantic communication system with a learnable channel estimator for OFDM transmission," *IEEE Trans. Cognit. Commun. Netw.*, vol. 11, no. 5, pp. 3361–3370, 2025.

[30] Z. Weng, Z. Qin, H. Xie, X. Tao, and K. B. Letaief, "Semantic MIMO systems for speech-to-text transmission," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18697–18710, Dec. 2024.

[31] Y. Sun, H. Shen, W. Xu, N. Hu and C. Zhao, "Robust MIMO Detection With Imperfect CSI: A Neural Network Solution," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 5877-5892, Oct. 2023.

[32] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[33] X. Ma, Y. Xin, Y. Ren, D. Burghal, H. Chen, and J. C. Zhang, "Diffusion model based channel estimation," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 1159–1164.

[34] B. Fesl, M. Baur, F. Strasser, M. Joham, and W. Utschick, "Diffusion-based generative prior for low-complexity MIMO channel estimation," *IEEE Wireless Commun. Lett.*, vol. 13, no. 12, pp. 3493–3497, Dec. 2024.

[35] Z. Chen, H. Shin and A. Nallanathan, "Generative Diffusion Model-Based Variational Inference for MIMO Channel Estimation," *IEEE Trans. Commun.*, vol. 73, no. 10, pp. 9254-9269, Oct. 2025.

[36] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[37] W. Liu, N. Ma, J. Chen et al., "Flow matching-based generative models for MIMO channel estimation," *arXiv preprint* arXiv:2511.10941, 2025.

[38] J. Fu, M. Xiao, M. Skoglund, and D. In Kim, "Land-then-transport: A Flow Matching-Based Generative Decoder for Wireless Image Transmission," *arXiv preprint*, arXiv:2601.07512, 2026.

[39] G. Kerrigan, G. Migliorini and P. Smyth, "Dynamic Conditional Optimal Transport through Simulation-Free Flows," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024.

[40] K. Kim and J. C. Ye, "Noise2Score: Tweedie's approach to self-supervised image denoising without clean images," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021.

[41] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon and B. Poole, "Score-Based Generative Modeling through Stochastic Differential Equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[42] Y. Song and S. Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[43] Study on Channel Model for Frequencies From 0.5 to 100 GHz, 3GPP,document 38.901, 2020.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, 2009, pp. 248–255.

[45] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015.

[46] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc.20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.