Morphe：借助视觉基础模型实现高保真生成式视频流媒体传输

龚天一[1,2], 曹子健[1,2], 张子兴[1,2] *Jiangkai Wu*[3] *Xinggong Zhang*[3], 崔曙光[1,2], 王方欣[1,2]l香港中文大学（深圳）理工学院，[2]深圳市未来网络研究院，[3]北京大学王选计算机研究所

arXiv:2602.03529v1 [cs.NI] 3 Feb 2026

摘要

视频流传输是一项基础的互联网服务，但其质量仍无法得到保障，尤其是在带宽受限和偏远地区等网络条件较差的环境中。现有研究主要沿着两个方向展开：传统像素编解码流媒体传输几乎已达到极限，压缩率难以进一步提升；新兴的神经增强或生成式流媒体传输通常在延迟和视觉保真度方面表现不足，阻碍了其实用化部署。

受视觉基础模型（VFM）近期成功的启发，我们致力于利用 *VFM* 强大的视频理解与处理能力，实现实时视频流传输的泛化、高保真度和丢包韧性，同时进一步提高压缩率。我们提出了首个革新性范式，旨在实现基于 VFM 的端到端生成式视频流媒体传输。具体而言，Morphe[1] 采用视觉分词器联合训练和可变分辨率时空优化，并在模拟的网络约束下进行优化。此外，我们还构建了一个稳健的流媒体传输系统，通过智能丢弃数据包来抵御现实世界中的网络扰动。大量评估结果表明，与 H.265 相比，Morphe 在实现相当视觉质量的同时节省了 62.5% 的带宽，并在复杂网络环境中实现了实时、抗丢失的视频传输，为基于 VFM 的多媒体流媒体传输解决方案树立了里程碑。

1 引言

如今，视频内容占据了互联网流量的主导地位，根据 [26]，其占比超过 65%。然而，客户端与服务器之间的网络带宽不足且不稳定，常常妨碍用户访问高质量视频内容；对于高分辨率直播视频流媒体传输而言，情况更是如此。在高速铁路等网络受限的场景中以及偏远地区，人们对优质媒体内容的需求仍难以得到满足。

主要路线图： 为解决这一问题，多媒体流媒体传输采用编解码器优化、自适应码率控制以及网络感知协议，以在受限条件下保持视频质量。传统方法（例如，采用 H.264/265 [29, 34] 并结合自适应码率 [27]）专注于像素级编码与压缩，追求像素级完美重建。然而，当面临带宽不足或波动时，它们难以保持最佳性能，从而导致压缩伪影或播放中断。尽管像素级编解码器方法已广泛应用于现代流媒体服务，但在带宽受限的环境中，视觉质量妥协成为必要，这类方法已几乎达到其物理极限。

神经增强流媒体 [13, 14, 42] 作为一种有前景的方向应运而生，它利用基于学习的后处理来实现显著的带宽节省，但通常会带来较大的计算开销和更高的延迟。生成式流媒体方法 [2, 32]试图从紧凑的语义表示中重建视频内容，本质上是根据高层描述生成帧，而不是传输像素 [47]。这些方法具有提高带宽效率的潜力，但通常会遭受难以接受的视觉质量下降，这通常源于较差的泛化能力。此外，由于语义与像素之间存在密集映射且缺乏恢复机制，轻微的丢包就可能导致严重失真。

视频流传输的核心原则是**以尽可能高的视频压缩率，流传输高保真、实时且抗丢失的视频内容**。然而，在复杂的网络条件下，现有的流传输系统——无论是传统的基于像素的编解码器，还是新兴的神经网络方法——都难以有效平衡这一三角关系。在任何一个维度上妥协，都会持续严重降低当前所有流媒体范式的服务质量（QoS）。

机遇与挑战： 视觉基础模型（视觉基础模型） [4] 因其

前三位作者龚天一、曹子健和张子兴对本文的贡献相同。王方欣是通讯作者。[1]我们的项目页面位于 https://inml-tygong.github.io/Morphe

在视频理解方面的卓越能力 [44]和生成能力 [36]而备受关注。视觉基础模型凭借其独特的“先理解后生成”范式，为内容感知视频流系统带来了变革性机遇。视觉基础模型通过无监督学习在大规模视觉语料库上进行预训练，擅长将冗余视觉数据提炼为紧凑的语义表示，使其天然契合人类的感知优先级——优先关注人脸、显著物体和文本元素，而非低层次的纹理细节。这种语义感知编码能够实现非凡的压缩效果（近期研究中超过 300× ）[23]，同时在保持感知优化质量的同时，超越传统的像素级指标。此外，视觉基础模型固有的零样本泛化能力，确保了其面对真实世界数据变化时仍能稳健运行。受视觉基础模型非凡潜力的启发，我们希望探究一个问题：我们能否实现基于视觉基础模型的视频编码，从而提供高保真、抗丢失且实时的流传输服务？

遗憾的是，根据我们在§2.4中的分析，将视觉基础模型应用于视频流传输仍面临三重挑战：

C1：如何在实现高保真视觉恢复的同时最大化基于视觉基础模型的压缩？ 视觉基础模型从根本上是为语义理解和生成任务而设计的，主要关注低频内容，这与要求保留精细纹理的高保真视频传输存在固有冲突。除了严重的细节丢失之外，时空不连续性在将视觉基础模型适配到视频流传输时也构成重大挑战，其可能源于较高的时空压缩比以及不当的视频分块策略。语义智能与高保真传输需求之间的根本错配，解释了为何简单地将视觉基础模型迁移到流传输场景无法重建照片级真实感视频内容。

C2：如何在计算密集型视觉基础模型下保障低延迟流传输？ 视觉基础模型通常在分词和推理方面需要大量计算开销，这与实时流媒体传输需求形成固有张力。计算瓶颈主要源于复杂的编码器-解码器架构，这类架构通过多阶段操作处理视觉输入，包括三维哈尔小波变换、因果残差块以及因果时空注意力机制 [1]。将视觉基础模型直接应用于高清视频流会加剧这一问题，因为顺序处理特性与注意力复杂度会造成显著的处理延迟。

C3：如何利用基于视觉基础模型的编解码器确保抗丢失且网络自适应的流传输？ 传统流媒体系统依赖基于冗余的纠错，但基于视觉基础模型的系统由于其语义压缩范式而面临独特的脆弱性。当网络状况剧烈波动时，过大的带宽会导致利用率不足，而带宽不足则可能造成数据包丢失。语义压缩流中的数据包丢失尤其具有破坏性，因为这可能触发灾难性的质量下降和跨帧传播的时间不一致。当前的视觉基础模型缺乏错误恢复机制和网络感知的自适应能力，无法根据信道状况动态调整压缩策略。

表 1：不同流媒体范式的比较。

技术范式	保真度	效率	鲁棒性
传统编解码器 [34] [29]	Low	High	High
神经增强 [43]	中等	中等	High
生成式神经编解码器 [7]	中等	中等	High
基于扩散的 [35]	中	Low	Low
基于视觉基础模型的 Morphe	High	High	High

设计与贡献： 为应对上述挑战，我们提出了 Morphe[2]，这是一种由视觉基础模型赋能的创新端到端视频流媒体系统。据我们所知，Morphe 是首个利用视觉基础模型实现高保真、低延迟且抗丢失的先进视频服务的系统，这一点也再次体现在表1中。我们的设计核心包括三个方面：由视觉基础模型微调而来的生成式编解码器，可同时确保高压缩比与高保真度；采用神经网络增强处理的分辨率缩放加速器，可显著降低计算复杂度；以及能够响应网络波动的自适应流媒体控制器。具体而言，Morphe 采用以下三项创新设计组件，以高效地将视觉基础模型应用于多媒体网络。

• **视觉增强生成式编解码器**：为应对 *C1*，我们设计了一种非对称时空压缩策略，对视觉基础模型（VFMs）进行微调，并最大限度地减少编解码过程中的视觉伪影。该模块通过先进的生成式重建和帧间平滑，在保留视觉质量的同时自动优化压缩效率，即使在极高压缩比下，也能实现前所未有的视频保真度和照片级真实感的播放质量。

• **分辨率缩放加速器**：为解决*C2*，我们提出了一种全面的加速框架，可显著提升系统效率。该模块通过预处理阶段的智能下采样和后处理阶段的基于深度学习的超分辨率，自动平衡计算复杂度与视觉质量，从而在资源限制下实现实时1080p流媒体。

• **网络自适应流媒体控制器**：为解决*C3*，我们构建了稳健的流媒体架构，采用可扩展码率控制和恢复机制，以应对网络动态和带宽限制。该模块会通过自动调整流媒体参数来适应网络条件

[2]Morphe 源自掌管梦境、塑造人类形体的希腊神祇。这体现了其核心功能：从高度压缩的潜在特征中，智能地塑造并重建高保真视频。

通过自适应码率算法和自纠错策略，即使在严重丢包的情况下，也能确保视频顺畅传输。

Morphe 在保持相当视觉质量的同时，将比特率较 H.265 降低了 62.5%。该系统在配备单块 NVIDIA RTX 3090 的服务器上实现了 65 帧每秒的实时流媒体传输，并在真实网络传输中达到 94.2% 的带宽利用率。本文的贡献总结如下：

• 我们指出了将视觉基础模型应用于视频流媒体系统的关键挑战，并证明现有压缩与流传输方法不足以在网络约束下实现高保真、实时、抗丢包视频传输（§2）。

• 随后，我们设计了Morphe，这是首个由视觉基础模型驱动的端到端视频流媒体系统，它融合了时空视觉增强模块以提升编解码质量，采用全面的预处理与后处理优化以提高系统效率，并配备智能流媒体控制器以应对网络波动（§4、§5、§6）。

• 我们提供了全面的实验与仿真，以证明该系统的适用性与可扩展性，结果表明其在带宽限制下可实现卓越质量与实时高清流媒体，并在严苛环境中面对丢包时仍保持稳健性能（§8）。

2 背景与动机

2.1 带宽受限的多媒体流媒体传输

全球视频通信中的带宽问题。设想身处乡村地区的商务旅客通过退化的 2G 网络参加重要的视频会议，而该网络仅支持 100 kbps 的带宽，远低于实现清晰视频通话所需的 300 kbps 最低带宽。尽管视频通信已无处不在，但由于带宽严重受限，数十亿人仍无法享受高质量的视觉连接。

这一限制并不只存在于发展中地区。应急响应人员需要通过受损的灾害网络获取清晰无比的视频流。偏远工业设施即使连接能力有限，也需要实时设备监控。高速交通系统面临间歇性连接，但乘客安全取决于不受影响的视频质量。如图 1 所示，我们对高速铁路旅程和边缘地区车辆行驶过程中的网络轨迹进行调查和收集，发现许多现实场景仍面临恶劣的网络条件。这些场景面临一个共同挑战：如何在带宽极度受限的情况下传输高质量视频内容。

2.2 传统视频流传输

低带宽下性能枯竭。当面临极端带宽限制时，当前的视频编码技术暴露出架构局限。H. 等传统编解码器 264 [34] 和 H.265 [29] 采用基于像素的帧内编码和帧间预测范式针对中等压缩比进行了优化。这些编解码器主要通过利用相邻像素块之间的相关性消除空间冗余，并通过帧间预测技术利用连续帧之间的相似性消除时间冗余，从而实现压缩。然而，这种消除冗余的方法存在一个根本上限——信息论物理约束限定了可实现的压缩比。

神经网络嵌入式编解码器 [10, 17, 19, 20] 通过利用神经网络学习复杂的运动模式和相关性，改进了传统方法，但仍受限于传统的冗余消除框架。这种架构上的局限性限制了其潜力，因为它们仍依赖像素级重建假设。在低于 50 kbps 时，这些方法会彻底失效，生成的视频存在严重的块效应、重影和细节丢失，因而无法用于实际观看。

2.3 基于学习的视频流

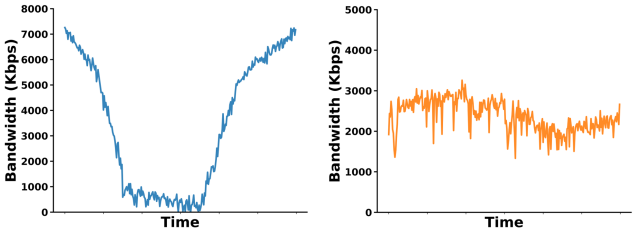
2.3.1 神经增强流媒体

学习不足与泛化能力有限。神经增强视频压缩利用深度学习，通过传输失真严重的低码率视频，再利用神经网络后处理进行恢复，从而实现更优的率失真权衡。与传统的基于冗余的方法不同，这种方法会主动丢弃大部分信息，依靠神经网络强大的恢复能力来找回丢失的细节，同时实现较高的压缩比。

然而，由于训练数据集与真实部署场景之间存在领域差异，这些 AI 驱动的网络面临显著的脆弱性 [40]。遇到分布外内容时，性能会下降，并产生不必要的高频细节等伪影以及幻觉效应，损害内容的真实性。尽管 NAS 等解决方案 [14, 42, 43] 提议针对每个视频片段微调网络，但传输这些适配后的模型会增加比特率，形成适应能力与带宽效率之间的根本性权衡。

2.3.2 抗丢失生成式编解码器

抗丢包生成式编解码器的发展经历了几个不同的技术范式。早期系统如 GRACE [7] 开创了基于丢弃的训练，但依赖过度简化的模型，假设数据包均匀随机丢失；在真实网络条件下，遇到时间上的丢包聚集时，性能便会下降。Reparo [18]等基于令牌的方法使用向量量化生成对抗网络 [11] 码本，将视频内容映射为离散索引，在保持感知质量的同时大幅降低比特率。除了前述泛化能力较弱和任务特定的局限性外，这种方法还存在其他根本性问题：



（a）乘坐火车时的网络轨迹 （b）乡间自驾游途经隧道时的网络轨迹

图 1：带宽受限的多媒体通信案例研究。

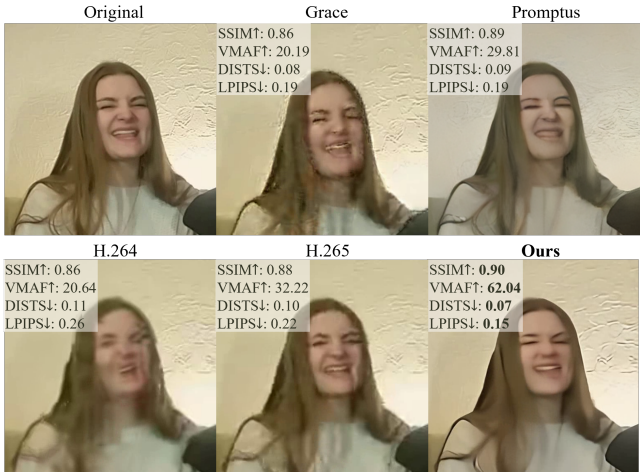


图 2：400 kbps 下不同流媒体技术的视觉感知效果。大多数现有方案在网络条件较差时都会产生严重伪影。

运动建模失效。 GRACE模型等系统对帧进行独立建模，导致时间一致性较弱，在极端压缩条件下，运动区域周围会产生严重的马赛克伪影。这种帧独立建模方法没有将视频作为整体进行考虑，从而导致灾难性的误差传播——这是视频传输中的典型问题：受损帧会沿时间依赖关系逐步扩散。

帧分辨率缺乏灵活性。 基于令牌的系统存在严格的分辨率依赖关系，受训练配置限制。在720p内容上训练的模型无法适应1080p等更高分辨率的输入，除非对码本和网络进行完整的重新训练。这种缺乏灵活性的特点限制了部署可扩展性，因此需要针对不同分辨率部署多个专用模型，并增加了系统复杂性。

2.3.3 基于扩散的生成式流媒体

基于扩散的方法在视频流传输领域展现出巨大潜力，DiffusionStream [15] 证明了实时传输的可行性。Promptus 等系统 [35] 利用紧凑的语义提示引导基于扩散的生成，推动了这一范式的发展，并实现了

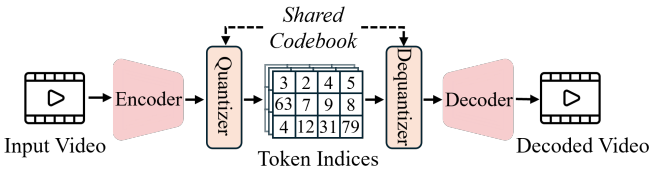


图 3：视觉基础模型中的视频令牌化流程。输入视频被编码为令牌，其体积通常比输入视频小得多。随后，解码器根据这些令牌重建输入视频。

在多种内容类型和任意分辨率下均表现出卓越的泛化能力。然而，基于扩散的流传输面临两项关键限制：

效率与质量之间的权衡并不理想。 扩散模型依赖大量去噪步骤，需要复杂的优化和硬件加速才能接近实时性能。要实现出色的视觉质量，需要长时间训练，因此只能部署于离线播放场景。此外，扩散可控性仍然较弱，经常产生人工智能伪影——例如时间上的不一致或不真实纹理——这些伪影很容易被察觉，且达不到专业应用所要求的照片级真实标准。

网络韧性较差。 现有的基于扩散的设计忽视了波动的网络状况，缺乏可靠的丢包韧性。扩散生成的序列特性容易造成脆弱性：提示损坏或传输不完整会进一步导致整帧重建失败。当提示信息部分丢失时，基于扩散模型的系统难以维持语义连贯性，因此不适用于移动流媒体常见的不可靠网络环境。

随着视频流传输技术的发展，我们仍不断看到许多局限性。如图2所示，当前流媒体系统在带宽不足和网络抖动的情况下，视频质量会显著下降。

2.4 推进视觉基础模型迈向高质量视频流媒体

视觉基础模型（VFMs）已成为变革性架构，为视频压缩带来了前所未有的语义理解能力和泛化能力。如图3所示，这些模型采用编码器-解码器架构，在压缩视觉数据的同时保留语义内容。编码函数f_ϕ 将输入视频$x_{0:T} \in \mathbb{R}^{(1+T)\times H\times W\times 3}$ 转换为紧凑的潜在表示$z_{0:T'} \in \mathbb{R}^{(1+T')\times H'\times W'\times C}$，通过多维下采样实现空间压缩因子$s_{HW}$ 和时间压缩因子s_T，随后由解码器g_θ进行重建。该架构能够实现语义压缩，并根据感知重要性带来显著收益。

然而，直接将视觉基础模型移植到多媒体

表 2：用于视频编码和解码的视觉基础模型对比分析。

模型	精度	编码（帧每秒）	解码（帧每秒）
VideoVAE Plus [37]	fp16	2.12	1.47
Cosmos [1]	fp16	6.21	5.08
CogVideoX-VAE [41]	fp16	5.52	1.95

流传输中并不可行。视觉基础模型以理解为导向的设计与流传输对质量的核心诉求之间存在根本矛盾，暴露出深层的不兼容性。如下所述，这些差距构成了我们必须应对的核心挑战：

高保真与感知有损的视觉基础模型。 视频流传输要求提供照片级真实感的播放体验，但视觉基础模型主要专注于视频理解与生成，其中低频内容表示发挥着至关重要的作用，却不可避免地丢失大量对视觉保真度至关重要的高频细节。当前的视觉基础模型分词器从根本上牺牲高质量细节以实现语义理解，使高保真流传输应用面临难以弥合的鸿沟。此外，视觉基础模型分词器固有的有损压缩会造成感知伪影——空间模糊、时间闪烁和幻影细节，这些都违背了人类视觉感知的预期。持续的视觉失真和较差的帧间一致性会带来令人不适的观看体验。

低延迟与计算密集型视觉基础模型。 视频流传输要求实现无缝流畅的播放体验，但视觉基础模型本质上是计算密集型设计，与实时流媒体传输需求存在根本矛盾。我们的研究发现，视觉基础模型处理高分辨率视频内容所需的时间对于实时应用而言高得难以接受。这些 AI 原生架构的推理延迟，在以标准帧率处理 1080p 视频流时会造成难以逾越的瓶颈，我们的研究结果如表 2 所示。这表明，视觉基础模型的计算需求与实现无缝视频流传输所必需的实时约束之间存在固有的不兼容性。

丢包韧性与网络无关型视频基础模型。 视频流媒体系统必须适应存在严重丢包的严峻网络环境，因此需要精细的质量控制来维持服务质量。然而，当前的视觉基础模型完全缺乏网络感知能力，作为孤立的黑盒运行，无法整合网络状况。传输过程中的丢包可能会彻底破坏语义令牌，导致重建图像完全失败，而非实现平滑质量降级。若要让基于视觉基础模型的系统在真实网络环境中稳定运行，抗丢包策略和网络自适应令牌化机制必不可少，但视觉基础模型从根本上缺少构建稳健流媒体系统所需的这些前提条件。

3 Morphe概述

针对上述挑战与洞见，我们为Morphe设计了三个模块：视觉增强生成式编解码器（§4）、分辨率缩放加速器（§5）和网络自适应流媒体控制器（§6）。 Morphe的工作流程如图4所示，具体细节如下：

视觉增强生成式编解码器（VGC）通过残差编解码、时间平滑和智能令牌丢弃，对预处理后的视频进行高保真压缩。与传统的视觉基础模型不同，经过微调、专注于视觉质量的 VGC 能保留重要的视觉细节，并灵活控制比特率。该编解码器生成变码率数据流，可根据 NASC 的网络反馈动态调整，从而在画质损失极小的情况下实现极高的压缩比。

分辨率缩放加速器（RSA）根据带宽动态选择适用于 VGC 编码的分辨率，以平衡画质与延迟，从而在资源受限的条件下实现实时 1080p 流媒体。预处理后的降采样帧会发送至 VGC 进行压缩；对于解码后的流，神经超分辨率网络会结合利用 VGC 的解码器特征来恢复全分辨率。

网络自适应流媒体控制器（NASC）负责协调 VGC 压缩令牌在波动的网络条件下进行流传输。NASC 通过联合管理 VGC 的比特率调整，实现可扩展码率控制，并提供具有自恢复机制的鲁棒流媒体协议。当带宽变化或发生丢包时，NASC 会优先处理关键语义令牌，并通过时空相关性重建缺失信息，确保即使在有损网络条件下也能稳定传输视频。

4 视觉增强生成式编解码器

如图 4 VGC 编码器和解码器所示，本节分三部分介绍 VGC 编解码器。首先介绍编解码器模型的整体架构（§4.1），然后说明如何抑制视频抖动并提高稳定性（§4.2），最后介绍如何实现可伸缩视频编码，以获得可调节的压缩比（§4.3）。

4.1 生成式编解码器架构

为兼顾高压缩比与视觉质量，VGC 采用具有非对称压缩特性的多接口编解码器框架。本节首先介绍编解码器架构，然后说明多接口框架。

VGC 的视频生成编解码器。原始视频基础模型主要用于提取视频的语义特征，以支持后续的视频生成或理解。为适应多种下游任务，它们通常提供两种标准压缩设置：（1） $8\times$ 时间下采样与 16×16 空间下采样，可实现较高的压缩比，但会带来明显的时间抖动

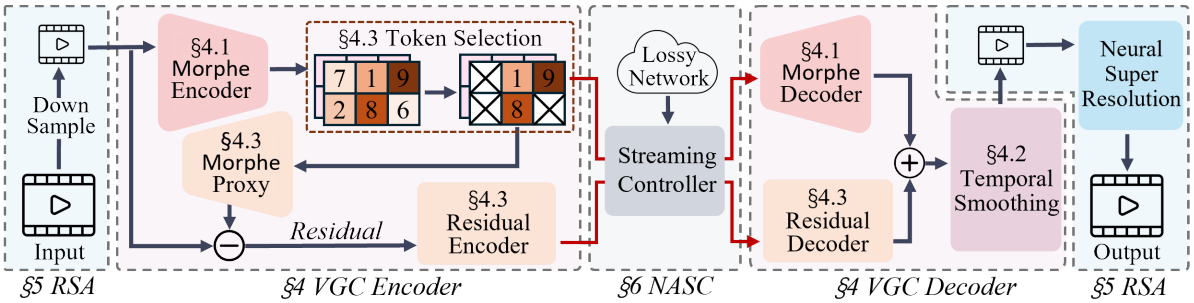


图4： Morphe 概览。它由三个核心模块组成：视觉增强生成式编解码器（§4）、分辨率缩放加速器（§5）以及网络自适应流媒体控制器（§6）。

并导致空间细节丢失。（2） 4× 时间下采样与 8×8 空间下采样，能够保留更多细节并实现更好的时间平滑，但压缩比较低，不适用于带宽受限的网络。压缩与视觉质量之间这一根本矛盾引出了一个关键的设计问题：我们能否找到一种平衡，既保留关键的空间细节、尽量减少时间抖动，又能适应窄带传输？

通过对流媒体视频流量的系统分析，我们发现，提升空间细节所需的数据远多于维持时间平滑所需的数据。这一发现为我们的设计提供了指导：优先保证空间细节，并在时间维度上加大压缩力度。具体而言，我们提出了一种非对称配置，在保持 8×8 空间压缩的同时，将时间压缩提高至 8×。这一选择既能保持较高的压缩比，又能保留高频空间细节，以满足窄带链路上的高保真视频需求。尽管更强的时间压缩会引入抖动，但§4.2 中的时间平滑增强技术能够有效缓解这一问题，使Morphe 能够在压缩效率与感知质量之间优化权衡。

多目标优化框架。 为提升视频重建质量和系统效率，我们在 VGC 中引入了一个多接口框架，其中包含两个关键接口：用于迭代优化的内部梯度接口，以及用于视觉增强的外部数据接口。梯度接口支持在不同阶段采用自适应训练目标，而外部接口则集成视觉增强模块和网络控制组件。VGC 的架构支持对视觉质量优化和网络自适应进行联合调度，为优化算法提供统一控制，并确保在严苛的网络条件下实现高质量流传输。

4.2 时间一致性增强

为减轻激进时间压缩造成的抖动，VGC 设计了一种利用编解码器协同优化的像素混合方案。该方案将模型优化与帧间平滑相结合，以抑制时间伪影。

像素混合与编解码器优化。受传统视频编解码器的启发，VGC 采用图像组（GoP）进行基于片段的编码和解码。然而，逐图像组处理与强时间压缩相结合，不可避免地会在图像组边界处造成时间抖动。一个直接的想法是在相邻图像组的边界附近混合像素，以实现平滑过渡。然而，由于 VGC 对每个图像组进行独立编码，会出现空间错位和亮度不一致。在这些情况下直接混合，往往会加重伪影和模糊，而非改善视觉连贯性。

为解决这一问题，VGC 使用了 §4.1 中介绍的内部梯度接口，并提出一种针对边界帧混合的迭代式编解码器协同优化方案。该方案分为两个阶段。

首先，为利用现有视觉基础模型中的先验并抑制闪烁，我们添加了一项专为图像组边界对齐设计的约束。具体而言，我们在边界区域计算像素级丢包率：

$$L_{\text{temporal}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{x}_{\text{curr},i} - x_{\text{prev},T-n+i}\right\| \tag{1}$$

其中，$\hat{x}_{\text{curr},\,i}$ 表示当前图像组重建出的前n 帧，$x_{\text{prev},\,T-n+i}$ 表示前一个图像组的最后n 帧。虽然这种像素差异监督无法完全体现时间连贯性，但它能有效利用模型先验并降低训练成本。加入这一约束后，图像组边界附近的帧在像素空间中会更加接近。

随后，对优化后的边界区域进行混合：

$$\hat{x}_{\text{blend},i} = \alpha_i \cdot \hat{x}_{\text{prev},T-n+i} + (1-\alpha_i)\cdot \hat{x}_{\text{curr},i} \tag{2}$$

其中， α_i 表示线性插值权重，我们将其设为$\alpha_i = {}_{n-\text{内}}$。由于训练会在图像组边界处确保像素邻近性，这种平滑处理使前一个图像组与当前图像组之间实现无缝过渡。借助这种平滑机制，VGC 显著提升了时间平滑度，且不会产生额外的传输成本。

4.3 可伸缩视频编码

为拓展编解码器的工作比特率范围并提升视觉质量，Morphe VGC采用可伸缩视频编码。在本

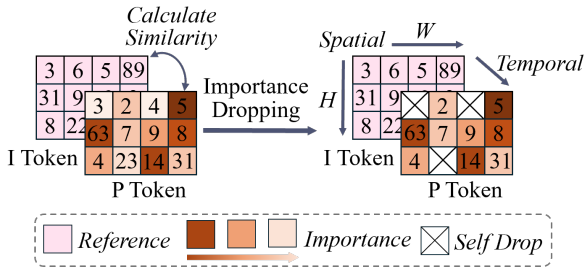


图5：基于相似度的令牌选择示意图。

节中，我们首先介绍基于相似度的令牌选择，然后说明像素残差计算与传输。**基于相似度的令牌选择**。 在低带宽网络中，即使采用激进的编解码器压缩，也可能超出可用带宽。一种直接的方法是随机丢弃语义令牌，但这会忽视语义重要性，可能导致关键视觉内容丢失和质量下降。受传统编解码器中 I/P帧结构的启发，我们发现 GoP语义表示中存在显著冗余。这使得进一步压缩成为可能：通过识别并有选择地丢弃相似令牌，即使带宽极其有限，我们也能保留核心视觉内容的传输。

VGC采用类似传统编解码器的I/P式结构。我们将每个图像组设为包含9帧。第一帧作为参考I帧，仅进行空间压缩，从而生成完整的语义令牌矩阵。后续8帧（P帧）则在8×处进行联合空间和时间压缩，生成另一个令牌矩阵。这种设计自然提供了令牌重要性估计的参考基准。

如图5所示，对于P帧令牌矩阵，我们根据每个令牌与相同空间位置上的I帧令牌之间的相似度来评估其重要性。具体而言，对于第k个图像组，令T_{Pk}表示由8帧联合压缩得到的令牌矩阵，令T_k^I表示第一帧（I帧）的令牌矩阵。我们计算每个令牌的余弦相似度：

$$S_{i,j} = \frac{T_k^{P(i,j)} \cdot T_k^{I(i,j)}}{\|T_k^{P(i,j)}\| \, \|T_k^{I(i,j)}\|} \tag{3}$$

其中，$T_{P(i,j)k}$和$T_{I(i,j)k}$分别是P帧和I帧令牌在位置（i,j）处的令牌向量。较高的$S_{i,j}$表示相对于参考帧具有更高的时间冗余；这类令牌的重要性可降低，并可在带宽压力下优先丢弃。

根据计算出的相似度，系统会根据当前可用带宽选择一个动态阈值τ，并将得分为$S_{i,j} > \tau$的令牌标记为可丢弃。该基于余弦相似度的评分方法效率很高，只需计算向量内积和范数，因此几乎不会带来额外开销。

像素残差压缩。 为增强高频细节并提供可扩展码率控制方案，我们引入了一种像素残差计算机制。

在编码器端，如图4所示，我们采用代理模型，实时将编码特征转换回像素。将原始帧x 与其重建图像$\hat{x}$进行比较，可得到逐像素残差$r = x - \hat{x}$，它能精确捕捉压缩过程中丢失的信息。然而，直接传输原始残差是不可能的：对于 1080p、30帧/秒的视频，残差流约为 1.39 Gbps，远超窄带链路预算，也与我们的压缩目标背道而驰。

因此，我们设计了一种新颖的残差处理流程。首先，我们发现相邻帧之间的残差往往相似，尤其是在静态或慢动作场景中。利用这一特点，我们通过平均对长度为T 的时间窗口内的残差进行聚合：

$$r_{\text{compressed}}(h,w,c) = \frac{1}{T}\sum_{t=1}^{T} r(h,w,c,t) \tag{4}$$

其中，h 和w 分别表示视频的高度和宽度，c 表示残差的通道数。这种时间压缩既减少了残差的体积，也降低了噪声，因为随机噪声会在平均过程中减弱。随后，将每个像素处压缩后的残差分配回窗口内的所有帧，以确保高视觉质量。

接下来，我们利用残差的稀疏性实现进一步压缩。由于 VGC 已保留大部分显著内容，残差数量通常较少。我们使用 θ 进行阈值处理，将绝对值低于 θ 的元素置零，从而得到高度稀疏的残差矩阵。

最后，我们采用传统视频编解码器中的算术熵编码，对稀疏残差进行无损压缩。由于稀疏度很高，算术编码能够实现出色的压缩效率。通过这些步骤，残差可压缩近 2000× ，且失真处于可接受范围内，因此即使在带宽有限的网络中也能进行传输。

5 分辨率缩放加速器

为解决原始视觉基础模型计算效率低下的问题，本节介绍一种基于分辨率缩放的加速器，以提高 VGC 的系统效率。

自适应分辨率控制。降低输入视频的分辨率是减少编码计算量和传输比特率最直接的方法。这带来多重益处：较低的比特率使分辨率控制成为 Morphe 中主要的速率调节手段，从而能够快速响应带宽变化；同时还能降低编码和解码复杂度，提升资源受限设备上的效率。为此，我们开发了一个自适应分辨率控制框架。

然而，传统的先下采样再进行超分辨率（SR）的流程在语义视频编码中面临两大挑战。首先，VGC 解码器生成的重建结果通常包含噪声和伪影，直接应用超分辨率会放大这些劣化，损害感知质量。其次，由于编解码器基于使用高质量数据预训练的视觉基础模型构建，下采样后输入低分辨率内容会降低重建性能，无法达到预期的视觉质量。

为解决这些问题，我们提出分阶段优化方案：训练编解码器生成适合超分辨率处理的输出，而不是强迫轻量级超分辨率模型适应编解码器伪影。这一设计是出于实时约束的考虑：轻量级超分辨率网络深度有限、参数较少，难以学习复杂的去噪映射。相比之下，编解码器的结构更丰富、容量更大，因此更适合学习如何生成符合目标分布的输出。

该过程采用两阶段方法：第一阶段在大规模数据上独立训练轻量级超分辨率模型，训练数据通过加入随机噪声、压缩伪影和质量退化进行增强，以建立稳健先验并定义预期的输入分布。第二阶段冻结超分辨率权重，并在高清视频上对 VGC 进行联合微调，使其重建结果与超分辨率模型的训练分布对齐，将目标函数重新定义为分布对齐，而不是从头学习复原映射。

我们通过第 §4.1 节中的内部梯度接口，端到端地进行联合微调。由于超分辨率模型已经针对模拟退化完成预训练，编解码器的梯度空间大幅缩小，从而使训练更加稳定高效。这种反向适配策略改善了质量与计算权衡。

6 网络自适应流媒体控制器

本节针对 Morphe 提出网络集成技术，其中 NASC 在实现丢包韧性的同时，支持可扩展比特率控制和稳健的流传输。

6.1 可扩展比特率控制

为适应动态网络带宽，Morphe实现了连续且可扩展的码率自适应。本节首先介绍 Morphe 如何通过协调三种码率控制技术来控制比特率：自适应分辨率控制（§5）、像素残差传输和智能令牌丢弃（§4.3）。随后，我们将介绍如何根据网络带宽自适应地选择目标比特率。

码率控制框架。 Morphe 采用基于锚点的线性控制方案。我们定义了两个关键锚点—— 3×下采样锚点（R_{3x}）和 2× 下采样锚点（R_{2x}）——它们作为启用不同码率控制策略的阈值。根据测得的可用带宽B_{avail}，控制器实时选择策略组合，相关细节见附录A.1。

在极低带宽模式下，系统以 3× 下采样来编码令牌，然后应用 §4.3中的相似性感知令牌选择，根据当前带宽调整丢弃率。在低带宽模式下，系统保留完整的 3× 下采样语义表示，并将剩余带宽分配给 §4.3中的像素残差，确保显著区域获得更高保真度的重建图像。带宽充足时，控制器会切换到 2× 下采样，以获得更高的基础

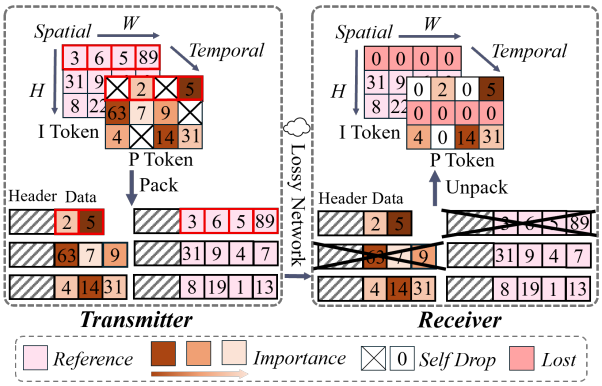


图6：令牌数据包化与混合丢包处理。

质量，并继续将多余带宽用于残差传输，以进一步提升细节。为确保观看流畅，模式切换采用滞回机制，避免因带宽抖动而发生振荡。分辨率变化时，系统会应用 §4.2 中的时间平滑技术，以保持帧间连贯性。

这种分层控制策略使Morphe 能够智能地组合多种机制，并在多样化的网络条件下，相较于传统编解码器提供更高的视觉质量和带宽效率（§8.3）。

自适应码率选择。 NASC 采用接收端驱动的控制架构，并使用 BBR（瓶颈带宽和往返传播时间） [6] 进行带宽估计。接收端每 100 毫秒报告一次估算的可用带宽，随后发送端调用上述码率控制器，动态调整编码参数。

6.2 稳健的流传输协议

传统视频流媒体系统采用多种丢包恢复机制，通常结合前向纠错（FEC）、自动重传请求（ARQ）和解码端错误隐藏 [30]。然而，在不利条件下，这些机制往往会导致延迟过高或视觉质量剧烈波动 [16]。

传统编解码器通过简单插值处理缺失的像素。相比之下，语义编解码器利用高层表示，并通过利用周围的语义和时空上下文来重建缺失内容。因此，我们设计了一种针对语义编解码器的流媒体协议，将传输与语义重建紧密结合。该协议包含两个核心组件：智能分包策略和混合丢包处理策略。这些组件协同工作，在不引入额外冗余的情况下实现具有韧性的视频传输。

智能分包。 NASC 提出了一种面向令牌的分包方案，将丢包重新定义为网络已经学会处理的去噪问题。如图6所示，对于编码器输出的 $H \times W$ 令牌矩阵，我们按行打包成 H 个数据包，每个数据包长度为 W。每个

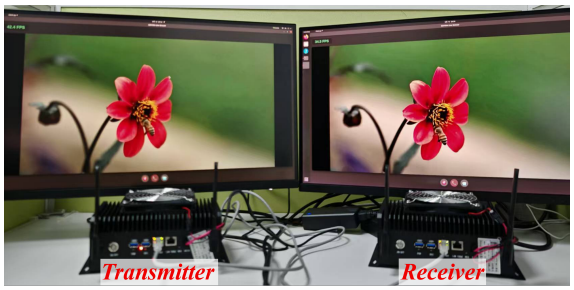


图7：系统原型中所用设备。两台 NVIDIA Jetson 分别作为发送端和接收端，通过网线连接，实现端到端传输。

数据包都包含一个包头和由有效令牌组成的有效载荷。包头记录原始矩阵中的行索引，以及数据包中令牌的位置掩码，该掩码表示为长度为W的二进制向量，其中 1 表示有效令牌，0 表示主动丢弃令牌。

一项关键创新是对缺失信息进行统一处理。在接收端，解码器利用包头中的行索引和掩码重组每一行收到的数据，将有效令牌放置在对应位置，并在掩码标记的位置填入零值，这些位置对应主动丢弃的令牌。如果数据包丢失，则整行填零。从解码器的角度来看，基于相似度的令牌丢弃和网络丢包表现得完全相同，都属于零值噪声。

由于Morphe 经过联合训练，能够从不完整令牌矩阵中重建内容（§7），因此丢包只会增加待处理噪声的数量，而不会引入新的故障模式。该系统利用编解码器的内在能力，无需额外的纠错层即可保持稳健性。

混合丢包设计。Morphe对语义令牌和残差采用不同的丢包策略。对于语义令牌，系统直接解码部分数据，仅当丢包率超过预设阈值（通常为 50%）时才触发重传。对于残差，策略则更宽松。如果残差数据包丢失，对应的帧只需跳过残差增强，而不必等待重传。其基本思路是，语义令牌承载核心内容，而残差主要用于补充细节。在网络受限或不稳定的情况下，持续传输核心内容比完美重建细节更为重要。

这种内容感知流式协议在条件恶化时会带来平缓的质量下降，而不是严重卡顿，从而显著改善用户体验。与此同时，由于系统可以在不完整数据上运行，重传得以减少，从而降低端到端时延并提高带宽利用率。

7 实现Morphe

Morphe 使用约 3.5 万行 Python 和 C++ 代码实现，包括其编解码器模型、可扩展码率自适应、流媒体协议控制以及基于 WebRTC 的原型。

编解码器训练。 Morphe 编解码器架构基于最新的 Cosmos 视觉基础模型 [1]。我们微调 Morphe ，从其在 Ultravideo-Long 数据集上对该视觉基础模型的预训练开始 [38]，采用两阶段训练策略。第一阶段在无丢包条件下训练基础模型，以优化帧间时间平滑性并支持自适应分辨率。第二阶段引入随机令牌丢弃训练（§4.3），丢弃率从 [0%，25%] 均匀采样。与仅解码器错误隐藏不同，Morphe 采用联合训练，使编码器与解码器在处理令牌丢失时保持对齐。训练目标函数结合了多种损失函数。[3] 我们在 16 块 NVIDIA A100 GPU 上训练 VGC，总计约消耗 195 个 GPU 小时。由于篇幅限制，关于Morphe 训练的更多细节见附录 A.2。**编解码器封装**。为实现端到端视频流传输，我们实现了一个封装模块，将 VGC 集成到传输流水线中。我们使用 PyTorch JIT 2.6 版本编译Morphe，并使用 BFP16 进行编码和解码计算。该封装在发送端和接收端维护模型参数状态，从而消除初始化开销。例如，当编码器调整其下采样因子时，解码器的超分辨率模块可以立即切换到相应的模型参数，而无需通过 PCIe 重新加载网络权重。

Morphe 原型。 我们构建了一个Morphe 原型，基于 WebRTC [3],分别运行在两台 NVIDIA Jetson AGX Orin（32 GB）设备上。如图 7所示，一台 Jetson 作为发送端，另一台作为接收端，两台设备均运行封装后的 Morphe 编解码器模型。运行期间，编解码器通过外部数据接口与可扩展码率控制（§6.1）模块交互，以实现动态码率自适应。两台设备通过网线连接。为在有损网络条件下测试我们的鲁棒流媒体协议（§6.2），我们还构建了一个基于轨迹的网络模拟器。该网络模拟器充当中继，在发送端和接收端之间接收并转发数据包。在模拟器中，数据包会被随机丢弃，并使用 Mahimahi [22] 重放来自 Puffer 的真实网络轨迹 [39]。发送端和接收端均使用 Qt 渲染播放界面，该界面在左上角显示实时编码帧率。

8 评估

8.1 设置

测试视频：我们的测试数据集由 100 个独特的 10 秒视频组成，采用 1080p 分辨率，包含 30 帧每秒和 60帧/秒的视频（总计 1000 秒）。这些视频经过精心挑选，来源于四个公开数据集：UVG [21],UHD [38],用户生成内容 [31],和 Inter4K [28]。这个全面且多样化的数据集使我们能够测试Morphe的

[3]请注意，ℓ_1ˆx 与x 之间的范数与ˆx （相对于x）的峰值信噪比高度相关。为避免这种偏差，我们在 §8 的评估中，使用 VMAF 等替代指标报告质量提升。

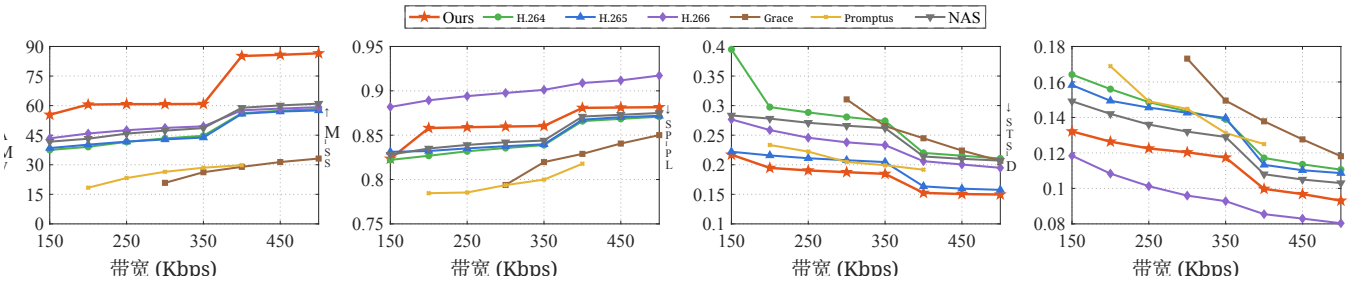


图 8：我们的生成式编解码器的率失真性能（UGC数据集）。

在现实视频流传输任务中的性能。

基线：为全面评估 Morphe 的性能，我们选取了一组具有代表性的视频编解码器作为基线。其中，首先纳入了应用最广泛的编解码器 H.264、H.265 和 H.266 [5]，（通过 FFmpeg v7.1.3 实现）。此外，我们还测试了近期的先进技术，例如代表神经编解码器的 Grace [7]，以及基于扩散模型的 Promptus [35]。**指标**：遵循视频流传输性能评估的最新实践 [8]，，我们选取了三个方面：

- **视觉保真度** 通过四项主要指标进行衡量：结构相似性指数[33],VMAF [24], LPIPS [46],和 DISTS [9]。为评估神经流式传输的时间稳定性，我们引入了帧间时间一致性指标。我们的首要目标是在低比特率下维持高质量的用户观看体验。这些指标综合反映了人类对视频播放性能的主观感知质量。
- **实时性能**通过测量复杂网络条件下编码器和解码器的逐帧处理时间，以及逐帧传输时延来评估Morphe 。

- **视频抗丢包能力**指系统在面对丢包条件时的鲁棒性和稳定性。我们报告丢包情况下的视觉质量，以及网络条件较差时客户端渲染的帧率。

8.2 视觉性能评估

压缩与视觉质量。我们进行了大量实验，以验证系统能够有效平衡激进压缩与高保真重建之间的权衡。图8表明，在所选带宽范围内，我们的生成式编解码器在用户生成内容数据集的大多数指标上显著优于传统编解码器和生成式基线。在400千比特/秒下，我们的方法取得了85.17的VMAF分数，而H.266为57.61，H.265为55.85，表明在带宽限制较低时具有更优的感知质量。尽管H.266在结构相似性指数和DISTS指标方面表现出色，但其计算复杂度以及当前硬件加速发展的局限性限制了其实用效率。值得注意的是，通过质量控制机制将带宽从350千比特/秒提升至400千比特/秒时，视频下采样比率

已从 3 倍调整为 2 倍。得益于高效的

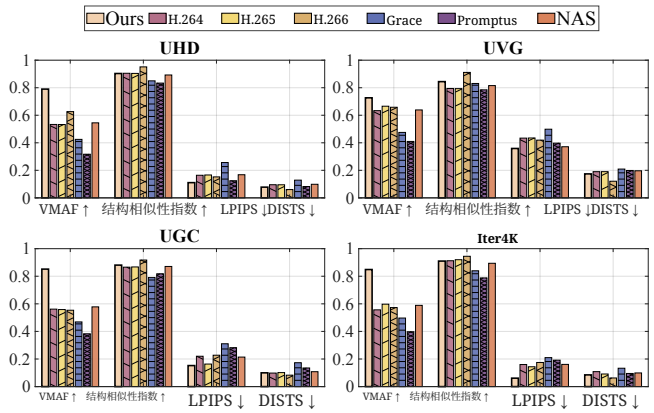


图9：四个数据集上的视觉指标。

时空压缩和下采样预处理，实现超过千倍的压缩，展现出广阔的应用前景。

Morphe泛化能力。我们在四个不同的数据集上验证了模型的性能。图9 表明，与其他方法相比，我们的方法始终取得更高的平均 VMAF 分数；同时，在所有感知度量（结构相似性指数、LPIPS、DISTS）上，其性能均与生成式基线方法相当，从而验证了其稳健的跨域适用性。关于以 400千比特/秒比特率传输的视频帧中 Morphe 与其他基线方法的视觉比较，更多详情见附录 A.3。

视频一致性。 如图 10 所示，我们给出了视觉时间一致性在两项指标上的分布曲线。我们计算每个视频连续帧之间的帧间残差，并通过将其与原始视频的残差进行比较来评估质量。对连续帧之间像素级变化的分析表明，包括 GRACE 和 Promptus 在内的所有基于神经网络的系统都比传统编解码器表现出明显更多的时间闪烁。为增强生成模型的时间连贯性，我们引入了一种时间平滑技术，以改善视觉感知体验。所提出的 Morphe 系统实现了更优的像素结构稳定性与一致性，大幅优于现有的基于神经网络的流媒体方法。

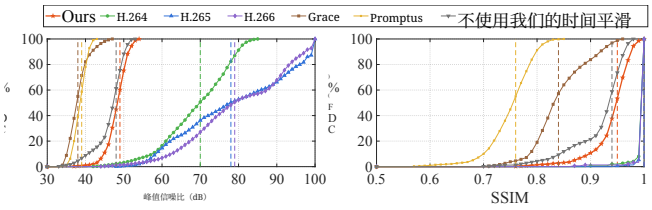


图 10：视频时间一致性对比。

表 3：不同设备的计算开销

设备	规模	GPU内存 (GB)	编码器 (帧每秒)	解码器 (帧每秒)
RTX3090	3x	8.86	98.51	65.74
	2x	17.09	47.14	32.03
A100	3x	7.96	101.23	83.33
	2x	16.24	52.54	40.19
Jetson	3x	15.21	61.17	43.45
	2x	23.87	31.87	24.93

8.3 网络系统评估

实时性能。我们在不同平台上进行了全面实验，包括 NVIDIA RTX 3090、A100 和 Jetson Orin 设备，结果见表 3。我们的评估聚焦于关键流传输指标，包括编码/解码帧率、内存消耗，以及不同分辨率缩放因子下的可扩展性（2×和 3×）。结果表明，我们的系统在所有平台上均实现实时性能：A100 可提供 101 帧每秒的编码和 83 帧每秒的解码，在 3× 缩放时，而 RTX 3090 则提供适合消费级流媒体传输的均衡性能。值得注意的是，即使资源受限的 Jetson Orin 仍能保持实用帧率（61/43 帧每秒，在 3×缩放时），从而验证了我们的系统在边缘设备上的可部署性。这些测量结果证明了该系统进行实时视频流传输的能力，并为从云服务器到边缘设备的各类流传输应用提供了最优部署策略。

丢包韧性。为了评估各系统在复杂网络环境下的丢包韧性，我们测试了 400千比特/秒下丢包率为 5% 至 25% 时的帧延迟和视频卡顿情况（以平均渲染帧率衡量）[25]。（Morphe 可容忍高达 30% 的丢包率，但为确保视觉质量，我们将其上限设为 25%。）图11 显示，即使在严重丢包的情况下，我们的系统仍能保持低延迟。在丢包率为 25% 时，Morphe 和 Grace 的帧中，超过 90% 的帧延迟低于 150 毫秒；而 H.266 则性能大幅下降，仅有 40% 的帧达到这一阈值。这验证了我们的抗丢包流媒体传输设计的有效性。图12 展示了丢包情况下的渲染帧率表现，采用两种目标帧率下的平均 FPS 进行衡量。我们的系统和 Grace 在所有丢包率下均保持稳定，即使丢包率达到 25%，帧率仍接近目标值。相比之下，H.266 性能严重下降，在丢包率为 25% 时帧率降至接近 1 帧每秒，视频几乎无法

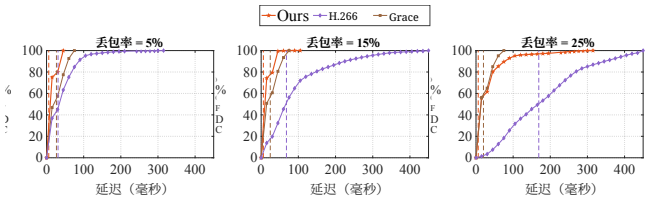


图 11：视频流传输的帧延迟分布在不同丢包率下。虚线表示中位数。

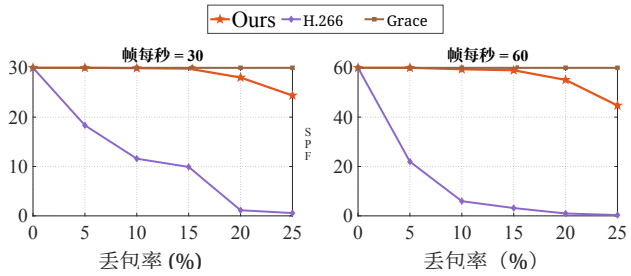


图12：比较以 30帧/秒和 60帧/秒编码的视频的解码帧率与渲染帧率。渲染帧数下降表明，丢包要么完全阻止了帧渲染，要么导致了严重的视觉质量下降。

无法观看。

除了评估复杂网络环境下的系统传输性能外，本研究还评估了丢包情况下的视觉质量。如图13所示，我们在 400千比特/秒下测试了丢包率为 5% 至 25% 时的多种视觉指标。对于衡量视频语义信息和客观质量的关键指标——VMAF、LPIPS 和 DISTS，Morphe不仅表现最佳，而且性能下降幅度最小，展现出很强的稳定性。相比之下，传统视频编解码器的性能随着丢包率上升而显著下降。GRACE模型的性能下降较为平缓，但其基于像素的特性使其在较差的网络环境下视觉质量本就较低。

通过将语义流传输与视觉基础模型的生成能力相结合，Morphe 在低带宽、有损网络中实现了卓越的高保真视频恢复，同时保持低延迟和流畅播放，确保出色的视觉质量和用户体验。

码率控制。为了评估系统对动态网络条件的适应性，我们使用 MahiMahi网络仿真器模拟了 200–500千比特/秒之间、周期为 30秒周期的码率波动。由于没有可用的开源码率控制，未将 Grace 纳入评估。如图 14所示，Morphe展现出出色的码率跟踪能力，其曲线与目标值高度吻合。这表明其可扩展的比特率控制效果良好，能够准确适应带宽变化，确保在复杂网络条件下稳定流传输。相比之下，传统编码器的性能波动显著。H.264 和 H.266 对目标的跟踪相对良好，但仍出现明显的波动和偏差。

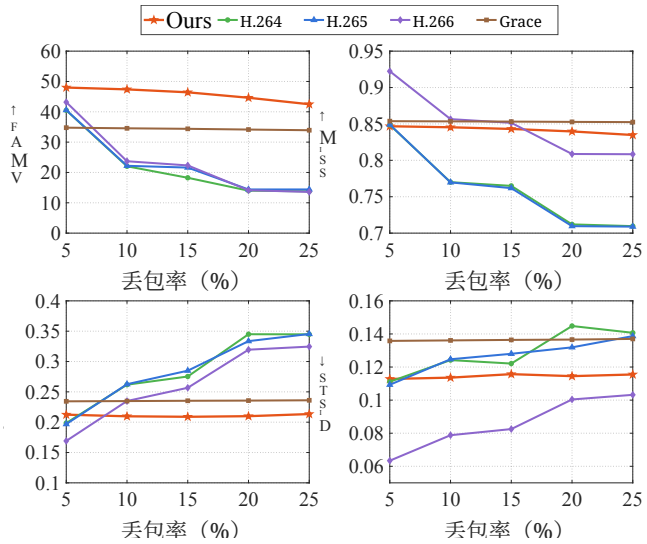


图13：不同丢包率下的视觉性能基准测试。

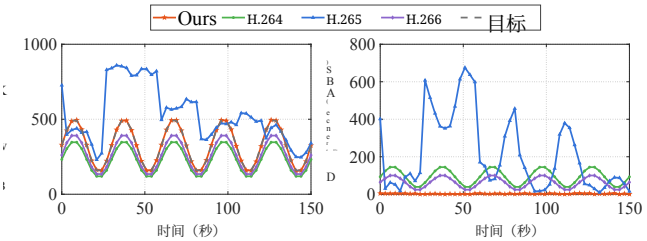


图14：网络带宽轨迹跟踪实验。

H.265 出现严重振荡，与目标比特率存在较大偏差，在网络快速变化期间码率最高达到 859.5 kbps，导致网络拥塞和丢包。

8.4 消融研究

表4 展示了消融研究，评估各组件对系统性能的贡献。表中的延迟指编码器/解码器延迟。移除智能自丢弃（§4.3）会导致最严重的性能退化：VMAF 从 60.76 降至 20.31，LPIPS/DISTS 也显著恶化，这表明该机制对于维持视觉质量至关重要。RSA（§5）模块通过降低延迟提高系统效率，使计算密集型人工智能模型更适用于实时流媒体。虽然残差（§4.3）会因代理模型而增加计算开销，但它显著提升了带宽利用率，并实现了更精细的质量调度。总体而言，这些结果验证了各组件在权衡质量与延迟方面的必要性。

我们在 50% 丢包率下测试了 Morphe，分别采用智能自丢弃和随机丢弃策略。如附录 16 图 A.4所示，智能丢弃在视觉质量和定量指标方面均优于随机丢弃。基于语义重要性的自丢弃机制

表 4：各模块贡献的消融研究

方法	VMAF↑	结构相似性指数↑	LPIPS↓	DISTS↓	延迟（毫秒）
不含 *RSA*	59.72	0.84	0.22	0.14	644.61 / 874.76
无残差	60.54	0.85	0.20	0.13	**78.21 /98.36**
无自丢弃	20.31	0.73	0.41	0.23	89.82 / 136.87
Morphe	**60.76**	**0.86**	**0.18**	**0.11**	91.36 / 136.91

注： 延迟值分别表示对一个包含 9 帧的视频块进行编码/解码所需的延迟（以毫秒为单位）。

有效降低了关键区域的视觉质量劣化。

为评估时间平滑的效果，我们对视频片段进行时间分析和可视化。如附录 A.4中的图 17以及图 10中的定量结果所示，不使用此机制的系统会出现明显更严重的时间闪烁，严重损害观看体验。

9 结论与讨论

我们介绍了Morphe，这是首个由视觉基础模型驱动的视频流媒体传输系统，可在严苛的网络条件下实现高保真、低延迟传输。与 H.265 相比，我们的系统在保持相当视觉质量的同时，将比特率降低了 62.5%；在实际部署中实现了实时流媒体传输和 94.2% 的带宽利用率。

我们由视觉基础模型驱动的方法也存在权衡：小型标识等细节可能会丢失，而由于当前视觉基础模型的文本训练有限，文本密集型内容仍然具有挑战性——这是生成式流媒体传输方法普遍存在的局限。尽管存在这些限制，Morphe 仍代表着视频流传输领域的一项重大进展，表明基于视觉基础模型的视频传输已可用于实际应用。

致谢

我们衷心感谢 shepherd Junchen Jiang 以及匿名审稿人提供的宝贵意见。本研究部分得到以下项目资助：中国国家自然科学基金委员会（项目编号：62293482）、河套深港科技创新合作区基础研究项目（项目编号：HZQB-KCZYZ-2021067）、中国国家重点研发计划（项目编号：2024YFB2907000）、国家自然科学基金委员会（项目编号：62471423）、深圳市科技计划（项目编号：JCYJ20241202124021028 和 JCYJ20230807114204010）、广东省人才计划（项目编号：2024TQ08X346）、深圳市杰出人才培养基金（项目编号：202002）、中国科协青年人才托举工程（项目编号：2022QNRC001）、广东省智能未来网络重点实验室（项目编号：2022B1212010001）以及深圳市大数据与人工智能重点实验室（项目编号：SYSPG20241211173853027）。Xinggong Zhang 和 Jiangkai Wu 的研究工作得到国家自然科学基金委员会（项目编号：62431017）的资助。

参考文献

[1] Niket Agarwal、Arslan Ali、Maciej Bala、Yogesh Balaji、Erik Barker、Tiffany Cai、Prithvijit Chattopadhyay、Yongxin Chen、Yin Cui、Yifan Ding 等。《Cosmos：面向物理人工智能的世界基础模型平台》。arXiv 预印本，arXiv:2501.03575，2025年。

[2] Emanuele Artioli。用于 HTTP 自适应流传输的生成式人工智能。载于第 15 届 ACM 多媒体系统会议论文集，第 516–519 页，2024年。

[3] Adam Bergkvist、Daniel C. Burnett、Cullen Jennings、Anant Narayanan 和 Bernard Aboba。WebRTC 1.0：浏览器之间的实时通信。W3C 建议，W3C，2021年。访问日期：2025年9月19日。

[4] Rishi Bommasani。论基础模型的机遇与风险。arXiv 预印本 arXiv:2108.07258，2021年。[5] Benjamin Bross、Ye-Kui Wang、Yan Ye、Shan Liu、Jianle Chen、Gary J Sullivan 和 Jens-Rainer Ohm。通用视频编码（VVC）标准及其应用概述。IEEE 电路与系统视频技术汇刊，31(10)：3736–3764，2021年。[6] Neal Cardwell、Yuchung Cheng、C Stephen Gunn、Soheil Hassas Yeganeh 和 Van Jacobson。BBR：基于拥塞的拥塞控制：测量瓶颈带宽和往返传播时间。Queue，14(5)：20–53，2016年。[7] Yihua Cheng、Ziyi Zhang、Hanchen Li、Anton Arapin、Yue Zhang、Qizheng Zhang、Yuhan Liu、Kuntai Du、Xu Zhang、Francis Y Yan 等。 {GRACE}：{具备丢包鲁棒性的}{实时} 视频神经编解码器。载于第 21 届 USENIX 网络系统设计与实现研讨会（NSDI 24），第 509–531 页，2024年。[8] Luca De Cicco 和 Saverio Mascolo。自适应视频流控制系统：建模、验证与性能评估。IEEE/ACM 网络汇刊，22(2)：526–539，2013年。[9] Keyan Ding、Kede Ma、Shiqi Wang 和 Eero P Simoncelli。图像质量评估：统一结构相似性与纹理相似性。IEEE 模式分析与机器智能汇刊，44(5)：2567–2581，2020年。

[10] Abdelaziz Djelouah、Joaquim Campos、Simone Schaub- Meyer 和 Christopher Schroers。《用于视频编码的神经帧间压缩》。载于 IEEE/CVF 国际计算机视觉会议论文集，第 6421–6429 页，2019年。

[11] Patrick Esser、Robin Rombach 和 Bjorn Ommer。《驯服用于高分辨率图像合成的 Transformer》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 12873–12883 页，2021年。

[12] 何恺明、张祥雨、任少卿和孙剑。《深度残差网络中的恒等映射》。载于《欧洲计算机视觉会议》，第630–645页。施普林格，2016年。

[13] Seunghwa Jeong、Bumki Kim、Seunghoon Cha、Kwang- gyoon Seo、Hayoung Chang、Jungjin Lee、Younghui Kim 和 Junyong Noh。《神经网络增强型自适应直播的实时 CNN 训练与压缩》。IEEE 模式分析与机器智能汇刊，46(9):6023–6039，2024年。[14] Jaehong Kim、Youngmok Jung、Hyunho Yeo、Juncheol Ye 和 Dongsu Han。《神经网络增强型直播：通过在线学习改进实时视频接入》。载于 ACM数据通信特别兴趣小组关于计算机通信的应用、技术、架构与协议年度会议论文集，第107–125页，2020年。[15]Akio Kodaira、Chenfeng Xu、Toshiki Hazama、Takanori Yoshimoto、Kohei Ohno、Shogo Mitsuhori、Soichi Sugano、Hanying Cho、Zhijian Liu 和 Kurt Keutzer。《StreamDiffusion：实时交互式生成的流水线级解决方案》。arXiv预印本 arXiv:2312.12491，2023年。[16] Marc Lambooij、Wijnand IJsselsteijn、Marten Fortuin、Ingrid Heynderickx 等。《立体显示器的视觉不适与视觉疲劳：综述》。《成像科学与技术杂志》，53(3):30201–1，2009年。[17] 李嘉豪、李斌和卢衍。《采用特征调制的神经视频压缩》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第26099–26108页，2024年。[18] 李天鸿、Vibhaalakshmi Sivaraman、Pantea Karimi、范立杰、Mohammad Alizadeh 和 Dina Katabi。《Reparo：适用于视频会议的抗丢包生成式编解码器》。arXiv预印本 arXiv:2305.14135，2023年。[19] 林建平、刘栋、李厚强和吴枫。《M-LVC：用于学习式视频压缩的多帧预测》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第3546–3554页，2020年。[20] 陆国、欧阳万里、徐东、张晓云、蔡春雷和高志勇。《DVC：端到端深度视频压缩框架》。载于

IEEE/CVF计算机视觉与模式识别会议论文集，第11006–11015页，2019年。

[21] Alexandre Mercat、Marko Viitanen 和 Jarno Vanne。UVG 数据集：用于视频编解码器分析与开发的 50/120 帧每秒 4K 序列。载于第 11 届 ACM 多媒体系统会议论文集，第 297–302 页，2020年。

[22]Ravi Netravali、Anirudh Sivaraman、Somak Das、Ameesh Goyal、Keith Winstein、James Mickens 和 Hari Balakrishnan。Mahimahi：用于 HTTP 的精准记录与重放。载于 2015 年 USENIX年度技术会议（USENIX ATC 15），第 417–429 页，圣克拉拉，加利福尼亚州，2015 年 7 月。USENIX协会。

[23] Huu-Tai Phung、Yu-Hsiang Lin、Yen-Kuan Ho 和 Wen-Hsiao Peng。探索用于图像压缩的自回归视觉基础模型。arXiv 预印本 arXiv:2509.05169，2025年。

[24] Reza Rassool。VMAF 可复现性：验证一种感知实用型视频质量指标。载于 2017 年 IEEE宽带多媒体系统与广播国际研讨会论文集，第 1–2 页。IEEE，2017 年。

[25]Michael Rudow、Francis Y Yan、Abhishek Kumar、Ganesh Ananthanarayanan、Martin Ellis 和 KV Rashmi。Tambur：通过流式编码为视频会议高效恢复丢包。载于第 20 届 USENIX网络系统设计与实现研讨会（NSDI 23）论文集，第 953–971 页，2023 年。

[26] Sandvine。2023 年全球互联网现象报告。技术报告，2023 年。

[27] Kevin Spiteri、Rahul Urgaonkar 和 RameshK Sitaraman。BOLA：在线视频的近乎最优码率自适应。IEEE/ACM 网络汇刊，28(4):1698–1711，2020年。

[28] Alexandros Stergiou 和 Ronald Poppe。Adapool：用于保留信息的下采样的指数自适应池化。IEEE图像处理汇刊，32:251–266，2022年。

[29] Vivienne Sze、Madhukar Budagavi 和 Gary J Sullivan。高效视频编码（HEVC）。集成电路与系统、算法和架构，39:40，2014。

[30] Yao Wang 和 Qin-Fan Zhu。视频通信中的差错控制与错误隐藏：综述。IEEE会刊，86(5):974–997，1998。

[31] Yilin Wang、Sasi Inguva 和 Balu Adsumilli。《用于视频压缩研究的 YouTube用户生成内容数据集》。载于《2019年 IEEE 第21届多媒体信号处理国际研讨会（MMSP）》，第1–5页。IEEE，2019年。

[32] Yuancheng Wang、Haoyue Zhan、Liwei Liu、Rui- hong Zeng、Haotian Guo、Jiachen Zheng、Qiang Zhang、Xueyao Zhang、Shunsi Zhang 和 Zhizheng Wu。《Maskgct：基于掩码生成式编解码器变换器的零样本文本转语音》。arXiv 预印本，arXiv:2409.00750，2024年。[33] Zhou Wang、Alan C Bovik、Hamid R Sheikh 和 Eero P Simoncelli。《图像质量评估：从误差可见性到结构相似性》。IEEE图像处理汇刊，13(4):600–612，2004年。

[34] Thomas Wiegand、Gary J Sullivan、GisleBjontegaard 和 Ajay Luthra。H.264/AVC视频编码标准概述。IEEE 电路与系统视频技术汇刊，13(7):560–576，2003年。

[35] Jiangkai Wu、Liming Liu、Yunpeng Tan、Junlin Hao 和 Xinggong Zhang。Promptus：提示词流传输能否借助 Stable Diffusion 取代视频流传输？arXiv预印本 arXiv:2405.20032，2024年。

[36] Yecheng Wu、Zhuoyang Zhang、Junyu Chen、Haotian Tang、Dacheng Li、Yunhao Fang、Ligeng Zhu、Enze Xie、Hongxu Yin、Li Yi 等。Vila-U：融合视觉理解与生成的统一基础模型。arXiv 预印本 arXiv:2409.04429，2024年。

[37] Yazhou Xing、Yang Fei、Yingqing He、Jingye Chen、Ji- axin Xie、Xiaowei Chi 和 Qifeng Chen。基于跨模态视频变分自编码器的大运动视频自编码。arXiv 预印本 arXiv:2412.17805，2024年。

[38] Zhucun Xue、Jiangning Zhang、Teng Hu、Haoyang He、Yinan Chen、Yuxuan Cai、Yabiao Wang、Chengjie Wang、Yong Liu、Xiangtai Li 等。Ultravideo：包含全面字幕的高质量超高清（UHD）视频数据集。arXiv 预印本 arXiv:2506.13691，2025年。

[39] Francis Y Yan、Hudson Ayers、Chenzhi Zhu、Sadjad Fouladi、James Hong、Keyi Zhang、Philip Levis 和 Keith Winstein。《实地学习：视频流传输中的一项随机实验》。载于第 17 届 USENIX网络系统设计与实现研讨会（NSDI 20），第 495–511 页，2020年。

[40] Xi Yang、Wangmeng Xiang、Hui Zeng 和 Lei Zhang。《真实世界视频超分辨率：基准数据集与一种基于分解的学习方案》。载于 IEEE/CVF 国际计算机视觉会议论文集，第 4781–4790 页，2021年。

[41] Zhuoyi Yang、Jiayan Teng、Wendi Zheng、Ming Ding、Shiyu Huang、Jiazheng Xu、Yuanming Yang、Wenyi Hong、Xiaohan Zhang、Guanyu Feng 等。CogVideoX：采用专家变换器的文生视频扩散模型。arXiv 预印本 arXiv:2408.06072，2024年。

[42] Hyunho Yeo、Chan Ju Chong、Youngmok Jung、Jun-cheol Ye 和 Dongsu Han。Nemo：在商用移动设备上实现神经增强视频流传输。载于第26届移动计算与网络国际年会论文集，第1–14页，2020年。

[43] Hyunho Yeo、Youngmok Jung、Jaehong Kim、Jinwoo Shin 和 Dongsu Han。神经网络自适应内容感知互联网视频传输。载于《第13届USENIX操作系统设计与实现研讨会》（OSDI 18），第645–661页，2018年。

[44] Lu Yuan、Dongdong Chen、Yi-Ling Chen、Noel Codella、Xiyang Dai、Jianfeng Gao、Houdong Hu、Xuedong Huang、Boxin Li、Chunyuan Li 等。Florence：面向计算机视觉的新型基础模型。arXiv 预印本 arXiv:2111.11432，2021年。

[45] Aiping Zhang、Zongsheng Yue、Renjing Pei、Wenqi Ren 和 Xiaochun Cao。利用扩散先验进行退化引导的一步图像超分辨率。arXiv 预印本 arXiv:2409.17058，2024年。

[46] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman和Oliver Wang。深层特征作为感知度量的超乎寻常的有效性。载于IEEE计算机视觉与模式识别会议论文集，第586–595页，2018年。

[47] Pengyuan Zhou、Lin Wang、Zhi Liu、Yanbin Hao、Pan Hui、Sasu Tarkoma和Jussi Kangasharju。关于生成式人工智能与大语言模型在视频生成、理解及流传输中应用的综述。arXiv预印本 arXiv:2404.16038，2024年。

A 附录

A.1 可伸缩比特率控制算法

算法 1 给出Morphe的比特率控制流程。ComputeResidual 与 TokenSelection 例程来自§4.3所述的可伸缩视频编码技术。

算法 1 可扩展码率控制

要求： 可用带宽 B_{avail}，视频帧F
确保： 编码比特流
如果 $B_{\text{avail}} < R_{3x}$ **则**
  $y \leftarrow$ 编码_3X(F)
  $y_{\text{filtered}} \leftarrow$ TOKENSELECTION(y, B_{avail})
  返回 y_{filtered}
否则如果 $R_{3x} \le B_{\text{avail}} < R_{2x}$ **则**
  $y \leftarrow$ 编码_3X(F)
  残差 $\leftarrow$ 计算残差($F, y, B_{\text{avail}} - R_{3x}$)
  返回 {y，残差}
否则
  $y \leftarrow$ 编码_2X(F)
  残差 $\leftarrow$ 计算残差($F, y, B_{\text{avail}} - R_{2x}$)
  返回 {y, 残差}
结束如果

A.2 Morphe训练流程

基础编解码器训练框架。Morphe 通过 §4.1中描述的内部梯度接口，对预训练的视觉基础模型（VFM）骨干网络进行微调。第一阶段训练聚焦于两项核心能力：块间时间平滑（§4.2）和自适应分辨率控制（§5）。优化目标函数定义如下：

$$L = \alpha L_{\text{pixel}}(\hat{x}, x) + (1-\alpha) L_{\text{flow}}(\hat{x}, x) \tag{5}$$

其中L^{pixel} 表示 ℓ_1 像素重建损失，L_{flow}表示光流一致性损失，而权重参数 α 设为 0.8。光流损失有助于在图像组 (GoP) 边界之间保持时间连贯性，而像素损失则确保空间细节保留。

为适应自适应视频编码，我们通过将原始 1080P 视频调整为 540P 和 720P 分辨率来扩充训练数据。训练期间，随机抽取不同分辨率的视频帧，以增强模型对多分辨率输入的适应能力。

此外，还引入了对抗性 GAN 损失 [45] 项L_{adv}，以提升视频细节的保真度。超参数$\gamma = 0.1$ 用于平衡 GAN 损失的贡献。这一对抗性目标函数促进生成视觉上合理的细节，使模型能够捕捉更精细、更自然的纹理和动态模式。

因此，整体训练损失可表示为：

$$L_{\text{total}} = L + \gamma \cdot L_{\text{adv}} \tag{6}$$

该训练方案使 Morphe 能够保持时空一致性，同时灵活支持多分辨率视频推理，显著提升生成视频细节的丰富度和真实感。

视频超分辨率模型训练。我们首先使用上一阶段训练的编解码器模型，以最低比特率对 540P 视频进行编码。这一编码和解码过程会引入一些丢包和压缩伪影。由此生成的劣化视频构成超分辨率模型的训练数据集。超分辨率（SR）模型采用残差卷积网络构建 [12]。模型首先在该数据集上进行训练，学习如何从模糊、抖动和压缩输入中复原原始视频。此阶段的丢包率采用与公式（5）相同的形式，将像素重建和光流一致性损失结合起来，以保留空间细节和时间连贯性。预训练后，将超分辨率模型的权重与解码器融合。我们冻结超分辨率模型

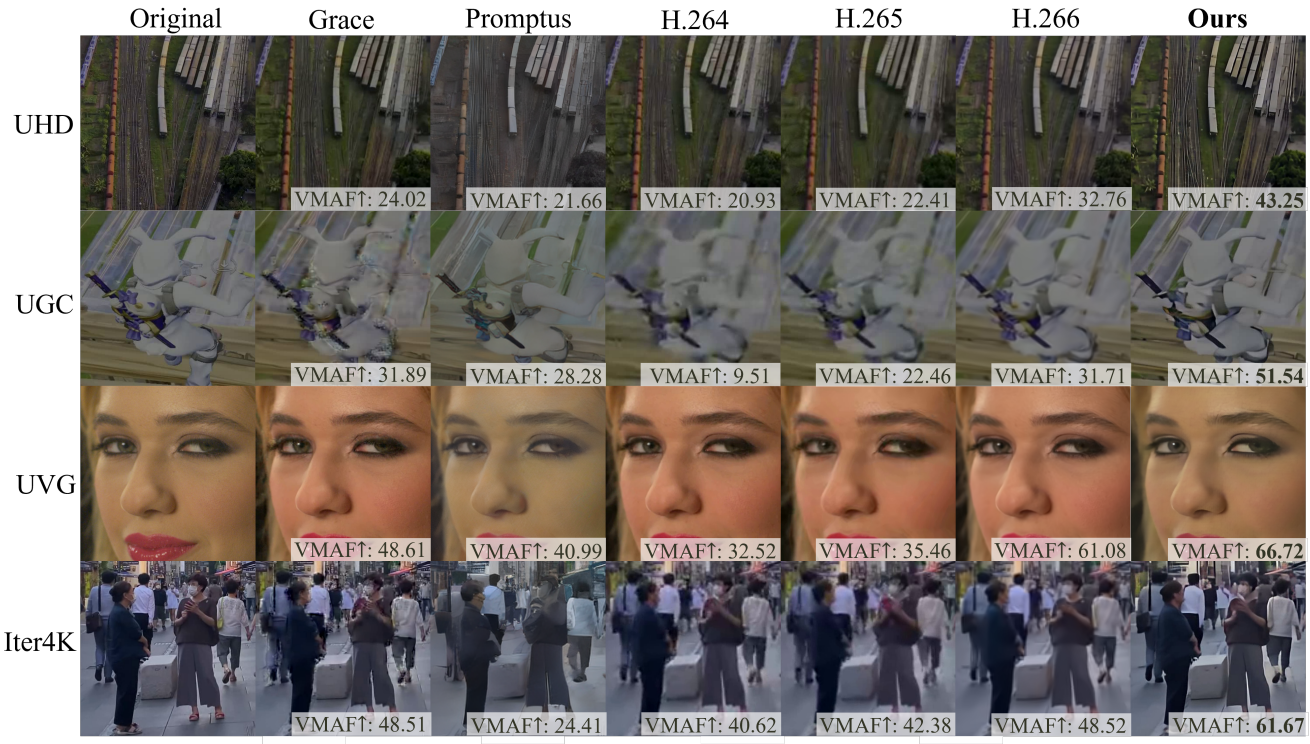


图 15：四个数据集上的视觉质量对比。

权重，仅微调编解码器权重。在这一联合训练过程中，原始 1080P 视频按随机选取的 2 倍或 3 倍比例进行线性下采样。随后将这些低分辨率帧输入编解码器模型进行推理，生成超分辨率视频。为进一步提升超分辨率性能，使用公式（6）中定义的 GAN 损失，以对应的原始高分辨率视频监督输出视频。训练过程利用 §4.1 中介绍的内部梯度接口进行梯度反向传播和迭代优化。通过冻结超分辨率模型、仅更新编解码器权重，编码器能够学习生成符合超分辨率模型预期分布的像素特征。这有助于提升超分辨率复原质量和最终的视频保真度。

通过随机丢弃进行鲁棒性训练。 在建立基本的编解码能力后，第二阶段引入令牌丢弃训练，以增强丢包韧性。

对于 P 帧语义张量 $y \in \mathbb{R}^{H\times W\times T}$，给定相似度分数 $S_{i,\,j}$ 和目标丢弃率，我们有选择地丢弃相似度较高的令牌。将被丢弃位置的值置零，形成部分表示 $\tilde{y}$。解码器 $g_\theta(\cdot)$使用$\tilde{y}$，并学习重建原始视频，采用与第一阶段相同的目标函数：

$$L_{\text{robust}} = \alpha L_{\text{pixel}}\big(g_\theta(\tilde{y}),x\big) + (1-\alpha)\,L_{\text{flow}}\big(g_\theta(\tilde{y}),x\big) \quad (7)$$

关键的是，通过内部梯度接口，梯度会反向传播至编码器，使其学会生成能够抵御令牌丢失的语义表征-

。在训练期间，我们首先以最低比特率执行推理。在此过程中，我们应用基于相似度的令牌丢弃方法 §4.3 来模拟自主丢包。为适应不同的网络丢包条件，我们从范围 $[0,25\%]$中均匀采样丢弃率。这使模型具备处理不同程度缺失信息的能力。

为什么 Morphe 具有更强的丢包韧性？与仅依赖解码器进行差错隐藏不同，联合训练使编码器和解码器协同适应令牌丢弃。首先，解码器学会利用 I 帧语义矩阵中的参考信息，推断并补全 P帧中缺失的令牌。其次，编码器学会组织语义空间，使 I 帧和 P帧共享的冗余内容彼此更接近，从而提高相似度估计的准确性。

这种协同优化在带宽限制下具有直接优势。系统可以安全地丢弃冗余的高相似度令牌，并优先保留携带新信息的低相似度令牌。由于编码器施加了更有结构的语义组织，解码器可以利用保留的 I 帧信息，更准确地恢复丢弃的 P 帧内容。这种自适应分配与协同训练相结合，使 Morphe 即使在令牌丢失率高达 30% 时，也能维持可接受的视觉质量，并在语义层和像素层实现鲁棒性。

在整个训练过程中，我们使用从 1×10^{-5} 逐渐衰减到 2×10^{-8}的学习率。这

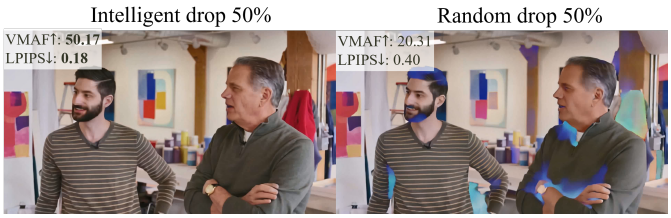


图16：我们的令牌丢弃方案与随机丢弃。

学习率调度有助于稳定训练：在早期阶段允许更大的参数更新以加快收敛，同时在后期逐步减小更新幅度，以便微调模型并避免过冲。因此，模型能够实现更好的泛化并提升性能。

A.3 视觉比较详情

视觉质量比较。图15 比较了Morphe 与四种基线方法在不同数据集上的视觉质量。我们在四个具有代表性的数据集上评估了每种方法。所有方法均在相同比特率（400千比特/秒）下进行比较，并使用 VMAF 作为感知质量指标。Morphe 在所有测试场景中均取得最高的 VMAF 分数。需要注意的是，虽然Morphe 编解码器能够对任意大小和尺寸的视频进行编码和解码，但为便于展示，我们将 1080p 视频裁剪为正方形格式。视觉比较表明，即使面对 UGC 等具有挑战性的用户生成内容，Morphe 仍保持显著优势，展现出对不同视频类型的鲁棒性。传统编解码器 H.264、H.265 和 H.266 在所有测试中的表现最差。神经编解码器 Grace 优于传统方法，但仍明显落后于我们的方法。对于 UVG 数据集中的 YUV 视频，除我们的方法外，所有方法都会造成严重色偏，损害观看体验。

传统编码器呈现出均匀的质量下降，并伴有可感知的压缩伪影和块效应。GRACE 在运动物体与背景之间出现明显的马赛克失真，而 Promptus 虽然保留了精细细节，但由于扩散模型的局限性而存在语义不一致。

A.4 消融研究

为评估我们的智能令牌丢弃机制的效果，我们在图中将其与降幅为 50% 的朴素随机丢弃进行比较16。结果显示出显著的性能差异：我们的方法取得了 50.17 的 VMAF 和 0.18 的 LPIPS，而随机丢弃的 VMAF 仅为 20.31，LPIPS 则为 0.40。视觉对比清楚地表明，随机丢弃会引入严重的损坏伪影，尤其会影响服装和背景细节等纹理丰富的区域。相比之下，我们的智能方法通过

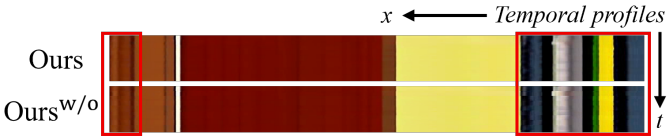


图 17：时间平滑的消融研究

优先保留语义显著令牌，实现了约 2.5× 倍的 VMAF 得分提升，并将感知失真降低了 55%。这些结果凸显了内容感知令牌选择对于满足质量要求的视频传输的重要性。

图 17 展示了在采用与不采用我们的平滑机制时时间一致性的视觉比较。时间轮廓可视化表明，未使用时间平滑的系统（我们的方法 w/o）在连续帧之间表现出显著更多的色彩变化和不连续性，轮廓带中的明显波动即证明了这一点。相比之下，我们的完整系统在各帧之间展现出明显更平滑的过渡，整个序列中的色彩模式也更为一致。帧间变化的减少直接带来闪烁伪影的减少和观看体验的提升。