VideoTokenCom：基于文本意图引导的多速率视频令牌通信与基于 UEP 的自适应信源信道编码

Jingxuan Men，*IEEE* 学生会员，Mahdi Boloursaz Mashhadi，*IEEE* 高级会员，Ning Wang，*IEEE* 高级会员，Yi Ma，*IEEE* 高级会员，Mike Nilsson 和 Rahim Tafazolli，*IEEE* 会士

arXiv:2603.02470v1 [cs.IT] 2 Mar 2026

摘要—**令牌通信（TokenCom）是一种新兴范式，受大型人工智能模型（LAM）和多模态大语言模型（MLLM）近期取得的成功所推动。在这一范式中，令牌作为统一的通信与计算单元，使未来无线网络能够高效开展面向语义和目标的信息交换。本文提出一种新颖的视频 TokenCom 框架，利用基于不等差错保护（UEP）的信源信道编码自适应，实现文本意图引导的多速率视频通信。该框架将用户指定的文本描述与离散视频令牌化和不等差错保护相结合，以在带宽限制严格的条件下提升语义保真度。首先，通过预训练的视频令牌化器提取离散视频令牌，同时结合文本条件视觉语言建模与光流传播，识别视频在时空维度上与用户意图语义相对应的令牌。接下来，我们提出一种语义感知多速率比特分配策略：与用户意图高度相关的令牌采用完整码本精度进行编码，而非意图令牌则通过缩减码本精度差分编码表示，从而在节省速率的同时保持语义质量。最后，我们设计一种信源与信道编码自适应方案，根据不断变化的资源和链路状况调整比特分配与信道编码。在多个视频数据集上的实验表明，在广泛的信噪比范围内，该框架在感知质量和语义质量方面均优于传统基线方法和语义通信基线。**

索引术语—**令牌通信、视频语义通信、多模态大语言模型（MLLMs）、自适应编码与调制、不等差错保护（UEP）。**

I. 引言

TOKENS 已成为现代生成式基础模型中用于跨模态感知、推理和内容生成的表征与处理基本单元 [1], [2]。离散令牌表示能够将高维信号压缩为紧凑且具有语义意义的序列，视觉和多模态学习中的向量量化及离散潜变量建模方法已证明了这一点 [1], [2]。随着多模态大语言模型（MLLM）的快速发展，

J. Men、M. Boloursaz Mashhadi、Y. Ma 和 R. Tafazolli 就职于 5GIC & 6GIC、通信系统研究所（ICS）、萨里大学，英国（电子邮箱：j.men, m.boloursazmashhadi, y.ma, r.tafazolli@surrey.ac.uk）。N. Wang 就职于智能互联网实验室、布里斯托大学，英国（电子邮箱：n. wang@bristol.ac.uk）。M. Nilsson 就职于英国电信研究实验室、英国电信集团，英国（电子邮箱：mike.nilsson@bt.com）。

学习式标记化机制进一步支持文本与视觉之间的语义对齐，同时保留感知细节和高层语义概念 [3]–[5]。这些特征表明，基于令牌的表示与无线通信系统之间存在天然的接口，离散且语义结构化的信息单元可以直接映射至资源受限的传输流水线。受基础模型与无线通信融合趋势的启发，基于令牌的通信已成为一种新的设计范式，它摒弃传统的比特级传输，转而采用具有语义意义的令牌表示。

TokenCom框架最初于 [1], 提出，旨在弥补传统语义通信和目标导向通信中泛化能力方面的差距 [6]–[8]。TokenCom建立了一种上下文感知架构，利用基于Transformer的基础模型直接在标记域中提取、变换和传输多模态信息。后续研究从多个维度推动了以令牌为中心的无线技术发展，包括标记域多址接入 [2], [9],、通过遗传搜索或前瞻搜索实现的语义分组聚合 [10], [11],、面向目标导向通信或基于Transformer的边缘推理的自适应标记选择 [12], [13], 、文本引导的基于令牌的图像通信 [14],、采用选择性训练标记的多模态语义通信 [15],，以及抗损失掩码视觉标记建模 [16]。研究还从信息瓶颈视角探讨了TokenCom [17] ，并将其扩展至由多模态大模型赋能的资源受限多用户网络 [18]。最近，研究人员针对TokenCom探讨了联合语义信道编码与自适应调制策略，并在标记语义与物理层传输之间进行了跨层设计 [19]。最后，TokenCom-UEP [20], 通过为语义重要性不同的令牌分配差异化的信道保护，展现出更强的鲁棒性。尽管取得了这些进展，视频TokenCom仍鲜有研究。

大多数现有的视频语义通信系统基于连续特征表示或任务特定潜在编码运行，并未显式地将离散视频令牌作为压缩与通信的统一单元加以利用。尽管离散视频标记化已成为高维视频数据的一种有效表示，现有框架并未利用标记级语义结构来实现高效通信，例如对视频源的最优分配

以及信道编码率。本文针对上述不足，提出一种融合视频令牌化、文本条件意图相关性提取、多速率比特分配策略和基于语义UEP的源信道自适应的多速率文本意图引导信源信道编码自适应视频令牌通信框架。与许多现有的端到端学习或基于深度联合源信道编码（DJSCC）的语义通信方案 [21], [22], 不同，所提出的框架借助预训练离散标记化，基于源信道分离进行设计，从而在不同信道/网络状况下具备灵活性，并采用与ITU-T OSI兼容的多层设计以实现可扩展性。我们首先将视频分解为帧内编码帧（I帧）和预测编码帧（P帧），然后采用预训练视频令牌化器将视频帧映射为离散令牌。随后，我们利用视觉语言模型在空间和时间维度上识别视频中的用户意图指定语义区域。将这些语义掩码映射到令牌网格上，得到与文本意图显式关联的用户意图指定令牌集合和非意图令牌集合。接下来，我们设计一种语义感知多速率比特分配方案：用户意图区域内的令牌以完整比特精度传输，而非意图令牌则相对于参考帧以降低的精度进行差分编码，从而在保持意图相关内容质量的同时节省比特率。最后，我们在固定资源预算下构建源信道编码自适应问题，分别为意图令牌类别和非意图令牌类别分配源比特精度和调制编码配置，以在UEP框架内平衡任务失真与传输时延。本文的主要贡献由此总结如下：

- 我们基于视觉语言建模和视频帧上的光流轨迹，构建了一个意图相关性提取框架。该框架利用用户的文本意图描述，将视频内容转换为与用户意图指定和非意图指定语义显式关联的离散令牌类别。

- 为实现语义感知信源编码，我们提出了一种多速率比特分配策略，根据不同视频令牌类别（即意图指定类与非意图指定类）的语义相关性，为其分配不同的信源编码比特精度。意图指定类采用完整的令牌码本精度比特进行编码，而非意图指定类则通过紧凑的精简码本，以降低后的精度进行差分编码。这种设计在保留意图相关语义区域视觉质量的同时，显著提升了速率效率。仿真结果表明，所提出的 TokenCom 框架在每像素比特数（BPP）极低时，性能优于传统 H.265 [23]、[24] 和基于生成扩散的 VC-DM [25] 视频编码基线方法。

- 我们提出了一种源信道优化方案，在有限资源约束下显式权衡语义失真与端到端传输时延。遵循 UEP 方法，所提出的方案针对目标令牌类别和非目标令牌类别分别调整调制编码方案（MCS），同一类别中的所有令牌采用相同的 MCS。该优化方案为每个令牌类别选择一种配置，在满足固定资源预算和可靠性约束的前提下，最小化失真与时延的加权和。与传统 H.265 视频压缩及传输相比，在各种信噪比（SNR）下，Video TokenCom 在峰值信噪比（PSNR）、结构相似性指数（SSIM）、学习感知图像块相似度（LPIPS）、Fréchet 视频距离（FVD）和基于 CLIP 的语义相似度（CLIP）方面均取得了稳定提升；在信噪比为 6 dB 时，FVD 降低了近 1500。

本文其余部分安排如下：第二节介绍所提出的文本意图引导的多速率视频令牌通信框架和基于UEP的自适应信源信道编码系统模型，并给出问题表述。第三节和第四节分别介绍实验评估和消融研究。最后，第五节给出结论。

II. 系统模型

图1展示了整体系统架构，该架构由三个主要部分组成：(i) 基于令牌的文本意图引导信源编码器，包含多模态用户意图令牌提取和语义感知多速率比特编码；(ii) 基于令牌的信源解码器；(iii) 基于UEP的信源信道编解码自适应。所提出的架构具有灵活性，允许不同实体发布文本意图，无论是内容生产者还是内容消费者。例如，内容提供商可以在各个视频片段中嵌入文本意图，以便在不同信道条件下进行流传输时，引导自适应信源信道编码。另一方面，允许内容消费者在内容消费过程中表达文本意图，可以更灵活地支持不同的观看重点，但会产生少量意图反馈开销。

A. 视频令牌化

我们部署视频令牌化器，将视频转换为离散令牌表示。给定输入视频$\mathbf{V} \in \mathbb{R}^{T\times H\times W}$，其帧数（时间长度）为 T，空间分辨率为 $H \times W$，我们使用预训练视频令牌化器 $\mathcal{E}(\cdot)$以及预先学习的离散令牌码本$\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N\} \in \mathbb{R}^{N\times D}$，其中 N为码本大小，即离散令牌的总数，D 为每个令牌向量的嵌入维度。每个条目$\mathbf{c}_i \in \mathbb{R}^D$ 对应于一个学习得到的原型，用于表示一个时空视频块。

在时间下采样因子 d_t 和空间下采样因子 d_s下，令牌化器将视频编码为由离散令牌构成的网格$\mathbf{Z} = \mathcal{E}(\mathbf{V}) \in \mathbb{R}^{t\times h\times w}$，其中$t = \frac{T}{d_t}$，$h = \frac{H}{d_s}$，$w = \frac{W}{d_s}$。此处，$\mathbf{Z}$是取值范围 $\{1, \ldots, N\}$内的整数索引向量。通过传输离散令牌而非原始像素，由于采用低比特整数索引，数据量显著减少，同时借助学习得到的码本嵌入保留高层语义。

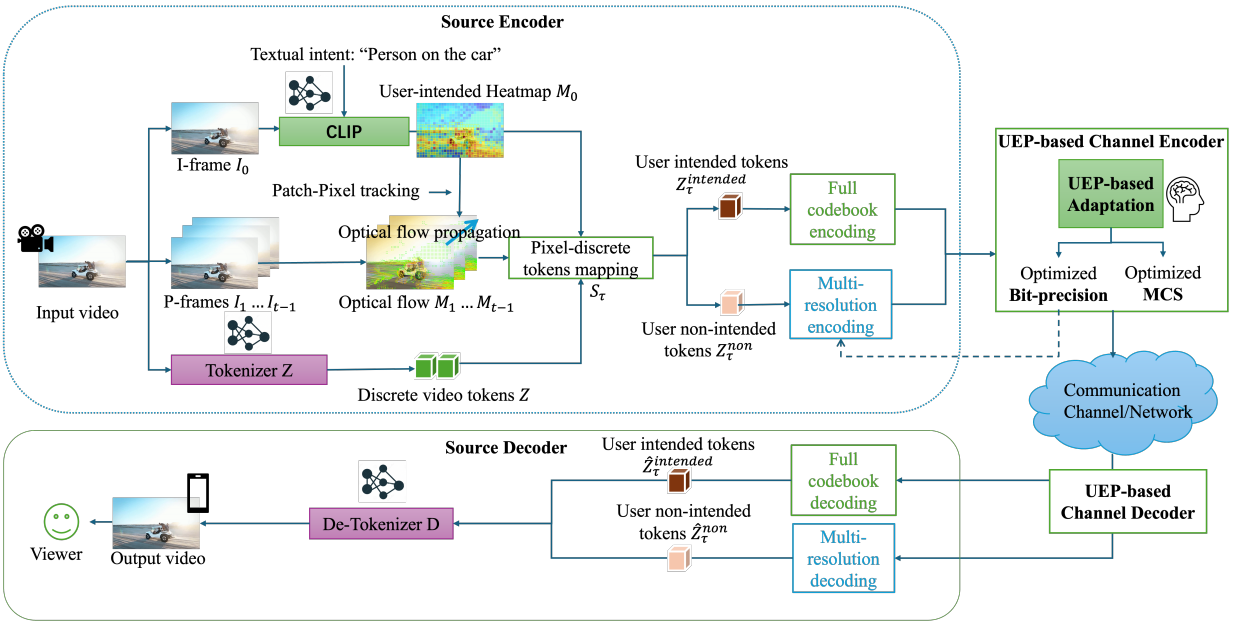


图1. 所提出的具备多速率文本意图信源信道编码自适应的视频令牌通信框架总体架构。该流程包括：(1) 所提出的基于令牌的文本意图引导信源编码器，包含多模态用户意图令牌提取和语义感知多速率比特编码；(2) 所提出的基于令牌的信源解码器；以及(3) 所提出的基于UEP的信源信道编解码自适应。

B. 多模态用户意图令牌提取

用户意图语义通信 [26], [27] 已成为一个有影响力的研究方向，其驱动力在于需要使传输优先级与用户感知、注意力或任务特定目标相一致，而不是统一对待所有内容。在本工作中，我们在多模态场景中实例化这一原则，根据文本意图从视频内容中提取用户意图令牌，从而在令牌级实现差异化处理与传输。

1) 文本条件热图生成：为根据文本意图在输入视频中查找相关内容，对于序列中的首帧$\mathbf{I}_0 \in \mathbb{R}^{H\times W}$，我们计算由视觉语言模型生成的、针对文本意图描述$\mathbf{p}$的文本条件热图。视觉语言模型的图像块大小为 p（例如，ViT-B/32的图像块大小为 p=32 ）。在网格上生成的文本条件热图为

$$\hat{\mathbf{H}} \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}}, \qquad \hat{\mathbf{H}}_{i,j} = \cos\big(f_{\text{img}}(\mathbf{I}_0^{(i,j)}),\, f_{\text{text}}(\mathbf{p})\big), \tag{1}$$

其中，$\mathbf{I}_0^{(i,j)}$ 表示网格索引 $(i,\ j)$ 处的 $p \times p$ 图像块，f_{img}, f_{text} 分别表示视觉语言模型的图像编码器和文本编码器，cos(·, ·)表示余弦相似度。然后，我们对热图$\mathbf{H}_{i,\ j}$进行归一化。

$$\tilde{\mathbf{H}}_{i,j} = \frac{\hat{\mathbf{H}}_{i,j} - \min(\hat{\mathbf{H}})}{\max(\hat{\mathbf{H}}) - \min(\hat{\mathbf{H}}) + \varepsilon} \in [0,1], \tag{2}$$

其中，ε 是一个较小的常数，用于避免除以零。

为识别与文本相关的区域，我们对归一化热力图应用相似度阈值 ℓ 。具体而言，如果一个图块的 CLIP 相似度超过 ℓ，则将其选为用户意图指定区域，从而得到二值图块级掩码

$$\mathbf{M}_0^{\text{patch}}(i,j) = \mathbb{1}\big[\tilde{\mathbf{H}}_{i,j} > \ell\big], \tag{3}$$

其中，$\mathbb{1}[\cdot]$ 是示性函数。这里，$(h_p,\ w_p)$ $=(H/p,\ W/p)$ 表示图块级热力图的空间分辨率。所选图块的比例记为 ρ，其计算方式为

$$\rho = \frac{1}{h_p w_p}\sum_{i=1}^{h_p}\sum_{j=1}^{w_p}\mathbf{M}_0^{\text{patch}}(i,j), \tag{4}$$

其中，图块级掩码可选择通过形态学膨胀进行细化，以扩大高响应区域。最后，将图块级掩码上采样至像素分辨率

$$\mathbf{M}_0 = \mathcal{U}\big(\mathbf{M}_0^{\text{patch}};\, H, W\big) \in \{0,1\}^{H\times W}, \tag{5}$$

其中，$\mathcal{U}(\cdot)$表示最近邻上采样，它会将每个图块标签复制到对应的 $p \times p$ 像素块。

2) 动态光流传播：我们通过设置首帧的 $\{\mathbf{M}_k\}_{k=0}^{T-1}$，初始化像素级二值语义掩码序列。令$\mathbf{F}_{k\to k+1} \in \mathbb{R}^{H\times W\times 2}$表示从帧 k 到 k+1的前向光流 [28], [29]，其中每个向量$\mathbf{F}_{k\to k+1}(u,v)$ $=(\Delta x, \Delta y)$表示像素位移，(x, y)枚举像素坐标，而$(\Delta x, \Delta y)$取自$\mathbf{F}_{k\to k+1}(y, x)$。我们通过双线性采样进行变形来传播掩码 $\mathcal{W}$

$$\mathbf{M}_{k+1} = \mathcal{W}\big(\mathbf{M}_k, \mathbf{F}_{k\to k+1}\big), \quad (x', y') = (x + \Delta x,\, y + \Delta y), \tag{6}$$

其中，k索引原始视频帧（$k \in \{0, \dots, T-1\}$），$\mathcal{W}$ 是一个双线性采样器；对于每个像素，光流位移（Δx，Δy）将其映射到亚像素坐标（x'，y'），并通过对$\mathbf{M}_k$的四个相邻像素进行双线性插值来获取其值。

3) 内容离散令牌映射：我们将所有视频令牌划分为两个互不相交的令牌类别 $c \in \{s, n\}$，分别对应用户意图令牌（s）和非意图令牌（n），

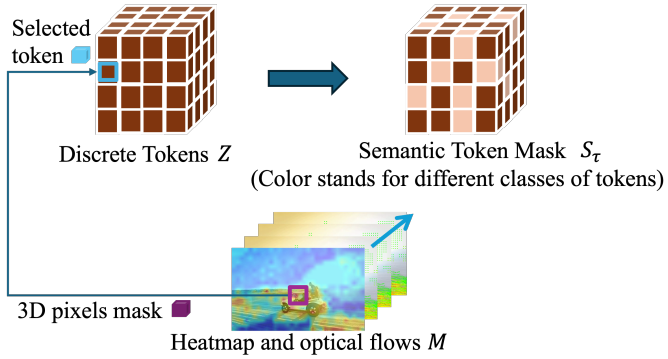


Fig. 2. 文本意图引导的多类别多速率令牌映射

。如图 2 所示，该分类在令牌级进行，方法是将文本相关内容像素映射到离散时空令牌网格。为了将文本相关内容像素映射到离散视频令牌，我们首先注意到，给定时间下采样因子 d_t，令牌的时间长度为 $t = T/d_t$ 。令 $\tau \in \{0,\dots,t-1\}$ 为时间令牌的索引。帧 $\tau\cdot d_t+f$ 上传播后的像素级内容掩码记为 $\mathbf{M}_{\tau\cdot d_t+f} \in \{0,1\}^{H\times W}$，其中 $f \in \{0,\dots,d_t-1\}$ 表示与第 τ 个令牌对应的时间窗口内原始视频帧的索引。时空池化将传播后的像素级掩码与三维令牌网格对齐

$$\bar{\mathbf{M}}_\tau = \frac{1}{d_t}\sum_{f=0}^{d_t-1} \mathrm{AdaptiveAvgPool2D}\big(\mathbf{M}_{\tau\cdot d_t+f},(h,w)\big), \quad (7)$$

其中，$\bar{\mathbf{M}}_\tau(i,j) \in [0,1]$ 表示在长度为 d_t 帧的时间窗口内，映射到令牌 (i,j) 的文本相关像素的平均比例。通过阈值处理可获得令牌级二值语义掩码

$$\mathbf{S}_\tau = \mathbb{1}\big[\bar{\mathbf{M}}_\tau > \theta\big], \quad (8)$$

其中，$\mathbf{S}_\tau \in \{0,1\}^{h\times w}$ 表示二值语义令牌掩码，$\theta \in (0,1]$ 控制将令牌分类为用户意图指定所需的最小文本相关像素比例。如果由$\mathbf{M}_\tau(i,j)$所表示的时空令牌内文本相关像素的平均比例超过阈值 θ，则将该令牌标记为用户意图指定。

所提出的多模态用户意图指定令牌提取方法的完整说明见**算法1**。

算法1： 令牌级语义提取与掩码生成

输入：视频帧 $\{\mathbf{I}_k\}_{k=0}^{T-1}$，文本意图 $\mathbf{p}$。
输出： 令牌级语义掩码 $\mathbf{S} = \{\mathbf{S}_\tau\}$；意向集合 $\mathcal{S}$ 与非意向集合 $\mathcal{N}$。

在首帧上计算视觉语言热图$\mathbf{I}_0$并获得初始像素级语义掩码$\mathbf{M}_0^{(\mathrm{px})}$。
对于 $k=0$ **到** $T-2$ **执行**
  使用光流传播像素级语义区域：

$$\mathbf{M}_{k+1}^{(\mathrm{px})} = \mathrm{Warp}\Big(\mathbf{M}_k^{(\mathrm{px})}, \mathbf{F}_{k\to k+1}\Big).$$

将 $\{\mathbf{M}_k^{(\mathrm{px})}\}_{k=0}^{T-1}$ 下采样至时空令牌网格，并获得令牌级语义掩码 $\mathbf{S} = \{\mathbf{S}_\tau\}_{\tau=0}^{t-1}$。定义意图指定令牌集合和非意图指定令牌集合：

$$\mathcal{S} = \{(\tau,i,j) : \mathbf{S}_\tau(i,j) = 1\}, \qquad \mathcal{N} = \{(\tau,i,j) : \mathbf{S}_\tau(i,j) = 0\}.$$

返回 S，$\mathcal{S}$，以及 $\mathcal{N}$。

C. 语义感知多速率比特编码

根据令牌化器生成的离散令牌索引图 $\mathbf{Z} = \{\mathbf{Z}_\tau\}$ 以及获得的语义令牌掩码 $\mathbf{S}_\tau$，我们根据语义重要性分配比特精度。用户意图指定区域内的令牌采用完整码本精度传输，而区域外的令牌则使用在令牌索引差分空间中定义的精简码本进行编码。这样便实现了数字语义通信 [30]–[33]，在降低比特率的同时保留语义重要内容。

a) 意图令牌的完整码本精度编码：对于用户意图令牌，其中$\mathbf{S}_\tau[i,j] = 1$，$B_{\text{full}} = \lceil \log_2 N \rceil$ 是令牌索引的完整比特精度（例如，B_{full}=16 对应 64000 大小的码本），因此，用户意图区域令牌$\mathbf{Z}_\tau^{\text{intended}}[i,j]$ 以 B_{full} 比特传输。

b) 非意图令牌的缩减码本精度差分编码：对于非用户意图令牌，我们采用精度为 B_Δ 的差分编码。令$\mathbf{Z}_{\text{ref}}$ 表示取自首帧$\mathbf{I}_0$的参考令牌图。我们定义原始差分值

$$x = \mathbf{Z}_\tau^{\text{non}}[i,j] - \mathbf{Z}_{\text{ref}}[i,j]. \quad (9)$$

一个 B_Δ 位有符号量化器可表示范围内的整数值

$$-Q \le x \le Q, \qquad Q = 2^{B_\Delta-1}-1, \quad (10)$$

由此得到共 $2Q+1 = 2^{B_\Delta}-1$ 个可表示符号。有效表示集合 $\{-Q,\dots,0,\dots,Q\}$ 构成令牌索引差分空间中的一个精简码本，其中包含 $2Q+1 = 2^{B_\Delta}-1$ 个有效差分符号，采用 2^{B_Δ} 级无符号表示。为确保差分值落在此域内，我们应用对称裁剪算子

$$\Delta_\tau[i,j] = \mathcal{C}_Q(x) = \min\big(\max(x,-Q),Q\big), \quad (11)$$

其中，$\mathcal{C}_Q(\cdot)$ 将差值裁剪至对称范围 $[-Q,Q]$。由于比特流无法在不进行额外符号处理的情况下直接存储有符号整数，我们通过加上 Q 将该范围移至非负区间

$$\tilde{\Delta}_\tau[i,j] = \Delta_\tau[i,j] + Q \in \{0,\dots,2Q\}, \quad (12)$$

随后使用恰好 B_Δ 比特传输移位后的差分值 $\Delta_\tau[i,j]$。这种多速率比特精度分配使编码器能够根据语义内容调整传输速率。

c) 令牌级意向比率与 *BPP* 分析：式 (4) 中的补丁级意向比率 ρ 表示首帧 CLIP 补丁网格上与意向相关的补丁。经过光流传播 (6) 和时空池化 (7) 后，在视频令牌时空网格 $(h,\ w,\ t) = (H/d_s,\ W/d_s,\ T/d_t)$ 上进行意向/非意向令牌分类。由此得到的令牌级意向比率定义为

$$\rho_s = \frac{N_s}{thw} = \frac{1}{thw}\sum_{\tau=0}^{t-1}\sum_{i=1}^{h}\sum_{j=1}^{w}\mathbf{S}_\tau(i,j), \tag{13}$$

其中，N_s 表示用户意图令牌的数量，$\mathbf{S}_\tau \in \{0,1\}^{h\times w}$ 为令牌级语义掩码。请注意，所得 ρ_s 值取决于 CLIP 阈值 ℓ 和令牌级阈值 θ。给定 ρ_s，视频的总比特载荷为

$$B_{\text{tokens}} = thw\big(\rho_s B_{\text{full}} + (1-\rho_s)B_\Delta\big), \tag{14}$$

因此，将其除以视频分辨率 $T\times H\times W$ 及 RGB 通道数进行归一化后，所得每像素比特数值为

$$\text{BPP} = \frac{1}{3d_t d_s^2}\big(\rho_s B_{\text{full}} + (1-\rho_s)B_\Delta\big). \tag{15}$$

所提出的语义感知多速率比特编码的完整描述见**算法 2**。

D. 基于 *UEP* 的失真与传输时延联合最小化

令 $\mathcal{T}$ 表示时间索引，(i,j) 表示空间令牌位置。语义重要性由视觉语言相关热图结合光流传播得出，从而生成意向令牌集合 $\mathcal{S}\subseteq\mathcal{T}\times\mathbb{Z}^2$ 和非意向集合 $\mathcal{N}$。每个类别 $c\in\{s,n\}$ 中的所有令牌共用一个从有限候选集 $\mathcal{K}_c$ 中选出的传输配置。

a) UEP 候选集合：对于意图令牌，仅使用完整语义比特精度 B_{full}，并从以下选取传输配置

$$\mathcal{K}_s = \{\text{QPSK}_{1/3},\ \text{QPSK}_{1/2}\}.$$

对于非意图令牌，编码器联合选择调制编码方案和差分比特精度 B_Δ。$\mathcal{K}_n$ 中的每个候选项形式如下

$$k = (B_\Delta^{(k)},\ m_k^{(n)},\ r_k^{(n)}),$$

其中，$B_\Delta^{(k)}$ 从降精度集合中选取

$$B_\Delta^{(k)} \in \mathcal{B}_n = \{16,15,14,13,12,11,10\}.$$

数对 $\big(m_k^{(n)},\ r_k^{(n)}\big)$ 由所选调制编码方案（MCS）确定，候选 MCS 集合定义如下

$$\mathcal{M}_n = \{\text{QPSK}_{1/3},\ \text{QPSK}_{1/2},\ \text{16QAM}_{1/2},\ \text{16QAM}_{3/4}\}.$$

其中，$m_k^{(n)}$ 表示调制阶数（每符号比特数），$r_k^{(n)}$ 表示信道编码率。只有当瞬时信噪比超过其激活阈值时，该配置才可行。

算法 2： 语义感知差分令牌编码与重建

输入 ：令牌索引 $\mathbf{Z} = \{\mathbf{Z}_\tau\}_{\tau=0}^{t-1}$，语义令牌掩码 $\mathbf{S} = \{\mathbf{S}_\tau\}$，位精度（$B_{\text{full}}, B_\Delta$）。

输出： 编码令牌流 $\mathbf{E}$ 和重建索引 $\widehat{\mathbf{Z}}$.

1. 参考令牌编码：

以全精度传输第一个令牌网格 $\mathbf{Z}_{\text{ref}}$：

$$\mathbf{E}_{\text{ref}} = \mathbf{Z}_{\text{ref}}.$$

2. Differential token coding (for $\tau \geq 1$): for $\tau = 1$ to $t-1$ do

  for *each token position* (i,j) **do**

    if $\mathbf{S}_\tau(i,j) = 1$ (user-intended) **then**

$$\mathbf{E}_\tau(i,j) = \mathbf{Z}_\tau(i,j) \quad (\text{use } B_{\text{full}} \text{ bits})$$

    else

      Compute differential w.r.t. reference:

$$x = \mathbf{Z}_\tau(i,j) - \mathbf{Z}_{\text{ref}}(i,j)$$

      Clip to $[-Q,Q]$, $Q = 2^{B_\Delta - 1} - 1$:

$$\Delta_\tau(i,j) = \mathcal{C}_Q(x)$$

      Shift and encode:

$$\mathbf{E}_\tau(i,j) = \Delta_\tau(i,j) + Q \quad (\text{use } B_\Delta \text{ bits})$$

3. 解码器重建： R

恢复 s igned 差异 异 ial用于非目标令牌的 ens:

$$\widehat{\Delta}_\tau(i,j) = \hat{\mathbf{E}}_\tau(i,j) - Q, \quad Q = 2^{B_\Delta - 1} - 1.$$

重建令牌索引：

$$\widehat{\mathbf{Z}}_\tau(i,j) = \begin{cases} \hat{\mathbf{E}}_\tau(i,j), & \mathbf{S}_\tau(i,j) = 1, \\ \text{clip}_{[0,N-1]}\big(\mathbf{Z}_{\text{ref}}(i,j) + \widehat{\Delta}_\tau(i,j)\big), & \mathbf{S}_\tau(i,j) = 0, \end{cases}$$

Return the encoded stream $\mathbf{E}$ and reconstructed tokens $\widehat{\mathbf{Z}} = \{\widehat{\mathbf{Z}}_\tau\}$.

---

b) PDU 聚合与候选项指标：令牌通过协议数据单元（PDU）传输。设 L_c 为每个 PDU 中的令牌数，H_c 为 PDU 的比特开销（例如报头、导频、循环冗余校验）。对于 $k\in\mathcal{K}_c$，在令牌比特精度为 $B_k^{(c)}$、码率为 $r_k^{(c)}$ 且调制阶数为 $m_k^{(c)}$ 时，频谱效率为

$$g_{k,\text{raw}}^{(c)} = \rho_{\text{oh}}\, m_k^{(c)} r_k^{(c)}, \tag{16}$$

其中，$\rho_{\text{oh}} \in (0,\ 1)$ 表示考虑参考信号、控制开销及其他非数据资源元素后的有效物理层资源利用因子。误块率以调制阶数 m 和编码率 r 为函数制成表格，并按照MCS值进行索引，与3GPP NR标准 [34], [35]类似。具体而言，相应的误块率为

$$p_k^{(c)} = \text{BLER}\Big(m_k^{(c)},\ r_k^{(c)} \mid \gamma\Big), \tag{17}$$

其中 γ 表示瞬时信噪比。在仿真中，$p_k^{(c)}$ 控制一个协议数据单元级的分组丢失过程，其中令牌被分组成每个包含 L_c 个令牌的协议数据单元，且每个协议数据单元均以

独立概率 $p_k^{(c)}$ 丢失。接收端会将丢失的协议数据单元中的令牌替换为参考令牌 Z_{ref}，以进行重建。

一个 PDU 总共承载 $B^{(c)}{}_{\mathrm{PDU}\,k} = L_c B_k^{(c)} + H_c$ 比特，包括有效载荷和开销。每个 PDU 的资源消耗（Hz·s）和传输时间（秒）为

$$\bar{C}_{\mathrm{PDU},k}^{(c)} = \frac{B_{\mathrm{PDU},k}^{(c)}}{g_{k,\mathrm{raw}}^{(c)}}, \qquad \bar{T}_{\mathrm{PDU},k}^{(c)} = \frac{\bar{C}_{\mathrm{PDU},k}^{(c)}}{W_B}, \tag{18}$$

其中，W_B 表示以 Hz 为单位的系统带宽。每个令牌对应的资源消耗和时延通过对 PDU 中承载的L_c 个令牌取平均值获得

$$c_k^{(c)} = \frac{\bar{C}_{\mathrm{PDU},k}^{(c)}}{L_c}, \qquad \tau_k^{(c)} = \frac{\bar{T}_{\mathrm{PDU},k}^{(c)}}{L_c}, \tag{19}$$

其中，$c_k^{(c)}$（单位为 Hz·s/令牌）和 $\tau_k^{(c)}$（单位为秒/令牌）正是联合失真–时延优化中使用的标量成本。量化失真采用指数模型近似

$$d_k^{(c)} = \alpha_c e^{-\beta_c B_k^{(c)}}, \tag{20}$$

该模型受到率失真感知理论 [36], [37], 的启发，用于描述失真和感知指标随速率变化时呈凸且单调不增的特性，其中 α，β 为通过实验中的曲线拟合确定的参数 [26]。

c) 类级决策变量：由于同一语义类别中的所有令牌共享一种传输配置，因此每个类别 $c \in \{s, n\}$ 都从其候选集 $\mathcal{K}_c$中恰好选择一个条目。这通过二元决策变量进行编码

$$z_k^{(c)} \in \{0,1\}, \qquad \sum_{k\in\mathcal{K}_c} z_k^{(c)} = 1, \tag{21}$$

其中，$z_k^{(c)} = 1$ 表示为属于类别 c 的每个令牌选择配置 k 。令 $N_s = |\mathcal{S}|$ 和$N_n = |\mathcal{N}|$ 分别表示意向集合和非意向集合中的令牌数量。由于类别c 中的所有令牌使用相同的配置，总失真、时延和资源消耗都成为选择变量的线性函数。具体而言，总失真为

$$D_{\mathrm{tot}} = N_s \sum_{k\in\mathcal{K}_s} d_k^{(s)} z_k^{(s)} + N_n \sum_{k\in\mathcal{K}_n} d_k^{(n)} z_k^{(n)}, \tag{22}$$

其中，类别 s 选择配置 k 时会产生 N_s 份失真 $d_k^{(s)}$，类别 n 同理。总时延为

$$\tau_{\mathrm{tot}} = N_s \sum_{k\in\mathcal{K}_s} \tau_k^{(s)} z_k^{(s)} + N_n \sum_{k\in\mathcal{K}_n} \tau_k^{(n)} z_k^{(n)}. \tag{23}$$

d) 资源约束：时频资源块中的固定比例 $\eta \in (0,1]$ 可用于令牌传输。由此得到的资源预算（单位：Hz·s）为

$$R_{\max} = \eta\, W_B\, T_{\mathrm{RB}}, \tag{24}$$

其中，W_B 和 T_{RB} 分别表示每个资源块的频率宽度和时间长度。总资源消耗（单位：Hz·s）为

$$N_s \sum_{k\in\mathcal{K}_s} c_k^{(s)} z_k^{(s)} + N_n \sum_{k\in\mathcal{K}_n} c_k^{(n)} z_k^{(n)} \le R_{\max}, \tag{25}$$

其应小于 $R_{\max}$。

e) 加权失真–时延优化：失真和时延通过其可行极值进行归一化

$$D_{\mathrm{norm}} = \frac{D_{\mathrm{tot}} - D_{\min}}{D_{\max} - D_{\min} + \epsilon_D}, \qquad T_{\mathrm{norm}} = \frac{\tau_{\mathrm{tot}} - T_{\min}}{T_{\max} - T_{\min} + \epsilon_T}. \tag{26}$$

在仿真中，归一化常数$\{D_{\min}, D_{\max}, T_{\min}, T_{\max}\}$ 根据有限候选集合上的类别极值计算。具体而言，令 $d^{(c)} = \min_{k\in\mathcal{K}_c} d_k^{(c)}$和 $d^{(c)} = \max_{k\in\mathcal{K}_c} d_k^{(c)}$，并以类似方式从 $\{\tau_k^{(c)}\}_{k\in\mathcal{K}_c}$定义$\underline{\tau}^{(c)}$和 $\overline{\tau}^{(c)}$。然后我们设定

$$D_{\min} = N_s\, \underline{d}^{(s)} + N_n\, \underline{d}^{(n)}, \quad D_{\max} = N_s\, \overline{d}^{(s)} + N_n\, \overline{d}^{(n)}, \tag{27}$$

$$T_{\min} = N_s\, \underline{\tau}^{(s)} + N_n\, \underline{\tau}^{(n)}, \quad T_{\max} = N_s\, \overline{\tau}^{(s)} + N_n\, \overline{\tau}^{(n)}. \tag{28}$$

这种按类别构造的方法在保持失真与时延之间相对尺度的同时，实现了数值稳定的归一化。优化目标为

$$\begin{aligned}
\min_{\{z_k^{(c)}\}} \quad & J = w_{\mathrm{D}} D_{\mathrm{norm}} + w_{\mathrm{T}} T_{\mathrm{norm}} \\
\text{s.t.} \quad & N_s \sum_{k\in\mathcal{K}_s} c_k^{(s)} z_k^{(s)} + N_n \sum_{k\in\mathcal{K}_n} c_k^{(n)} z_k^{(n)} \le R_{\max}, && \text{(29a)} \\
& \sum_k z_k^{(c)} = 1, && \text{(29b)} \\
& z_k^{(c)} \in \{0,1\}, && \text{(29c)} \\
& \sum_k p_k^{(s)} z_k^{(s)} \le p_{\max}^{(s)}, && \text{(29d)} \\
& \sum_k p_k^{(n)} z_k^{(n)} \le p_{\max}^{(n)}(\mathrm{SNR}), && \text{(29e)}
\end{aligned}$$

其中，$w_{\mathrm{D}} \ge 0$ 和 $w_{\mathrm{T}} \ge 0$ 是权重因子，用于控制语义失真与传输时延之间的权衡，且 $w_{\mathrm{D}} + w_{\mathrm{T}} = 1$。误块率约束确保所选工作点处于与瞬时信道条件相匹配的可靠性区域内，从而控制分组错误概率。

我们提出**算法3**，基于混合整数线性规划（混合整数线性规划） [38]来求解问题（29）中的联合比特精度与调制编码方案选择。一般而言，求解混合整数线性规划是NP难问题，其最坏情况下的复杂度会随整数决策变量的数量呈指数增长。在我们的建模中，包含二元变量$\{z_k^{(s)}\}_{k\in\mathcal{K}_s}$ 和 $\{z_k^{(n)}\}_{k\in\mathcal{K}_n}$，因此二元变量的数量为 $N_{\mathrm{bin}} = |\mathcal{K}_s| + |\mathcal{K}_n|$。由于候选集合 $\mathcal{K}_s$和 $\mathcal{K}_n$ 通常较小，且每个调度窗口只需求解一次混合整数线性规划，因此该优化在实际中仍具有计算可行性。

算法3： 联合位精度与MCS自适应

输入 ： 信噪比 γ，带宽 W_B，开销因子 ρ_{oh}，
资源预算 $R_{\max}$，令牌数量（N_s， N_n），
位精度集合 $\mathcal{B}_s$， $\mathcal{B}_n$，MCS集合 $\mathcal{M}_s$， $\mathcal{M}_n$，
PDU打包参数 $\{(L_c,\ H_c)\}_{c\in\{s,n\}}$，
权重（w_D， w_T）。
输出： （$b_S^\star$， $m_S^\star$， $b_N^\star$， $m_N^\star$）。

候选生成：
对 $c\in\{s,n\}$ **执行**
  如果 $c=s$ **则**
    $\mathcal{B}_s\leftarrow\{B_{\text{full}}\}$。
  对$(b,\ m)\in\mathcal{B}_c\times\mathcal{M}_c$ **执行**
    （*SNR*可行性）仅在 (b,m) 满足以下条件时保留： $\gamma\ge\gamma_{\min}(m)$。
    根据链路级表计算误块率：

$$p_k^{(c)}=\mathrm{BLER}(m_k^{(c)},r_k^{(c)}\mid\gamma).$$

    频谱效率：

$$g_{k,\text{raw}}^{(c)}=\rho_{\text{oh}}\,m_{\text{mod}}(m)\,r_m.$$

    PDU比特（有效载荷 +开销）： $B_{\text{PDU}}=L_c\,b+H_c$。
    每令牌资源代价： $c_k^{(c)}=\frac{B_{\text{PDU}}}{L_c\,g_{k,\ \text{原始值}}^{(c)}}$.
    逐token时延： $\tau_k^{(c)}=\frac{c^{(c)}}{k}$
    失真代理： $d=\alpha_c\exp(-\beta_c^{W_B}b)$。
    将 $(c_k^{(c)},\ \tau_k^{(c)},\ d,\ p_k^{(c)},\ b,\ m)$ 存储在类别-c
    候选列表中。
混合整数线性规划解： 求解联合比特精度与调制编码方案选择问题，以获得最优配置 $(b_S^\star,\ m_S^\star)$ 和 $(b_N^\star,\ m_N^\star)$。
返回 $(b_S^\star,m_S^\star,b_N^\star,m_N^\star)$。

E. 基于令牌的重建

在接收端，对于用户意图令牌，读取传输的完整值；对于非用户意图区域令牌，则通过逆偏移并加上参考帧的令牌来恢复。

$$\widehat{\Delta}_\tau[i,j]\ =\ \widetilde{\Delta}_\tau[i,j]-Q, \tag{30}$$

$$\widehat{\mathbf{Z}}_\tau[i,j]\ =\ \begin{cases}\mathbf{Z}_\tau^{\text{intended}}[i,j], & \mathbf{S}_\tau[i,j]=1,\\ \mathrm{clip}_{[0,N-1]}\big(\mathbf{Z}_{\text{ref}}[i,j]+\widehat{\Delta}_\tau[i,j]\big), & \mathbf{S}_\tau[i,j]=0.\end{cases} \tag{31}$$

此处，$\mathrm{clip}^{[0,N-1]}(\cdot)$ 如公式 (11) 所示，表示裁剪算子，用于将重建的令牌索引限制在有效码本范围内。用户意图令牌以全精度直接传输，而非用户意图令牌则根据其 B_Δ 位差分表示进行重建；该表示通过将原始差值裁剪至 $[-Q,Q]$ 获得，并传输其经过 Q 移位的量化形式。最后，我们使用令牌化器的解码器 $\mathcal{D}(\cdot)$ 对离散令牌进行解码

$$\hat{\mathbf{V}}=\mathcal{D}(\widehat{\mathbf{Z}}_\tau), \tag{32}$$

其中，重建视频为 $\hat{\mathbf{V}}\in\mathbb{R}^{T\times H\times W}$。

在解码器处，需要令牌级语义掩码$\mathbf{S}_\tau$来区分由完整索引重建的令牌与由差分重建的令牌。我们并非为所有帧传输$\mathbf{S}_\tau$，而仅发送首帧掩码，并利用令牌网格上的轻量级块级运动矢量在时间上对其进行传播。

对于大小为 $t\times h\times w$ 的令牌网格以及块大小为 $g\times g$的情况，所得相对边信息开销为

$$\frac{B_{\text{side}}}{B_{\text{tokens}}}=\frac{\frac{1}{t}+\frac{(t-1)}{t}\frac{B_{\text{mv}}}{g^2}}{\rho_s B_{\text{full}}+(1-\rho_s)B_\Delta}, \tag{33}$$

其中， $\rho_s=N_s/(thw)$ 表示用户意图令牌所占比例。采用较小的粗粒度块 g 和紧凑的运动量化B_{mv}时，辅助信息通常仅占令牌有效载荷的百分之几。

III. 实验结果

对于数据集，我们选用了广泛使用的 MCL-JCV [39] 和 UVG [40]视频数据集，格式为 YUV格式。我们的 MCL-JCV 子集经过裁剪，分辨率为 1024×640。对于UVG数据集，我们使用原生分辨率为1920×1080的视频。我们将每段视频中出现的主要物体配以说明文字，作为用户意图文本描述，并将其用作系统的文本意图输入。在本次仿真中，我们采用预训练的 Cosmos [41] DV-$8\times16\times16$ 和 DV-$4\times8\times8$ 模型作为视频令牌化器，并采用预训练的 CLIP [4] 作为视觉语言模型，用于语义相关性估计。值得注意的是，此处采用 CLIP 作为具有代表性的视觉语言模型；由于语义令牌提取仅依赖通用的图像文本相似度分数，因此可轻松替换为其他预训练视觉语言模型，而不会影响所提出的框架。此外，所提出的框架与令牌化器的具体选择无关，可适配任何能够生成令牌索引时空网格的离散视频标记化模型。为平衡速率与失真，我们将第 II-B1 节中的 CLIP 相似度阈值固定为 $\ell=0.6$，并将第 II-B3 节中的令牌级阈值固定为 $\theta=0.3$。此外，公式 (16) 中的物理层资源利用因子固定为 $\rho_{\text{oh}}=0.85$。公式 (20) 中的参数 α 和 β 通过曲线拟合进行经验确定；在所有实验中，我们均采用 $\alpha_c=1.0$ 和 $\beta_c=0.2$。为在最终失真与时延之间实现均衡折衷，我们将公式 (29) 中的权重因子设为 $w_{\text{D}}=0.5$ 和 $w_{\text{T}}=0.5$。

A. 与基线方法的比较

我们将所提方法与先进的、基于扩散模型的超低比特率生成式语义通信方案 VC-DM [25],以及传统且广泛采用的 H.265 [23], [24]数字视频压缩编解码器进行比较。如表 I 所示，VC-DM 在 UVG 视频上进行训练和评估，这些视频被裁剪至空间分辨率 128×128，时长为 30 帧。因此，为了公平比较，我们将 UVG数据集下采样并裁剪至相同的分辨率和帧数，并采用与 VC-DM 相同的峰值信噪比、学习感知图像块相似度和 FVD 计算流程。表 I 中的结果表明，所提出的框架在所有测试视频上的表现均持续优于 VC-DM 和 H.265。此外，与 VC-DM 和 H.265 的 0.02BPP 相比，我们的方法实现了更低的速率（0.013BPP），进一步证明了所提出的视频 TokenCom 框架具有更优的率失真效率。

表 I 所提出的视频 TokenCom 框架（每像素比特数=0.013）与 VC-DM [25] 和 H.265 [24] （每像素比特数=0.02）基线方法在 UVG数据集上的性能比较。

UVG 视频样本	PSNR			学习感知图像块相似度			FVD		
	TokenCom	VC-DM [25]	H.265 [24]	TokenCom	VC-DM [25]	H.265 [24]	TokenCom	VC-DM [25]	H.265 [24]
UVG-ShakeNDry	**26.06**	24.68	24.43	0.1296	**0.111**	0.158	**1076**	1400	2896
UVG-Beauty	**30.57**	28.45	25.28	**0.0469**	0.057	0.17	**554**	1416	3227
UVG-Bosphorus	**28.94**	24.55	26.70	**0.0863**	0.101	0.104	**1406**	2079	2951
UVG-HoneyBee	**22.49**	—	21.55	**0.1683**	—	0.196	**504**	—	1446
UVG-Jockey	**24.31**	22.82	20.95	**0.0933**	0.147	0.201	**965**	2349	6426
UVG-ReadySteadyGo	**22.38**	20.70	20.11	0.1307	**0.112**	0.316	**2213**	2832	6347
UVG-YachtRide	**25.94**	25.61	23.95	**0.087**	0.096	0.14	**1522**	2540	4282
平均值	**26.36**	24.47	23.28	**0.095**	0.104	0.184	**1289**	2087	4010



图 3. 所提出的 videoTokenCom 框架在不同文本意图下的性能：“女人正在敲打男人的手机。”和“天空。”传输用户意图指定令牌和非意图令牌的比特精度分别为每个令牌 16 比特和 11 比特。红色矩形显示由文本意图引导的用户意图指定区域。

B. 基于文本意图的源编码

首先，我们评估了不同用户指定文本的可控性，并使用热力图和光流图对结果进行可视化。在图 3 的热力图和光流图中，红色和橙色的点表示与用户意图高度相关的像素区域，黄色点表示中度相关，绿色则表示相关性较低。从图 3 的结果可以看出，当用户指定文本为 “女人正在打男人的手机”时，热力图中的光流更集中于与用户指定文本相关的 女人、男人、打和手机 区域。

传输用户意图指定令牌时，每个令牌的比特精度为 16 比特，而非意图令牌则使用差分缩减码本，以每个令牌 11 比特的精度传输。此外，与将“*Sky*” 作为意图文本相比，红色边界框中女性的重建质量有所提升，视频的整体性能指标也更高。相反，当用户指定文本为“*Sky*”时，语义重要性会转向与天空相关的区域，从而提高天空区域的重建质量，同时减少与之相关性较低区域的比特分配。尽管意图不同，两种情况下的整体比特率仍相近，



图 4. 所提出的 videoTokenCom 框架的性能。红色矩形标示出由文本意图“汽车和人”引导的用户意图区域。

约为 160 kbps，因为两种文本意图覆盖的像素数量相当。这表明，所提方法能够根据不同的文本意图，保持用户意图指定区域的重建质量。

在图 4 中，为了展示我们提出的文本意图引导视频 TokenCom 框架在较长视频上的性能，我们在来自 Pexels [42] 的一段 1920×1080 分辨率、包含 400 帧的视频上评估了不同的位精度。图 4(a) 使用高质量的 Cosmos 分词器 DV-4×8×8，其比特率为 4608 Kbps。当信道/网络状况显著恶化时，所提出的 TokenCom 框架可以切换到质量较低的分词器（此处为 Cosmos 分词器 DV-8×16×16 ），将传输比特率降至 691 Kbps。如果信道/网络状况进一步恶化，文本意图感知功能还能使我们提出的框架进一步将比特率降至 590 Kbps，同时保持意图区域的质量；这些区域由文本意图 “汽车和人” 指定。该结果展示了我们提出的框架出色的自适应能力：既可以更换分词器，也可以在分词器保持不变时利用文本意图。

C. 信道自适应视频*TokenCom*

为展示我们联合自适应信源信道编码方案的有效性，我们进一步分析在不同带宽（BW）预算下失真和时延的变化，以实现时延感知语义通信 [31], [43]。如图 6(a) 所示，我们评估了 MCL-JCV 数据集中的“女人正在打男人的手机”视频，条件如下

四种带宽设置，即 330 千赫、340 千赫、350 千赫和 360 千赫，同时保持时频资源块大小不变。在信噪比固定时，峰值信噪比会随着带宽预算的增加而单调提升。这是因为更大的带宽使优化器能够选择更高的比特精度和更强的 MCS 等级，从而降低量化失真并提升重建质量。此外，图 6(b) 展示了相应的时延。在信噪比固定时，增加带宽预算通常会缩短端到端传输时延。这是因为更大带宽带来的吞吐量增益超过了更高信源比特精度分配所导致的有效载荷大小适度增加。

我们进一步在 MCL-JCV 和 UVG 数据集上，将所提出的语义信源信道编码框架与自适应 H.265 基准进行比较。为确保比较公平，H.265 编码器的配置与我们的方法在总体比特率和资源预算上保持一致，其中 MCL-JCV 实验采用 350 千赫带宽，UVG 实验采用 1060 千赫带宽。具体而言，我们通过 FFmpeg 使用 H.265 的 x265（libx265）软件实现 [44]作为基线视频编解码器。编码器以恒定比特率模式运行，并将比特率配置为与每个（视频，信噪比）组合对应的 TokenCom 配置相匹配。如表 II 所示，H.265 基线采用传统的基于像素的信源编码，每帧使用单一 MCS；而我们的 TokenCom 系统则在相同资源预算下，采用内容感知词元编码、语义驱动比特分配和逐类不等差错保护

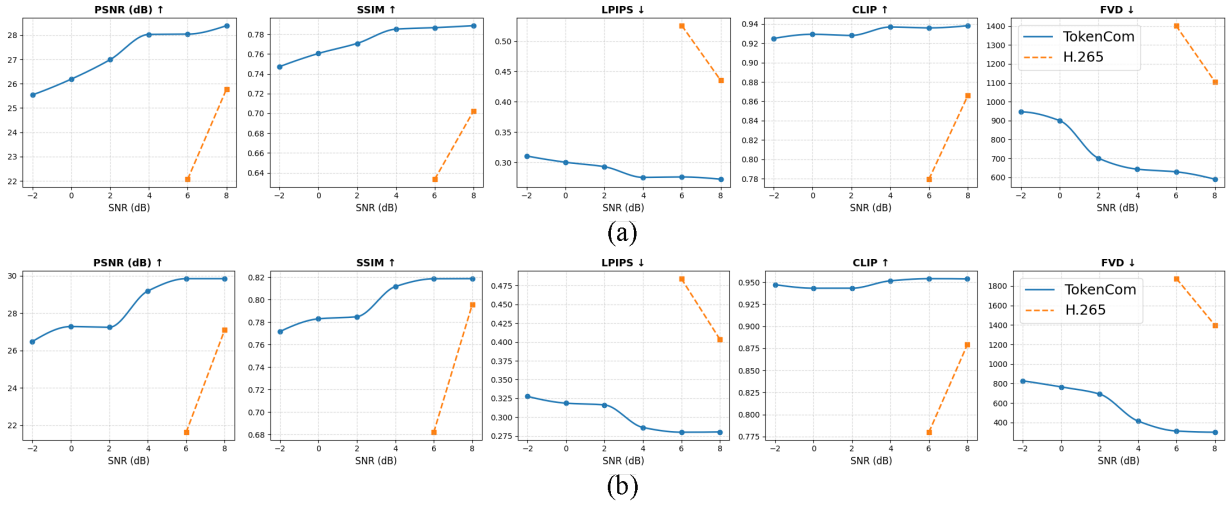


图5. 在 MCL-JCV (a) 和 UVG (b) 数据集上与基准方法的对比结果。请注意，在低信噪比下未显示部分 H.265 数据点，因为自适应 H.265 流水线经常无法解码超过 85% 的帧，这类情况在我们的评估中标记为无效。

表 II 所提出的视频 TokenCom 框架与 H.265 的关键参数对比

	TokenCom（所提出的方法）	**H.265（基线）**
信息单元	离散视频 token（码本索引）	像素域块（编码树单元/编码单元/变换单元）
编码粒度	令牌级全比特 / 差分比特	块级（16–64 px）
信源编码	多速率令牌编码（全精度）	变换 + 预测 + 上下文自适应二进制算术编码
语义感知	显式（CLIP 热力图 + 光流-传播）	None
码率自适应	逐token预期与非预期比特分配	基于QP恒定比特率模式；统一帧/切片
不等差错保护	语义不等差错保护（语义不等差错保令牌类别的速率 MCS）	无不等差错保护
传输单元	令牌 bits also 打包成协议数据单元	比特流分段为协议数据单元
协议数据单元大小	自 1024 调整为 1280 字节	1200 字节
资源预算	总计 $B \times T_{\mathrm{RB}}$	总计 $B \times T_{\mathrm{RB}}$
实验中的带宽-验	350 千赫（MCL-JCV），1060 千赫（UVG）	350 千赫（MCL-JCV), 1060 千赫 (UVG)

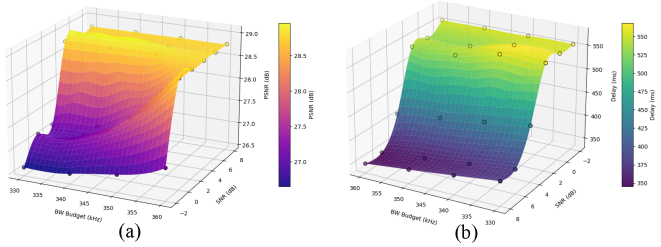


图 6. (a) 所提出的视频 TokenCom 框架在信源与信道编码失真优化方面可实现的峰值信噪比曲面；(b) 在信源与信道编码自适应方面可实现的时延曲面。

如图5(a)(b)所示，所提方法在所有信噪比水平和两种分辨率下，均在包括学习感知图像块相似度、CLIP 相似度和 FVD 在内的所有感知与语义指标上持续优于自适应 H.265。语义质量和感知质量的提升十分显著：我们的基于令牌的方案实现了明显更低的学习感知图像块相似度、更高的 CLIP 相似度以及大幅降低的 FVD，表明其语义保真度和时间连贯性更优。根据所采用的评估协议，只有至少 85% 的帧能够成功重建时，该工作点才被视为有效。引入这一阈值是出于务实的设计考量，以确保最低限度的时间

在视频播放中保持连续性和感知可用性，并始终应用于所有对比方案。在信噪比极低时，自适应 H.265 流水线往往会遭遇严重丢包，导致大量帧无法成功解码，因此图5中的 H.265 曲线不得不排除信噪比极低的点。相比之下，所提出的框架在所有信噪比和两种带宽下均能保持稳定的可解码性。这些结果表明，与 H.265 相比，所提出的语义信源信道编码框架能够实现稳健性显著更高、感知保真度更好的视频重建，尤其是在采用感知或语义指标进行评估，以及处理中高分辨率视频时。此外，表 III 列出了带宽为 340 千赫时各信噪比下的选择结果，证实所提出的优化器能够根据信道质量有效调整比特精度、调制方式和编码率。在不同带宽预算下，该算法始终能够选出充分利用可用资源并最大限度降低失真的调制与编码方案配置，展现了信源信道优化的实际效果。最后，在信噪比为 8 dB 时，所提出的视频 TokenCom 框架在 340 千赫带宽下，对视频序列“女人正在打男人的手机”的前 16 帧进行评估，达到 2.2×10^{-3} 的信道带宽比（CBR）。

表 III 所提出的视频 TokenCom 框架与 H.265 基线在典型视频上的并列比较。（“失败”表示成功重建的帧数不足 85%。）

(a) TokenCom（所提出的方法）

SNR	调制方式$_{int}$	$\mathbf{R}_{int}$	调制方式$_{non}$	$\mathbf{R}_{non}$	B_Δ	源比特率	PSNR
-2	QPSK	0.333	QPSK	0.500	12	165	26.69
0	QPSK	0.333	QPSK	0.500	12	165	27.40
2	QPSK	0.333	QPSK	0.500	12	165	27.40
4	QPSK	0.500	QPSK	0.500	16	173	28.82
6	QPSK	0.500	16正交幅度调制	0.500	16	173	28.82
8	QPSK	0.500	16正交幅度调制	0.500	16	173	28.82

(b) H.265（基线）

SNR	Mod	Rate	源比特率	PSNR
-2	QPSK	0.500	165	失败
0	QPSK	0.500	165	失败
2	QPSK	0.500	165	失败
4	QPSK	0.500	173	失败
6	16正交幅度调制	0.500	173	11.24
8	16正交幅度调制	0.500	173	27.77

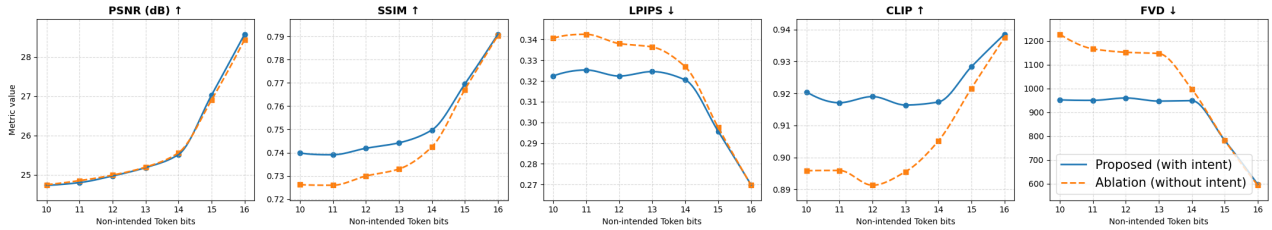


图 7. 在 MCL-JCV数据集上，所提出的视频 TokenCom 框架在有无文本意图引导时的消融结果。

D. 计算时延与开销

为评估我们方法的计算时延，我们使用单块 NVIDIA A6000 GPU，以分辨率为 1024×640 的视频作为输入，对 1000 帧视频进行了实验。对于全比特精度处理，每帧的平均计算时间为 65 毫秒；对于其他比特精度，每帧计算时间为 122 毫秒。此外，我们还测量了模型的计算复杂度。对于尺寸为 1024×640的输入视频，模型每帧需要 5，424 GFLOPs，包含 247。806 M 个可学习参数。

与视频载荷相比，将语义掩码传输至接收端的通信开销可以忽略不计。例如，对于包含 17 帧的 1024×640 视频片段，根据公式 (33) 计算显式发送语义掩码的开销；当 $B_{\text{full}} = 16$、$B_\Delta = 11$、块大小为 4×4 且预期令牌比例为$\rho_s = 0.7$时，掩码约占通信有效载荷的 1.7%。此外，还可以通过简单的游程编码或基于块的编码进一步降低这一开销。

E. 消融研究

作为消融研究，为证明所提出的文本意图感知速率自适应的有效性，我们将其与一种意图无关消融进行了比较；该消融移除了CLIP与光流模块，并改为随机选择非意图令牌。在MCL-JCV数据集上的定量结果如图7所示。我们测试了七种不同的 B_Δ取值，分别为10、11、12、13、14、15和16。正如预期，随着 B_Δ增大，各项指标总体呈改善趋势；在全精度下，所提出方案（含意图）与意图无关消融（不含意图）的性能相近。然而，在 B_Δ较低的受限速率取值下，所提出的意图感知方案在语义保真度与感知质量方面取得了明显优势，这在CLIP、LIPIPs和FVD指标中得到体现。需要注意的是，这三项指标评估与意图相关的语义保留，而峰值信噪比与结构相似性仅评估像素级失真。因此，CLIP、LIPIPs和FVD指标的提升表明，在受限比特率下，所提出的文本意图感知方案带来了语义保留增益。

IV. 结论

本文提出了一种由文本意图驱动的多速率视频令牌通信框架，可在未来的人工智能原生无线网络中实现语义与内容感知的速率自适应，从而高效传输多媒体。通过整合文本条件热图、光流传播和视频令牌化，我们建立了用户意图语义与离散视频令牌之间的细粒度映射。我们设计了一种多速率比特精度分配策略：为用户意图指定的令牌分配来自令牌码本的高精度索引，同时为非意图令牌采用来自精简码本的低比特差分索引，以降低传输开销。此外，我们设计了一种基于不等差错保护的信源—信道自适应方案，以在固定资源预算内平衡失真与时延。在多个视频数据集上开展的大量实验表明，所提出的视频 TokenCom 框架在广泛的信噪比范围内，在感知质量和语义质量指标上均优于传统基线方法和语义通信基线。值得注意的是，与传统的 H.265 视频压缩和传输方案相比，我们提出的 Video TokenCom 在峰值信噪比、结构相似性、学习感知图像块相似度、FVD 和基于CLIP的语义相似度方面均取得了稳定提升。特别是在典型的 6 dB 信道信噪比下，Video TokenCom 将 FVD 指标降低了近 1500。

参考文献

[1] L. Qiao, M. B. Mashhadi, Z. Gao, R. Tafazolli, M. Bennis 和 D. Niyato，“令牌通信：一种面向跨模态上下文感知语义通信的大模型驱动框架”，*IEEE*无线通信，第 32 卷，第 5 期，第 80–88 页，2025。

[2] L. Qiao, M. B. Mashhadi, Z. Gao, R. Schober 和 D. G¨und¨uz，“ToDMA：面向语义通信的大模型驱动令牌域多址接入”，arXiv 预印本 arXiv:2505.10946，2025。[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell 等，“语言模型是少样本学习器”，神经信息处理系统进展（NeurIPS），第 33 卷，第 1877–1901 页，2020。[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark 等，“通过自然语言监督学习可迁移视觉模型”，载于国际机器学习会议（ICML）。PMLR，2021，第 8748–8763 页。[5] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra 等，“语言并非你所需的一切：将感知与语言模型对齐”，神经信息处理系统进展（NeurIPS），第 36 卷，72 109，2023 72 096 页。[6] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen 和 C. Miao，“面向未来互联网的语义通信：基础、应用与挑战”，IEEE通信调查与教程，第 25 卷，第 1 期，第 213–250 页，2023。[7] D. G¨und¨uz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong 和 C.-B. Chae，“超越比特传输：上下文、语义与面向任务的通信”，IEEE通信选定领域期刊，第 41 卷，第 1 期，第 5–41 页，2023。[8] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran 和 D. Niyato，“面向语义通信的生成式人工智能：架构、挑战与展望”，IEEE无线通信，第 32 卷，第 1 期，第 132–140 页，2025。[9] L. Qiao, M. B. Mashhadi, Z. Gao 和 D. G¨und¨uz，“令牌域多址接入：利用语义正交性缓解碰撞”，载于 IEEE INFOCOM 2025——IEEE计算机通信会议研讨会（INFOCOM WKSHPS），2025，第 1–6 页。[10] S. Lee, J. Park, J. Choi 和 H. Park，“通过遗传束搜索实现令牌通信的语义分组聚合”，载于 2025 年 IEEE 第 26 届信号处理与无线通信人工智能国际研讨会（SPAWC），2025，第 1–5 页。[11] S. Lee, J. Park, J. Choi 和 H. Park，“通过前瞻搜索实现低复杂度令牌通信语义分组聚合”，arXiv 预印本，第 arXiv:2506.19451 号，2025。[12] A. Devoto, J. Pomponi, S. Petruzzi, P. Di Lorenzo 和 S. Scardapane，“面向人工智能原生目标导向通信的自适应语义令牌选择”，载于 IEEE Globecom研讨会（GC Wkshps），2024，第 1–6 页。[13] A. Devoto, J. Pomponi, M. Merluzzi, P. Di Lorenzo 和 S. Scardapane，“面向基于 Transformer的边缘推理的自适应语义令牌通信”，arXiv 预印本 arXiv:2505.17604，2025。[14] B. Liu, L. Qiao, Y. Wang, Z. Gao, Y. Ma, K. Ying 和 T. Qin，“面向无线图像传输的文本引导令牌通信”，载于 2025 年 IEEE/CIC 中国国际通信大会（ICCC），2025，第 1–6 页。[15] J. Peng, H. Xing, Z. Xiao, L. Xu 和 X. Lei，“利用选择性令牌进行训练的大模型赋能多模态语义通信”，IEEE信号处理快报，第 32 卷，第 2967–2971 页，2025。[16] S. Wang, J. Dai, X. Qin, K. Yang, K. Niu 和 P. Zhang，“Resicomp：通过双功能掩码视觉令牌建模实现抗丢失图像压缩”，IEEE视频技术电路与系统学报，第 35 卷，第 7 期，第 7181–7195 页，2025。[17] H. Wei, W. Ni, W. Wang, W. Xu, D. Niyato 和 P. Zhang，“大模型时代的令牌通信：一种基于信息瓶颈的方法”，arXiv 预印本 arXiv:2507.01728，2025。[18] J. Zhang, W. Ni, P. Wang 和 D. Wang，“资源受限多用户网络中由令牌通信驱动的多模态大模型”，arXiv 预印本 arXiv:2505.07841，2025。[19] J. Ying, Z. Qin, Y. Feng, L. Wang 和 X. Tao，“面向令牌通信的联合语义信道编码与调制”，IEEE无线通信汇刊，第 25 卷，第 8179–8193 页，2025。[20] K. Zhang, Z. Jin, Z. Cheng, M. Zeng, L. Qiao 和 Z. Fei，“TokCom-UEP：面向稳健图像传输的语义重要性匹配不等差错保护”，arXiv 预印本 arXiv:2511.22859，2025。[21] D. G¨und¨uz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong 和 C.-B. Chae，“超越比特传输：上下文、语义与面向任务的通信”，IEEE通信选定领域期刊，第 41 卷，第 1 期，第 5–41 页，2023。[22] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen 和 C. Miao，“面向未来互联网的语义通信：基础、应用与挑战”，IEEE通信调查与教程，第 25 卷，第 1 期，第 213–250 页，2023。[23] G. J. Sullivan、J.-R. Ohm、W.-J. Han 和 T. Wiegand，“高效视频编码（HEVC）标准概述”，IEEE视频技术电路与系统学报，第 22 卷，第 12 期，第 1649–1668 页，2012 年。[24] ISO/IEC，“信息技术——异构环境中的高效编码和媒体传输——第 2 部分：高效视频编码”，ISO/IEC 23008-2:2025，2025 年，国际标准。[在线]。可访问：https://www.iso.org/standard/90502.html [25] B. Li、Y. Liu、X. Niu、B. Bait、W. Han、L. Deng 和 D. G¨und¨uz，“利用预训练扩散模型进行预测的极限视频压缩”，载于《国际无线通信与信号处理会议（WCSP）》。IEEE，2024 年，第 1449–1455 页。[26] X. Liu、M. B. Mashhadi、L. Qiao、Y. Ma、R. Tafazolli 和 M. Bennis，“少传输，多合成：基于意图的时延感知生成式语义组播与扩散模型”，IEEE车辆技术学报，第 1–16 页，2026 年。[27] J. Men、M. B. Mashhadi、N. Wang、Y. Ma 和 M. Nilsson，“面向视线引导的立体媒体传输的自适应语义通信”，IEEE元宇宙计算、网络与应用国际会议（MetaCom），第 270–276 页，2025 年，IEEE。[28] V. Jampani、R. Gadde 和 P. V. Gehler，“视频传播网络”，载于《IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集》，2017 年，第 451–461 页。[29] Y. Weng、M. Han、H. He、M. Li、L. Yao、X. Chang 和 B. Zhuang，“用于高效视频语义分割的掩码传播”，《神经信息处理系统进展》（NeurIPS），第 36 卷，第 7170–7183 页，2023 年。[30] J. Huang、K. Yuan、C. Huang 和 K. Huang，“D^2-JSCC：面向语义通信的数字深度联合源信道编码”，IEEE通信选定领域期刊，2025 年。[31] L. Qiao、M. B. Mashhadi、Z. Gao、C. H. Foh、P. Xiao 和 M. Bennis，“基于预训练扩散模型的时延感知生成式语义通信”，IEEE无线通信快报，第 13 卷，第 10 期，第 2652–2656 页，2024 年。[32] Y. Oh、J. Park、J. Choi、J. Park 和 Y.-S. Jeon，“信道自适应数字语义通信的盲训练”，IEEE通信学报，2025 年。[33] C. Xu、M. B. Mashhadi、Y. Ma、R. Tafazolli 和 J. Wang，“基于基础模型的生成式语义通信：感知误差分析与语义感知功率分配”，IEEE通信选定领域期刊，第 43 卷，第 7 期，第 2493–2505 页，2025 年。[34] 3GPP，“TS 38.214：NR；数据的物理层过程”，第三代合作伙伴计划（3GPP），技术规范 38.214，2022 年。[35] 3GPP，“TS 38.212：NR；复用与信道编码”，第三代合作伙伴计划（3GPP），技术规范 38.212，2022 年。[36] Y. Blau 和 T. Michaeli，“重新思考有损压缩：率失真感知权衡”，载于《国际机器学习会议（ICML）》。PMLR，2019 年，第 675–685 页。[37] J. Chen、L. Yu、J. Wang、W. Shi、Y. Ge 和 W. Tong，“论率失真感知函数”，IEEE信息论领域选刊，第 664–673 页，2022 年。[38] G. L. Nemhauser 和 L. A. Wolsey，《整数与组合优化》，Wiley-Interscience离散数学与优化系列。Wiley-Interscience，1988 年，第 18 卷。[39] H. Wang、W. Gan、S. Hu、J. Y. Lin、L. Jin、L. Song、P. Wang、I. Katsavounidis、A. Aaron 和 C.-C. J. Kuo，“MCL-JCV：基于JND的H.264/AVC视频质量评估数据集”，载于《2016年IEEE国际图像处理会议（ICIP）》，2016 年，第 1509–1513 页。[40] A. Mercat、M. Viitanen 和 J. Vanne，“UVG数据集：用于视频编解码器分析与开发的50/120fps 4K序列”，载于《MMSys 2020——2020年多媒体系统会议论文集》。ACM，2020 年，第 297–302 页。[41] NVIDIA，“Cosmos分词器”，https://github.com/NVIDIA/Cosmos-Tokenizer，2024 年，访问日期：2025-11-30。[42] “Pexels”，https://www.pexels.com/。[43] B. Chong、H. Lu 和 G. Cheng，“利用语义通信实现超可靠低时延通信”，IEEE通信学报，第 73 卷，第 11 期，第 11 340–11 354 页，2025 年。[44] MulticoreWare公司和x265项目，“x265 HEVC编码器”，https://x265.readthedocs.io/，访问日期：2026-01-19。