视觉语言模型和世界模型赋能的视频语义通信

董莉[1]，邹健[1]，江沸菠[2]，王可之[3]，潘存华[4]
（1.湖南工商大学计算机学院，湖南 长沙 410205；2.湖南师范大学信息科学与工程学院，湖南 长沙 410081；3.英国伦敦布鲁内尔大学计算机科学系，英国 伦敦 UB8 3PH；4.东南大学移动通信国家重点实验室，中国 南京 210096）

摘要：智能视频业务的快速发展对无线视频传输的效率、鲁棒性和语义保真度提出了更高要求。然而，现有视频传输仍面临语义冗余高、生成式重构失真以及数据分布易偏移等问题。为此，本文提出一种视觉语言模型和世界模型赋能的多任务视频语义通信系统。在发送端，视觉语言模型将原始视频压缩为关键帧与文本语义，以减少冗余传输并保留核心内容。在接收端，世界模型基于接收的关键帧和文本语义重建高质量视频。针对动态环境中的任务数据分布偏移问题，本文进一步提出持续语义适应机制，通过跨任务优化对语义/信道编解码器进行稳定更新，从而保持跨任务语义理解能力。多种视频任务和信道条件下的实验结果表明，所提出的视频语义通信系统具有较高的语义一致性、稳定性和泛化能力。

关键词：世界模型；视觉语言模型；语义通信；视频语义通信；多任务学习
中图分类号：TN914
文献标志码：A

Vision-Language Model- and World Model-Empowered Video Semantic Communication

Dong Li[1], Zou Jian[1], Jiang Feibo[2], Wang Kezhi[3], Pan Cunhua[4]
1. School of Computer Science, Hunan University of Technology and Business, Changsha 410205, China
2. School of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
3. Department of Computer Science, Brunel University London, London UB8 3PH, UK
4. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

Abstract: The rapid development of intelligent video services has imposed increasingly stringent requirements on the efficiency, robustness, and semantic fidelity of wireless video transmission. However, existing video transmission systems still face several challenges, including substantial semantic redundancy, difficulties in generative reconstruction, and vulnerability to data distribution shifts. To address these challenges, a multi-task video semantic communication system empowered by a vision-language model (VLM) and a world model (WM) was proposed. At the transmitter, the original video was compressed by the VLM into keyframes and textual semantic descriptions, such that transmission redundancy was reduced while essential semantic content was preserved. At the receiver, high-quality videos were reconstructed by the WM based on the received keyframes and textual semantic descriptions. To mitigate task-specific data distribution shifts in dynamic environments, a continual semantic adaptation mechanism was further developed, through which the semantic and channel encoders and decoders were stably updated using cross-task optimization to preserve semantic understanding across different tasks. High semantic consistency, stability, and generalization capability under different video tasks and channel conditions were demonstrated by the experimental results.

Key words: World model, vision-language model, semantic communication, video semantic communication, multi-task learning

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX

0 引言

近年来，面向6G的高清视频、沉浸式交互和智能感知等业务快速发展，视频通信正由视觉信号传输向语义理解与内容交互演进，对高可靠、低开销及复杂信道下的稳定传输提出更高要求[1-2]。传统视频通信以信号保真为目标，缺乏对视频语义重要性的区分能力，难以主动识别任务相关的关键目标、动作关系和场景变化。在低带宽或信道波动场景下，传统方法往往需传输大量冗余视频帧以维持视觉质量，传输效率和资源利用率仍受到限制。

为缓解上述问题，视频语义通信逐渐成为研究热点[3]。其核心思想是在发送端提取与任务或用户感知相关的语义信息，并在接收端基于接收到的语义内容完成视频恢复或任务推理[4]。与文本、图像和语音等模态相比，视频数据不仅包含单帧画面中的目标、场景和动作语义，还包含帧间运动关系、主体交互过程和场景状态演化。同时，实际应用中的视频内容还会随着任务、场景和数据分布变化而不断改变[5]。因此，面向视频语义通信系统设计，仍需重点解决以下挑战：

（1）**视频语义冗余：**视频数据包含丰富的空间纹理、目标对象、动作行为和场景关系，其中既有与任务相关的核心语义，也存在大量冗余视觉信息[6]。传统视频传输方式主要依赖像素级压缩，难以从内容理解层面判断哪些信息应被保留、哪些信息可以舍弃。因此，如何从高维视频中提取紧凑、准确且任务相关的语义表示，是视频语义通信首先需要解决的问题。

（2）**生成式重构失真：**为降低视频传输负载，生成式视频语义通信通常只传输文本语义或紧凑视觉特征。然而，有限的语义与视觉线索难以完整描述连续运动、主体交互和场景状态变化。接收端若缺乏对视频动态规律的建模与推演能力，生成结果容易出现动作不连贯、主体关系变化不自然等问题。

（3）**分布偏移干扰：**实际视频语义通信系统往往部署于开放动态环境中，训练阶段覆盖的任务类型、动作类别和场景先验难以完全匹配后续应用分布。当输入视频的语义分布发生变化时，语义编码器与解码器已学习的跨模态映射关系可能失稳，导致新任务下的语义编码与视频重构性能下降；若直接利用新任务数据更新模型，又可能破坏已有语义表征，导致灾难性遗忘[7]。

为应对上述挑战，近年来已有研究开始将视觉语言模型（vision-language model，VLM）和世界模型（world model，WM）引入视频语义通信中，以提升视频内容理解、语义压缩和动态重构能力。VLM 侧重于跨模态语义理解，能够理解图像或视频中的目标、动作、场景及其关联关系，并提取更加准确、紧凑的高层语义表示[8]；WM则侧重于环境状态、因果关系及动态演化规律的世界建模，能够基于已有语义和关键视觉线索推演视频的后续发展变化，从而增强接收端对未传输帧或受损信息的预测、补全与一致性重构能力。

基于上述思想，现有工作已分别从VLM赋能的语义压缩与WM驱动的动态重构两个方向展开探索。在VLM方面，Zhao等人[9]提出LaMoSC，将VLM引入视觉语义通信系统，通过融合视觉特征与文本语义提升低信噪比条件下的视觉信息传输性能；Liang等人[10]则将VLM用于Massive MIMO语义通信，在发送端生成图像文本描述并优先传输语义文本，以降低带宽占用并保持较高的图像重构质量。在WM方面，Jiang等人[11]提出基于WM的语义通信框架，通过当前帧和文本语义预测未来帧，从而减少不必要的视频帧传输；Lokumarambage等人[12]进一步采用潜在预测机制，在接收端预测未来视频状态以降低传输开销。

然而，二者在视频语义通信中的协同应用仍缺乏深入研究。为此，本文构建VLM与WM协同的语义通信框架：发送端VLM提取文本语义与关键帧并传输，接收端利用恢复信息驱动WM完成时序一致的视频重构，并引入CSA缓解连续任务中的分布偏移与灾难性遗忘。主要贡献如下：

（1）**跨模态理解：**针对视频语义冗余问题，本文在发送端引入VLM对原始视频进行跨模态理解。该模块能够从视频中识别关键目标、动作行为、场景属性和代表性视觉信息，并将高冗余的视频帧序列转化为文本语义和关键帧等紧凑表示。与传统像素级压缩不同，该方法更加关注任务相关语义的筛选与表达，从而在减少传输数据量的同时保持视频内容的核心语义完整性。

（2）**生成式推演：**针对视频重构失真问题，本文在接收端引入WM进行视频推演与重构。该模型根据接收到的文本语义、关键帧和潜在视觉特

征，学习视频场景中的状态变化、动作连续关系和主体交互过程，并对未传输或受损的信息基于视觉知识进行预测与补全。通过这种方式，接收端不再只是被动解码接收信号，而是能够基于语义先验主动推演符合物理世界的视频演化过程，提高重构视频的时序一致性和语义连贯性。

（3）**持续语义适应：**针对分布偏移干扰问题，本文提出持续语义适应机制。该机制通过持续学习对信源和信道编解码器进行联合调整，并利用样本级自适应加权策略区分不同类型的训练样本：一方面增强模型对新任务语义分布的适应能力，另一方面保持对已有任务语义表征的记忆能力。通过这种方式，系统能够在学习新场景和新任务时减少灾难性遗忘，提升视频语义通信系统在动态环境中的稳定性和泛化能力。

1　系统模型

本文构建了一种端到端视频语义通信系统。不同于以逐帧像素保真为目标的传统视频通信，本系统更强调视频语义的准确传递与一致表达。发送端利用 VLM 对原始视频进行跨模态理解，提取场景、目标和动作关系等高阶文本语义，并筛选关键帧作为视觉结构线索；随后，双语义编码器对文本语义与关键帧信息进行压缩，并经信道编码映射为无线传输符号。接收端完成解码后，恢复高阶文本语义和低阶关键像素，并借助WM进行语义引导的视频重构，生成时空连贯的视频内容。

1.1　发送端

将发送端的原始视频序列记为 $\mathbf{V}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T\}$，其中 $\mathbf{x}_t\in\mathbb{R}^{H\times W\times C}$ 表示第 t 帧视频，视频序列 $\mathbf{V}\in\mathbb{R}^{T\times H\times W\times C}$，其中 T 表示帧数，H、W 和 C 分别表示帧高度、帧宽度和通道数。

为实现高效的视频语义压缩，本文采用语义提取器从视频V中获取两类信息：(1) 高阶文本语义（即视频场景与动作的文本语义描述 $\mathbf{s}$）；(2) 低阶关键像素（即保持视觉连贯性的关键帧 $\mathbf{f}$）。语义提取器的映射关系可表示为：

$$(\mathbf{s},\mathbf{f})=K_\theta(\mathbf{V},\mathbf{p}) \tag{1}$$

其中，K_θ 表示参数 θ 的语义提取器，$\mathbf{p}$ 是用于控制语义生成的提示词。接着，采用双语义编码器分别对文本语义 $\mathbf{s}$ 和关键帧 $\mathbf{f}$ 进行编码，将高阶与低阶语义信息投影至紧凑的潜在语义空间。编码输出随后被联合输入信道编码器，生成适用于无线传输的符号序列，具体形式如下：

$$\mathbf{z}=C_\gamma(S_\alpha(\mathbf{s}),S_\beta(\mathbf{f})) \tag{2}$$

其中，$\mathbf{z}$ 表示经过语义编码和信道编码后的符号序列；$S_\alpha(\cdot)$ 和 $S_\beta(\cdot)$ 分别为参数 α 和 β 的高阶与低阶语义编码器；$C_\gamma(\cdot,\cdot)$ 表示参数 γ 的信道编码器。

1.2　无线信道

发送端生成的符号序列 $\mathbf{z}$ 经无线信道传输后，接收端接收到的信号可表示为以下形式：

$$\hat{\mathbf{z}}=\mathcal{H}(\mathbf{z})+\mathbf{n} \tag{3}$$

其中 $\mathcal{H}(\mathbf{z})$ 为信道算子，$\mathbf{n}$ 代表加性高斯白噪声。

1.3　接收端

与发送端架构类似，接收端由信道解码器、语义解码器及语义重构器组成。信道解码器将接收到的符号序列映射为语义特征，其中高阶语义解码器负责重构文本语义，低阶语义解码器则负责重构视觉（关键帧）语义。解码过程可表述如下：

$$\hat{\mathbf{s}}=S_\varepsilon^{-1}\left(C_\delta^{-1}(\hat{\mathbf{z}})\right),\quad \hat{\mathbf{f}}=S_\eta^{-1}\left(C_\delta^{-1}(\hat{\mathbf{z}})\right) \tag{4}$$

其中，$C_\delta^{-1}(\cdot)$ 表示参数为 δ 的信道解码器，$S_\varepsilon^{-1}(\cdot)$ 和 $S_\eta^{-1}(\cdot)$ 分别表示参数为 ε 和 η 的高阶语义解码器与低阶语义解码器。此处 $\hat{\mathbf{s}}$ 和 $\hat{\mathbf{f}}$ 分别表示恢复的文本语义与重构的关键帧语义特征。

随后，语义重构器基于恢复的文本语义 $\hat{\mathbf{s}}$ 和低阶语义特征 $\hat{\mathbf{f}}$ 生成最终视频，其表达方式如下：

$$\hat{\mathbf{V}}=K_{\theta'}^{-1}\left(\hat{\mathbf{s}},\hat{\mathbf{f}}\right) \tag{5}$$

其中，$K_{\theta'}^{-1}(\cdot)$ 表示由 θ' 参数化的语义重构器，$\hat{\mathbf{V}}$ 为重构后的视频序列。

1.4　损失函数

为确保恢复过程中文本语义的一致性，论文引入交叉熵（CE）损失函数来量化 $\mathbf{s}$ 与 $\hat{\mathbf{s}}$ 之间的差异，从而确保传输与重构过程中的语义一致性。L 表示语义序列中的词元总数，w_l 表示原始语义序列中第 l 个真实词元，$\mathbf{y}$ 表示接收端输入语义解码器的条件语义特征，$p(w_l|\mathbf{y})$ 表示在条件语义特征 $\mathbf{y}$ 给定的情况下，语义解码器预测第 l 个真实词元 w_l 的概率。交叉熵损失函数定义为：

$$\mathcal{L}_{\mathrm{CE}}=-\frac{1}{L}\sum_{l=1}^{L}\log p\left(w_l|w_{<l},\mathbf{y}\right) \tag{6}$$

与此同时，为提升关键帧的重建质量，论文采用均方误差（MSE）损失函数，其定义如下：

$$\mathcal{L}_{\mathrm{MSE}}(\mathbf{f},\hat{\mathbf{f}}) = \frac{1}{N_p}\left\| \mathbf{f} - \hat{\mathbf{f}} \right\|_2^2 \tag{7}$$

其中 N_p 表示视觉特征维度，$\mathbf{f}$ 和 $\hat{\mathbf{f}}$ 分别表示真实值和预测值。

为联合优化文本与视觉语义的恢复性能，论文定义了以下组合损失函数：

$$(\alpha^*,\beta^*,\gamma^*,\delta^*,\varepsilon^*,\eta^*) = \arg\min_{\alpha,\beta,\gamma,\delta,\varepsilon,\eta} \mathbb{E}_{p(\hat{\mathbf{s}},\mathbf{s},\hat{\mathbf{f}},\mathbf{f})} \left[\lambda_1 L_{\mathrm{CE}}(\mathbf{s},\hat{\mathbf{s}}) + \lambda_2 L_{\mathrm{MSE}}(\mathbf{f},\hat{\mathbf{f}}) \right] \tag{8}$$

其中，α^*、β^*、γ^*、δ^*、ε^* 和 η^* 分别表示高阶语义编码器、低阶语义编码器、信道编码器、信道解码器、高阶语义解码器和低阶语义解码器的最优参数。权重系数 λ_1 和 λ_2 用于平衡文本语义与视觉语义的目标函数，$p(\cdot)$ 表示用于建模语义变量间依赖关系的联合概率分布。文本语义和关键帧在进入 WM 之前，分别通过交叉熵损失和均方误差损失约束其发送端与接收端恢复结果之间的一致性，从而形成 WM 输入端的显式监督。在 WM 内部，文本语义通过交叉注意力约束生成内容，关键帧通过视觉潜在条件提供主体外观、场景布局和时空锚点，二者共同作用于 DiT 的速度场预测，形成双条件生成的隐式约束。

2　VLM 与 WM 赋能的视频语义通信

2.1　系统概述

本系统以跨模态语义交互为核心，整合了以下三大模块：（1）基于 VLM 的高/低阶语义提取模块、（2）具备持续学习能力的双流语义编码/解码模块、以及（3）基于 WM 架构的语义引导视频重建模块（如图 1 所示）。整个工作流程包含五个协同运作的关键阶段。

2.1.1　视频语义提取

为实现更高效的视频压缩并提升源信息密度，发送端采用基于 VLM 的视频语义提取模块。该模块包含图像编码器、视频编码器及小型语言模型（small language model, SLM）。图像编码器与视频编码器首先提取语义线索、时空结构及视觉特征，并将关键帧作为低阶关键像素[13]。随后，基于视觉特征输入， SLM 生成概括视频场景与动作的文本语义描述，构成高级语义信息。

2.1.2　双语义编码

语义提取器生成的文本语义和关键帧会经过双语义编码器进行进一步压缩与编码处理。其中文本语义由 Transformer 编码器处理，关键帧则由 ViT 编码器处理。此外，编码器中还整合了多任务学习模块。根据任务特性，将样本划分为高性能、高方差、过拟合、噪声及正常类别，并分别赋予学习权重和知识保留权重，从而实现稳定高效的多任务优化。

2.1.3　信道编码与传输

经过编码处理的语义向量通过信道编码和调制映射为信道符号，从而确保在复杂信道环境下实现稳定可靠的传输。信号经物理信道传输至接收端后，接收端的解码器会将信号重新转换回语义特征空间。该信道编码/解码架构采用堆叠式自编码器实现，这种设计实现了可微分的端到端通信流程。

2.1.4　双语义解码

该语义解码器由基于 Transformer 的文本解码模块和基于 ViT 的图像解码模块组成。Transformer 解码器通过信道解码后的符号序列重构文本语义信息，而 ViT 解码器则还原关键帧图像，从而分别恢复高阶与低阶语义特征。

2.1.5　高保真视频生成

采用 WM 作为理解和推演视频的重构器，以实现高质量视频重构。首先，接收的关键帧经图像编码器编码为视觉隐变量，同时语义文本由文本编码器转化为文本嵌入向量。随后，对视觉隐变量与文本嵌入向量进行融合，构建条件向量并输入至基于流匹配的 DiT 网络。DiT 网络从高噪声潜在变量出发，预测当前噪声水平下的速度场，并由调度器沿该速度场逐步去噪至干净视频潜在表示。最后，利用 VAE 解码器对生成的隐式表征进行逐步上采样与解码，输出高质量视频帧序列。关键帧作为视觉和时空锚点，在生成全过程中约束视觉信息与语义的一致性，进而保证重构视频的高保真度与时序连续性。

2.2　基于 VLM 的视频语义提取器

如图 2 所示，在发送端，基于 VLM 的视频语义提取器用于将原始视频压缩为两类互补语义信息：一类是高阶文本语义描述，用于刻画视频中的主体目标、动作行为和场景关系；另一类是低阶关键帧集合，用于保留视频重构所需的主要视觉结构。该模块主要由双视觉编码器、特征投影融合模块和小语言模型组成，其目标是在降低传输负载的

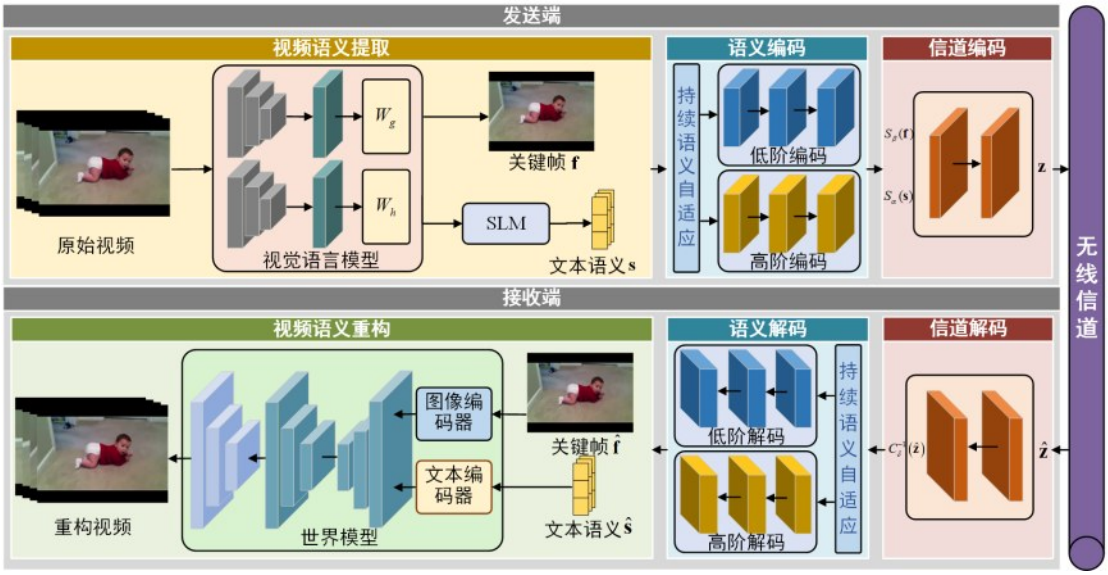

图1　基于VLM和WM的视频语义通信系统框架图

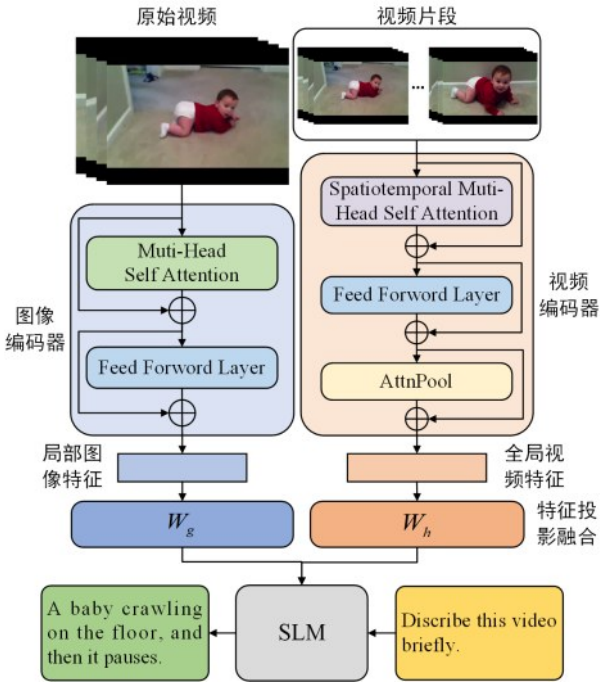

图2　基于VLM的视频语义提取器结构

同时保持视频内容的核心语义。

给定输入视频序列 $\mathbf{V}=\{\mathbf{x}_t\}_{t=1}^{T}$ 首先，将视频划分为 K 个时间片段，第 k 个片段的帧索引集合可表示为 $I_k=\left\{\left\lfloor\frac{(k-1)T}{K}\right\rfloor+1,\cdots,\left\lfloor\frac{kT}{K}\right\rfloor\right\}$ 随后，对每个时间片段进行降采样处理，得到低分辨率视频片段：

$$\tilde{\mathbf{V}}_k=D\left(\{\mathbf{x}_t|t\in I_k\}\right),\quad k=1,\cdots,K \tag{9}$$

其中 $D(\cdot)$ 表示降采样算子。通过片段级采样，系统能够在降低视频编码复杂度的同时保留主要时空演化信息。

随后，原始视频帧序列和降采样视频片段分别输入图像编码器 $\Phi_I(\cdot)$ 与视频编码器 $\Phi_V(\cdot)$，用于提取帧级空间语义特征和片段级时空语义特征：

$$\begin{aligned}\mathbf{H}^{\mathbf{I}}&=\Phi_I\left(\{\mathbf{x}_t\}_{t=1}^{T}\right),\\ \mathbf{H}^{\mathbf{V}}&=\Phi_V\left(\{\tilde{\mathbf{V}}_k\}_{k=1}^{K}\right)\end{aligned} \tag{10}$$

其中，图像编码器 $\Phi_I(\cdot)$ 采用基于ViT的帧级编码结构。每一帧图像首先被划分为若干非重叠图像块，并通过线性嵌入和空间位置编码转换为图像词元序列；随后，图像词元经过多层 Transformer 编码块，通过多头自注意力机制建模不同图像区域之间的空间关联关系，最终得到帧级局部图像语义特征。该编码器主要用于保留视频中的目标外观、场景布局和局部视觉细节。

视频编码器 $\Phi_V(\cdot)$ 采用时空 Transformer 结构[14]，用于建模视频片段中的动作变化和时间上下文。首先，每个降采样视频片段被划分为若干时空块。随后，时空块经过多层时空 Transformer 编码块，通过空间多头自注意力机制同时捕获同一帧内的空间相关性和不同帧之间的时间依赖关系，最终得到片段级时空语义特征。通过上述双编码结构，图像编码器侧重于高分辨率空间细节建模，视频编码器侧重于时间动态关系建模，二者共同为后续跨模态语义提取提供互补信息。

为实现视觉特征与语言空间的对齐，特征投影融合模块通过两个可训练投影层将图像特征和视频特征映射为语言嵌入：

$$\mathbf{E}^{\text{img}} = W_g(\mathbf{H}^{\text{I}}),\quad \mathbf{E}^{\text{vid}} = W_h(\mathbf{H}^{\text{V}}) \tag{11}$$

其中，$W_g(\cdot)$和$W_h(\cdot)$分别表示图像特征投影层和视频特征投影层，$\mathbf{E}^{\text{img}}$和$\mathbf{E}^{\text{vid}}$分别表示语言空间中的图像嵌入和视频嵌入。为降低训练复杂度并保持预训练视觉表征能力，本文冻结图像编码器和视频编码器的主体参数，仅更新特征投影融合模块及后续语义生成模块。

最后，将图像嵌入、分段视频嵌入和用户文本查询嵌入进行拼接，并输入轻量级语言模型生成视频语义描述：

$$\mathbf{s} = \text{SLM}\left(\left[\mathbf{E}^{\text{img}},\mathbf{E}_1^{\text{vid}},\cdots,\mathbf{E}_k^{\text{vid}},\mathbf{E}^{\text{text}}\right]\right) \tag{12}$$

其中，$\mathbf{s}$表示生成的视频语义文本，$\mathbf{E}^{\text{text}}$表示由用户提示词得到的文本查询嵌入，$\mathbf{E}_k^{\text{vid}}$表示第$k$个视频片段对应的语言空间嵌入。该过程使空间细节、时间上下文和用户查询能够在统一语言空间中进行融合，从而生成更加紧凑且任务相关的视频语义描述。同时，系统将每个视频片段的首帧作为关键帧集合$\mathbf{f}$。

2.3　基于WM的视频语义重构器

WM包括世界规划器、世界动作模型、世界合成器、世界模拟器四大类型。基于接收端恢复的语义文本$\hat{\mathbf{s}}$和关键帧$\hat{\mathbf{f}}$，如图3所示，本文采用世界模拟器作为视频语义重构器，在潜在空间中进行条件视频生成。与直接在像素空间逐帧重建不同，WM首先将文本语义和关键帧映射到统一条件潜在空间，再通过基于 DiT 的流匹配生成过程恢复完整视频序列。

2.3.1　潜在空间编码与去噪

首先，文本编码器将恢复的语义文本$\hat{\mathbf{s}}$映射为词元级文本潜在特征：

$$\mathbf{Z}_{\text{text}} = E_{\text{text}}(\hat{\mathbf{s}}),\quad \mathbf{Z}_{\text{text}} \in \mathbb{R}^{T_s \times d_s} \tag{13}$$

其中，$E_{\text{text}}(\cdot)$表示文本编码器，T_s为文本词元数量，d_s为文本嵌入维度。与此同时，VAE 编码器将关键帧$\hat{\mathbf{f}}$映射为词元级视觉潜在特征：

$$\mathbf{Z}_{\text{vision}} = E_{\text{VAE}}(\hat{\mathbf{f}}),\quad \mathbf{Z}_{\text{vision}} \in \mathbb{R}^{T_f' \times H' \times W' \times C'} \tag{14}$$

其中，T_f'、H'、W'和C'分别表示压缩后的潜在时间长度、高度、宽度和通道数。该视觉潜在特征保留关键帧中的主体外观、场景布局和局部结构信息，用于约束后续视频生成过程。

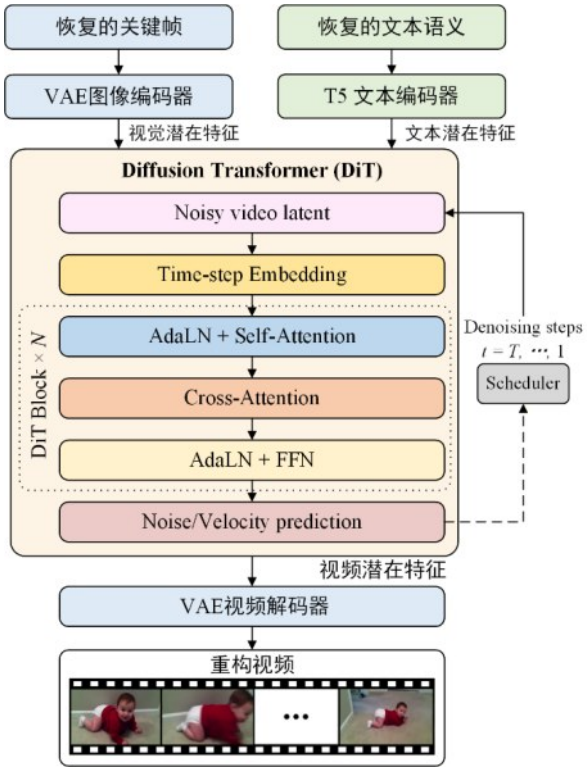

图3　视频语义重构器结构

为在潜在空间中生成长度为T_v的视频序列，将在流匹配采样i步时的潜在变量表示为$\mathbf{z}_{\sigma_i} \in \mathbb{R}^{T_v' \times H' \times W' \times C}$，其中$\sigma_i$表示对应的噪声强度。预测过程从高噪声状态$\mathbf{z}_{\sigma_N}$开始，逐步推理至$\mathbf{z}_{\sigma_0}$。随后按以下步骤构建初始潜在状态：

$$\mathbf{z}_{\sigma_N} = \grave{\mathbf{o}},\quad \grave{\mathbf{o}} \sim \mathcal{N}(0,\mathbf{I}), \tag{15}$$

其中，ϵ的维度与$\mathbf{z}_{\sigma_N}$相同，$\mathbf{I}$为单位协方差矩阵。

在第i个预测步骤中，DiT流场预测网络以$\mathbf{z}_{\sigma_i}$作为主要输入，并注入文本条件$\mathbf{Z}_{\text{text}}$和关键帧条件$\mathbf{Z}_{\text{vision}}$作为外部条件信号，用于在噪声水平$\sigma_i$下预测速度场$\hat{\mathbf{v}}_i$，$\mathcal{A}(\cdot)$表示对齐约束算子，$F_\theta(\cdot;\cdot)$表示参数$\theta$的DiT流场预测网络：

$$\hat{\mathbf{v}}_i = F_\theta(\mathbf{z}_{\sigma_i},\sigma_i;\mathbf{Z}_{\text{text}},\mathcal{A}(\mathbf{Z}_{\text{vision}})), \tag{16}$$

随后，根据当前噪声潜空间和模型预测的速度场$\hat{\mathbf{v}}_i$，先估计一个去噪潜空间$\hat{\mathbf{z}}_{0,i}$。扩散调度器$\mathcal{S}(\cdot)$按以下方式执行一步更新得到$\mathbf{z}_{\sigma_{i-1}}$：

$$\begin{aligned}\hat{\mathbf{z}}_{0,i} &= \mathbf{z}_{\sigma_i} - \sigma_i\hat{\mathbf{v}}_i,\\ \mathbf{z}_{\sigma_{i-1}} &= \mathcal{S}\left(\mathbf{z}_{\sigma_i},\hat{\mathbf{z}}_{0,i},\hat{\mathbf{v}}_i,\sigma_i,\sigma_{i-1}\right)\end{aligned} \tag{17}$$

经过N个采样步后，潜变量沿噪声水平序列$(1=\sigma_N \to \sigma_{N-1} \to \cdots \to \sigma_0 = 0)$，由高噪声状态逐步演化至干净视频潜在表示。最终生成的潜在变量可表示为：

$$\mathbf{z}_{\sigma_0} = \mathcal{S}^{(N)}(\mathbf{z}_{\sigma_N}, \{\hat{\mathbf{v}}_i\}_{i=N}^{1}) \tag{18}$$

其中，$\mathcal{S}^{(N)}(\cdot)$表示采样调度器在N个噪声水平上执行的复合更新过程。

2.3.2 DiT速度场预测

DiT采用基于Transformer的去噪架构。该模型通过输入$\mathbf{z}_{\sigma_i}$数据构建时空上下文信息，并在每个时间步i通过交叉注意力机制将$\mathbf{Z}_{\text{text}}$特征注入视频特征中以实现指令跟随功能。同时，时间步嵌入向量通过AdaLN[15]调节自注意力子层与前馈子层的参数，使去噪强度能够适应不同时间步i的噪声特征。在此架构下，Transformer输出对应上述方程中的速度场预测$\hat{\mathbf{v}}_i$。调度器根据当前噪声水平σ_i与预测$\hat{\mathbf{v}}_i$对潜变量进行更新，从而逐步将高噪声潜变量降噪至干净视频潜在表示。

2.3.3 VAE解码与视频重建

流匹配采样结束后，最终生成的潜在变量$\mathbf{z}_{\sigma_0}$通过VAE视频解码器$D_{\text{VAE}}(\cdot)$映射回像素空间：

$$\hat{\mathbf{V}} = D_{\text{VAE}}(\mathbf{z}_{\sigma_0}) \tag{19}$$

2.4 基于CSA的多任务联合编码

在动态环境中，模型通常会连续面对不同任务、场景和数据分布。若仅基于固定数据集训练语义编解码器，当后续视频内容、动作类别或场景先验发生变化时，模型已学习的语义映射关系可能失稳，导致新任务适应能力不足；而直接使用新任务数据微调模型，又容易破坏已有语义表征，引发灾难性遗忘。为此，本文提出持续语义适应机制（continual semantic adaptation, CSA），通过邻近知识保留和样本级自适应加权，在新任务学习与旧知识保持之间建立动态平衡。

在持续学习场景下，设系统依次接收任务序列$\{T_1, T_2, \cdots, T_t\}$。其中，第$t$个任务的数据集表示为：

$$\mathcal{D}_t = \left\{\left(\mathbf{s}_i^t, \mathbf{f}_i^t\right)\right\}_{i=1}^{N_t} \tag{20}$$

其中，$\mathbf{s}_i^t$表示第i个样本的视频字幕语义，$\mathbf{f}_i^t$表示对应关键帧集合，N_t为任务T_t中的样本数量。对于任意样本$\mathbf{x} = (\mathbf{s}, \mathbf{f})$，经过语义通信链路后，接收端得到重构结果$(\hat{\mathbf{s}}, \hat{\mathbf{f}})$。由于文本语义一致性损失、关键帧重构损失及其加权联合形式已在式(6)–式(8)中定义，本文在CSA部分不再重复展开，而将其统一记为联合重构损失$\mathcal{L}_{\text{rec}}(\mathbf{x}; \boldsymbol{\theta}_t)$，其中$\boldsymbol{\theta}_t$表示当前任务下语义编解码器及信道自编码器的可训练参数。

为抑制新任务学习过程中的知识漂移，CSA引入邻近知识保留约束。设当前模型和上一任务模型对同一样本$\mathbf{x}$提取的联合语义表示分别为：

$$\begin{aligned} \mathbf{z}_t(\mathbf{x}) &= \left[E_s(\mathbf{s}; \boldsymbol{\theta}_t), E_f(\mathbf{f}; \boldsymbol{\theta}_t)\right], \\ \mathbf{z}_{t-1}(\mathbf{x}) &= \left[E_s(\mathbf{s}; \boldsymbol{\theta}_{t-1}), E_f(\mathbf{f}; \boldsymbol{\theta}_{t-1})\right] \end{aligned} \tag{21}$$

其中，$E_s(\cdot)$和$E_f(\cdot)$分别表示文本和视觉语义编码器。邻近知识保留损失定义为：

$$\mathcal{L}_{\text{pr}}(\mathbf{x}; \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) = \left\| \mathbf{z}_t(\mathbf{x}) - \mathbf{z}_{t-1}(\mathbf{x}) \right\|_2^2 \tag{22}$$

该损失约束当前模型在潜在语义空间中不要过度偏离上一任务模型，从而降低新任务更新对已有语义知识的破坏。CSA根据每个样本的学习状态自适应调整训练权重。对于小批量样本$\mathcal{B}$，分别以联合重构损失和邻近知识保留损失衡量样本的新任务学习误差与旧知识漂移程度：

$$\begin{aligned} A(\mathbf{x}) &= \mathcal{L}_{\text{rec}}(\mathbf{x}; \boldsymbol{\theta}_t), \\ B(\mathbf{x}) &= \mathcal{L}_{\text{pr}}(\mathbf{x}; \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) \end{aligned} \tag{23}$$

根据小批量样本中$A(\mathbf{x})$和$B(\mathbf{x})$的均值与标准差，构建判别阈值：

$$\begin{aligned} F_A &= \mu_A - k\sigma_A, \quad G_A = \mu_A + k\sigma_A, \\ F_B &= \mu_B - k\sigma_B, \quad G_B = \mu_B + k\sigma_B \end{aligned} \tag{24}$$

其中，μ_A，σ_A和μ_B，σ_B分别表示$A(\mathbf{x})$与$B(\mathbf{x})$在当前小批量样本中的均值和标准差，k为阈值调节系数。为实现持续学习，CSA维护了一个历史记忆缓冲区：

$$\mathcal{M}_{t-1} = \bigcup_{j=1}^{t-1} \mathcal{M}_j, \quad \mathcal{M}_j \subset \mathcal{D}_j \tag{25}$$

其中，$\mathcal{M}_{t-1}$存储了先前任务的代表性样本（高性能样本），$\mathcal{M}_j$表示从第j个任务中选择的代表性样本子集，历史记忆缓冲区采样先入先出的策略。在任务$\mathcal{T}_t$的每次迭代中，从当前任务数据中随机抽取小批量$\mathcal{B}_t \subset \mathcal{D}_t$，并从记忆缓冲区随机重放小批量$\mathcal{B}_m \subset \mathcal{M}_{t-1}$。整体训练目标可表述如下：

$$\min_{\theta_t} \mathbb{E}_{x\sim\mathcal{B}_t}\left[w_{\text{lea}}(x)\mathcal{L}_{\text{rec}}(x;\theta_t)\right] +\lambda\mathbb{E}_{x\sim\mathcal{B}_m}\left[w_{\text{kep}}(x)\mathcal{L}_{\text{pr}}(x;\theta_t,\theta_{t-1})\right] \tag{26}$$

其中，$\lambda>0$为保留系数，用于调节新任务适应与旧知识保留之间的整体权衡关系。$w_{\text{lea}}(x)$和$w_{\text{kep}}(x)$分别表示学习权重与知识保留权重的调节系数。较大的$w_{\text{lea}}(x)$用于增强模型对当前任务分布的学习能力，而较大的$w_{\text{kep}}(x)$用于加强对已有语义知识的保持约束。

基于对$A(\mathbf{x})$与$B(\mathbf{x})$的联合判断,每个样本被归类为以下五种类型之一[16]：

（1）高性能样本：重建误差低$\left(\mathcal{L}_{\text{rec}}\leqslant F_A\right)$且漂移量小$\left(\mathcal{L}_{\text{pr}}\leqslant F_B\right)$，表明样本学习效果良好且稳定性强；需同时加强学习权重w_{lea}和知识保留权重w_{kep}。**（2）高方差样本：**重建误差低$\left(\mathcal{L}_{\text{rec}}\leqslant F_A\right)$但漂移量大$\left(\mathcal{L}_{\text{pr}}\geqslant G_B\right)$，表明样本学习效果良好但稳定性不足；需抑制学习权重w_{lea}并增强知识保留权重w_{kep}。**（3）过拟合样本：**重建误差高$\left(\mathcal{L}_{\text{rec}}\geqslant G_A\right)$但漂移量小$\left(\mathcal{L}_{\text{pr}}\leqslant F_B\right)$，表明样本受旧知识约束较强且新任务适应不足；需强化学习权重w_{lea}并适度减小知识保留权重w_{kep}。**（4）噪声样本：**重建误差高$\left(\mathcal{L}_{\text{rec}}\geqslant G_A\right)$且漂移量大$\left(\mathcal{L}_{\text{pr}}\geqslant G_B\right)$，表明样本可能造成振荡或有害更新；需同时减小学习权重w_{lea}和知识保留权重w_{kep}。**（5）正常样本：**不满足上述判别条件，表明样本误差与漂移处于正常范围；采用默认学习权重w_{lea}和知识保留权重w_{kep}。该目标函数通过语义一致性监督促进新任务学习，同时通过近端知识约束抑制先前习得知识的漂移。由此可有效缓解灾难性遗忘现象，并在语义/信道编解码器的持续优化过程中提升训练稳定性。

3　实验结果与分析

3.1　实验设置

为全面评估系统性能，论文在多个公开视频数据集上开展实验，包括 Kaggle 平台上的 UCF101[17]、Kinetics-700[18]以及 Charades [19]数据集。为验证所提出的视频语义通信系统的有效性，论文选取三个代表性基线方案进行对比：（1）基于深度学习的视频语义传输方案 DVST [20]，（2）基于生成模型的视频语义传输方案GVSC[21]，（3）传统编码传输流程H.264+ LDPC [22]。

基础模型方面，VLM选用VideoGPT+模型[23]，包含51亿参数量。VLM模型输出**s**的最大长度为128，**f**为2帧336*336的视频关键帧。WM使用WAN2.1模型[24]，包含140亿参数量。网络架构方面，文本语义编码器由三个Transformer编码器层构成，每层配备8个注意力头且特征维度为128；视觉语义编码器采用ViT架构，每个Transformer编码器层含4个注意力头，单头特征维度为16，最终每层总特征维度达64。信道编码器采用全连接隐藏层实现，首层含256个神经元，次层含128个神经元。为保持语义一致性，语义解码器与信道解码器采用与发送端相反的设计。

本文采用两阶段训练，VLM和WM始终冻结。第一阶段依次以互信息预训练信道编解码器、以式（8）训练双语义编解码器，并联合微调整条语义通信链路；第二阶段基于历史记忆缓冲区和式（26）开展CSA持续学习。所有实验均在统一软硬件环境下进行，硬件平台搭载英特尔(R)至强(R)Silver 4210R CPU（2.40GHz主频）及 NVIDIA A800 GPU（80G显存）。模型训练与推理通过PyTorch 1.8.0实现，并搭载 CUDA 11.6加速功能。

3.2　基础模型性能对比

为深入评估基础模型（VLM和WM）在视频语义提取与重构中的核心能力，在提出的视频语义通信系统中将三种不同VLM作为发送端语义提取器，包括 VideoChat 2[25]、VideoLLaMA 2[26]以及VideoGPT+模型，两种不同的WM作为接收的视频重构器，包括 CogVideoX 1.5[27]和WAN 2.1[24]。实验在 AWGN 信道环境下进行，数据集为UCF101，主要评估指标采用归一化的CLIPScore[28]。图4展示了六种方案在不同信噪比（SNR）条件下的CLIPScore变化趋势。结果表明，随着信道质量的提升，所有模型的语义一致性得分均呈上升趋势，但VideoGPT+和 WAN 2.1的组合在整个测试信噪比范围内始终保持最优性能。表明其更适合作为高保真视频语义通信系统的语义提取和重构模型。

然后，本文进一步对VLM，WM和CSA机制进行了消融实验。为保证系统链路的完整性，Ours w/o Prompt表示不用提示词对VLM的语义提取进行约束；Ours w/o VLM表示用S2VT[29]取代VLM模型；Ours w/o WM表示用SimVP[30]取代

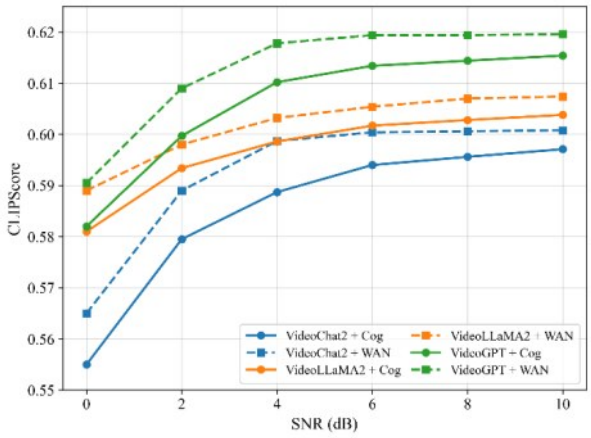

图4　基于不同 VLM 和 WM 的视频语义通信系统性能对比

WM 模型；Ours w/o CSA 表示不进行持续学习。评价指标为归一化 CLIPScore 和峰值信噪比（PSNR）。消融实验结果如表 1 所示。VLM 对实验结果的影响最大，表明了准确的语义提取是语义通信的基石，WM 的视觉预测能力也对实验结果有较大的影响，当系统不进行持续学习时，系统的性能也会下降。最后，提示词能够帮助 VLM 模型聚焦关键的语义信息，对系统的整体性能也有一定的影响。

表 1　　消融实验结果

方法	CLIPScore	PSNR
Ours	**0.62**	**35.37**
Ours w/o Prompt	0.60	33.25
Ours w/o VLM	0.34	17.36
Ours w/o WM	0.36	18.43
Ours w/o CSA	0.42	22.14

3.3　动态环境下 CSA 稳定性评估

为验证所提出的 CSA 模块在动态环境下的有效性，在 AWGN 通道下对配备 CSA 的视频语义通信系统与未配备 CSA 的基线系统进行了对比评估。模型依次在三个数据分布差异显著的视频数据集（UCF101、Kinetics-700 和 Charades[19]）上进行训练，以 BLEU 分数作为主要评估指标[31]。如图 5 和图 6 所示，两种配置在动态环境下展现出截然不同的性能轨迹。未配备 CSA 的基线系统在早期任务（UCF101）上的 BLEU 分数随后续任务的变化急剧下降，表明系统在适应新的任务分布时存在严重灾难性遗忘。相比之下，配备 CSA 的视频语义通信系统在执行后续任务（Kinetics 和 Charades）时，原有数据集（UCF101）上的性能仅出现微小下降，展现出显著提升的稳定性。

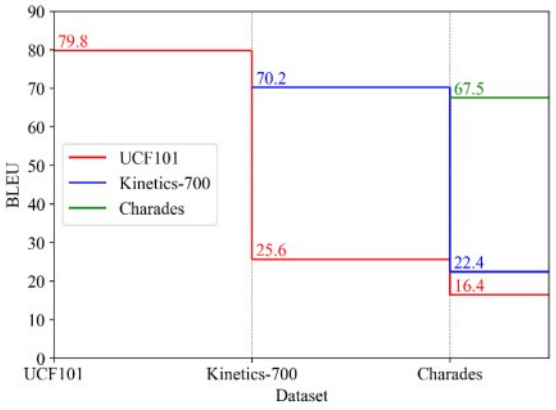

图5　动态环境下未采用 CSA 视频语义通信系统的 BLEU 性能

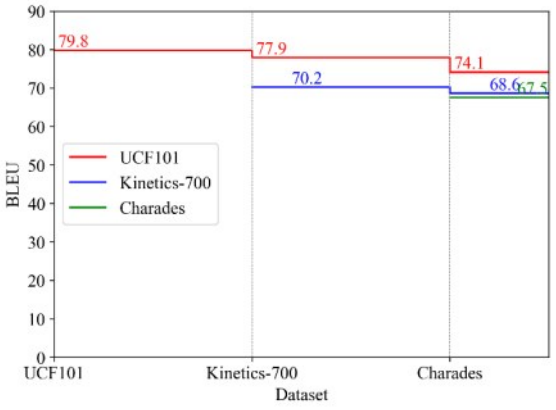

图6　动态环境下采用 CSA 视频语义通信系统的 BLEU 性能。

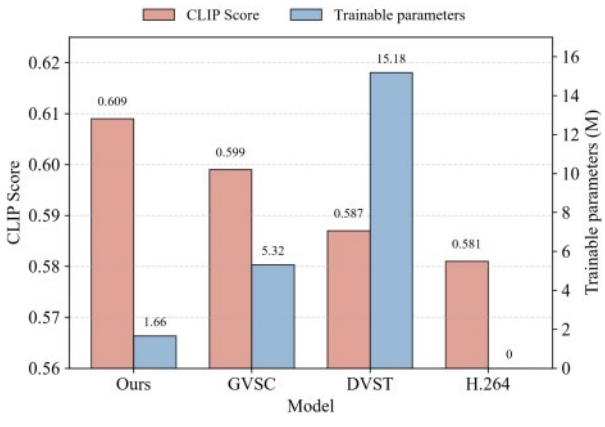

图7　不同视频通信系统之间可训练参数及语义一致性的比较

3.4　模型质量与复杂度评估

为全面评估系统性能，本文对比了 2 dB 低信噪比下不同通信系统的语义一致性与可训练参数

量，如图 7 所示。归一化 CLIPScore 用于衡量重构视频与发送端语义描述的匹配程度。所提系统通过仅传输关键帧和文本，并冻结基础模型，在学习型方案中取得了最高 CLIPScore 和最少的可训练参数。DVST 含 15.18 M 可训练参数，给资源受限平台带来较大计算压力；H.264+LDPC 虽无需训练，但 CLIPScore 最低。

此外，本文比较了不同系统的 PSNR、SSIM 和主观 QoE[32]。由表 2 可知，所提系统在三项指标上均获得最高得分，表明视频语义提取、重构及双语义压缩方案在低信噪比条件下具有更强的鲁棒性。

表 2　不同视频通信系统的主客观评价得分

方法	PSNR	SSIM	Subjective QoE
Ours	**34.64**	**0.81**	**53.22**
GVSC	32.48	0.79	51.64
DVST	29.32	0.72	47.45
H.264+LDPC	12.11	0.13	16.37

3.5　不同信道条件下视频生成质量评估

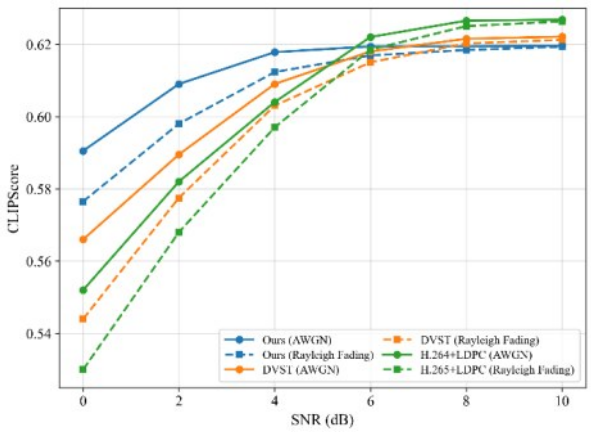

图 8　AWGN 与瑞利衰落信道下视频通信系统的性能对比

图 8 展示了不同信道条件下三种视频通信方案的 CLIPScore 性能。结果表明，所提系统在低信噪比条件下取得最优性能，有效缓解了传统通信系统的悬崖效应。该优势主要得益于“文本提示+关键帧”的生成式重构范式：世界模型利用已学习的时空演化规律和语义先验，对受损或缺失的视频信息进行预测与补全，从而降低信道噪声对像素细节的影响，并保持重构视频与文本语义的高度一致。在高信噪比条件下，H.264+LDPC 表现出一定优势，这是由于它们传输和重建的是像素语义，而本文系统优先保证主体、动作和场景等高层语义的一致性，并不以严格恢复原始视频的像素和纹理细节为目标，因此本文系统更适用于带宽受限或信道条件较差、但需要保持关键语义内容的通信场景。

4　结论

本文面向视频语义通信场景，提出了一种基础模型赋能的视频语义通信系统。该系统以 VLM 和 WM 分别作为语义提取与视频重构模块，提升了低信噪比条件下的语义保真度。此外，该系统引入 CSA 模块，通过自适应加权抑制灾难性遗忘，增强模型在多任务动态环境下的泛化能力。实验结果表明，本文所提出的视频语义通信系统在语义一致性和视频重构质量方面均优于传统方法，验证了其在鲁棒高效视频传输中的应用潜力。未来可进一步面向极端非平稳通信环境，研究更加稳健的 CSA 阈值自适应机制与高效编解码方法，以提升系统在复杂动态场景下的适应性与传输效率。

参考文献：

[1] 张平, 许晓东, 牛凯等. 现代语义通信与 6G 智简网络理论技术体系[J]. 北京邮电大学学报, 2025, 48(5): 1–16.
Zhang P, Xu X D, Niu K, et al. Theoretical and technical system of modern semantic communications and 6G intelligent-simple networks[J]. Journal of Beijing University of Posts and Telecommunications, 2025, 48(5): 1-16.

[2] Nan G, Li Z, Zhai J, et al. Physical-layer adversarial robustness for deep learning-based semantic communications [J]. IEEE Journal on Selected Areas in Communications, 2023, 41(8): 2592–2608.

[3] Zhang P, Xu W, Liu Y, et al. Intellicise wireless networks from semantic communications: A survey, research issues, and challenges [J]. IEEE Communications Surveys & Tutorials, 2024, 27(3): 2051–2084.

[4] 张平, 牛凯, 姚圣时等. 面向未来的语义通信：基本原理与实现方法[J]. 通信学报, 2023, 44(5): 1–14.
Zhang P, Niu K, Yao S S, et al. Semantic communications for future: basic principle and implementation methodology[J]. Journal on Communications, 2023, 44(5): 1-14.

[5] Cui Q, You X, Wei N, et al. Overview of AI and communication for 6G network: Fundamentals, challenges, and future research opportunities [J]. Science China Information Sciences, 2025, 68(7): 171301.

[6] Jiang F, Tu S, Dong L, et al. FlashSAM: Lightweight Vision Model for Multi-UAV Token Communication in Low-Altitude Wireless Networks [J]. IEEE Journal of Selected Topics in Signal Processing, 2026.

[7] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks [J]. Proceedings of the national academy of sciences, 2017, 114(13): 3521–3526.

[8] Jiang F, Tang C, Dong L, et al. Visual language model based cross-

modal semantic communication systems [J]. IEEE Transactions on Wireless Communications, 2025, 24(5): 3937–3948.

[9] Zhao Y, Yue Y, Hou S, et al. LaMoSC: Large language model-driven semantic communication system for visual transmission [J]. IEEE Transactions on Cognitive Communications and Networking, 2024, 10(6): 2005–2018.

[10] Liang S D. Vision language models for massive MIMO semantic communication[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2025: 1675-1685.

[11] Jiang P, Guo J, Wen C-K, et al. Semantic Communications with World Models [J]. `arXiv preprint arXiv:2510.24785, 2025.

[12] Lokumarambage M, Sivalingam T, Dong F, et al. Latent Prediction based Generative Semantic Communication for Video Transmission in Wireless Networks [J]. IEEE Open Journal of the Communications Society, 2026, 7: 3974–3986.

[13] Jiang F, Tu S, Dong L, et al. Tokencom: Vision-language model for multimodal and multitask token communications [J]. arXiv preprint arXiv:260300482, 2026.

[14] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 813-824.

[15] Peebles W, Xie S. Scalable diffusion models with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4195-4205.

[16] Zhang H, Lei Y, Gui L, et al. CPPO: continual learning for reinforcement learning with human feedback[C]//Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna: OpenReview, 2024.

[17] Soomro K, Zamir A R, Shah M. Ucf101: A dataset of 101 human actions classes from videos in the wild [J]. arXiv preprint arXiv: 1212.0402, 2012.

[18] Carreira J, Noland E, Hillier C, et al. A short note on the kinetics-700 human action dataset [J]. arXiv preprint arXiv:1907.06987, 2019.

[19] Sigurdsson G A, Varol G, Wang X, et al. Hollywood in homes: Crowdsourcing data collection for activity understanding[C]//European Conference on Computer Vision. 2016: 510-526.

[20] Wang S, Dai J, Liang Z, et al. Wireless deep video semantic transmission[J]. IEEE Journal on Selected Areas in Communications, 2023, 41 (1): 214-229.

[21] Yin H, Qiao L, Ma Y, et al. Generative video semantic communication via multimodal semantic fusion with large model [J]. IEEE Transactions on Vehicular Technology, 2025.

[22] Wang Y, Yu S, Yang X. Error robustness scheme for H.264 based on LDPC code[C]//2006 12th International Multi-Media Modelling Conference. 2006: 79.

[23] Maaz M, Rasheed H, Khan S, et al. Videogpt+: Integrating image and video encoders for enhanced video understanding [J]. arXiv preprint arXiv:2406.09418, 2024.

[24] Wan T, Wang A, Ai B, et al. Wan: Open and advanced large-scale video generative models [J]. arXiv preprint arXiv:2503.20314, 2025.

[25] Li K, Wang Y, He Y, et al. MVBench: A comprehensive multi-modal video understanding benchmark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 22195-22206.

[26] Cheng Z, Leng S, Zhang H, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms [J]. arXiv preprint arXiv:2406.07476, 2024.

[27] Yang Z, Teng J, Zheng W, et al. CogVideoX: Text-to-video diffusion models with an expert transformer[C]//International Conference on Learning Representations. 2025.

[28] Hessel J, Holtzman A, Forbes M, et al. CLIPScore: A reference-free evaluation metric for image captioning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 7514-7528.

[29] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence-video to text[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 4534-4542.

[30] Gao Z, Tan C, Wu L, et al. SimVP: simpler yet better video prediction [C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 3170-3180.

[31] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[32] Jiang F, Tu S, Dong L, et al. Large generative model-assisted talking-face semantic communication system [J]. IEEE Journal on Selected Areas in Communications, 2025.