arXiv:2602.17120v1 [eess.IV] 19 Feb 2026

HybridPrompt：为移动流媒体连接生成先验与传统编解码器

Liming LiuJiangkai Wu北京大学
Haoyang WangPeiheng Wang北京大学
Zongming GuoXinggong Zhang北京大学

摘要

在视频点播（VoD）场景中，传统编解码器凭借高效的解码性能，成为业界标准。然而，在低带宽条件下，它们会遭遇严重的画质下降。尽管新兴的生成式神经编解码器能显著提升感知质量，但其依赖繁重的逐帧生成，使其难以在移动设备上实现实时播放。我们提出一个问题：能否将传统标准的极快速度与神经网络方法的卓越视觉保真度结合起来？我们提出*HybridPrompt*，这是首个能够在商用智能手机上以超过150帧每秒的速度实现实时1080p解码的生成式视频编码系统。具体而言，我们采用混合架构，使用生成模型对关键帧进行编码，同时依靠传统编解码器处理其余帧。一项主要挑战在于，这两种范式的目标彼此冲突：生成模型"幻觉"出的细节往往与传统编解码器严格的预测机制不匹配，导致码率效率低下。为解决这一问题，我们证明传统解码过程是可微分的，从而能够进行端到端优化。这样一来，我们便可将后续帧用作额外监督，促使生成模型合成关键帧，使其不仅具有高感知保真度，也能成为传统编解码器在数学意义上的最优参考帧。通过整合两阶段生成策略，我们的系统在速度上比纯神经网络基线快出数个数量级，同时在200千比特每秒下，相较传统编解码器实现了LPIPS平均提升8%。

CCS概念

• **信息系统** → **多媒体流媒体**。

本作品采用知识共享署名 4.0 国际许可协议授权。*NOSSDAV '26*，香港，香港© 2026 版权归所有者/作者所有。ACM ISBN 979-8-4007-2534-0/26/04
https://doi.org/10.1145/3798065.3798083

关键词

混合视频编码、生成先验、移动流媒体、实时解码、码率套利、可微编解码器

ACM参考文献格式：

Liming Liu、Jiangkai Wu、Haoyang Wang、Peiheng Wang、Zong- ming Guo 和 Xinggong Zhang。2026。HybridPrompt：融合生成先验与传统编解码器的移动流媒体方案。载于数字音频与视频网络及操作系统支持研讨会（*NOSSDAV '26*），*2026*年*4*月*4*日至*8*日，香港，香港。ACM，纽约，NY，美国，7页。
https://doi.org/10.1145/ 3798065.3798083

1 引言

视频点播（VoD）服务占据了全球网络流量的主导地位 [12], ，然而，在不稳定的网络环境中为移动用户提供高质量视频流仍是一项根本性挑战。在带宽波动的场景中，例如拥挤的地铁或农村地区，传统视频编码标准（如 H.264 [1], H.265 [2], 和 VP9 [10]）往往无法维持可接受的视觉质量。这些标准通过将视频划分为独立的参考图像（称为**关键帧（I帧）**）和依赖帧（称为**预测帧（P帧/B帧）**）来工作，后者只记录变化。在严格的码率限制下，编解码器不得不大幅压缩 I 帧以节省数据。由于 P 帧和 B 帧完全依赖 I 帧进行重建，质量低、模糊的 I 帧会导致误差在整个视频片段中传播，从而造成糟糕的观看体验 [5]。

生成式语义模型提供了一种潜在解决方案，即从紧凑描述重建高保真度图像，但它们目前缺乏移动端播放所需的解码速度。Promptus [14] 等方法的核心机制依赖于"基于梯度的反演"：一种离线编码过程，通过迭代优化一段极小的类文本编码（提示词），以驱动模型（例如 Stable Diffusion [3]）合成图像。在编码某一帧时，它会调整这一小段类提示词编码，直到生成输出与源视频匹配。由于只需传输紧凑编码，这可以实现极低的带宽使用量。然而，代价在于解码的计算成本：现有生成式视频方案通常需要为每一帧运行这一繁重的生成过程。在算力有限的移动设备上，这种方法

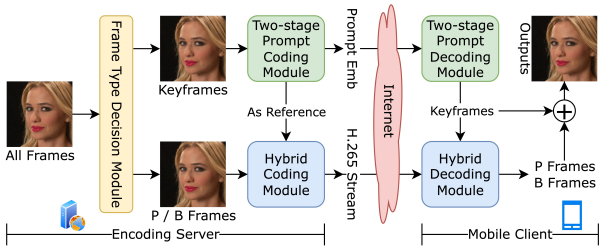


图1：HybridPrompt概览。

慢得令人难以接受，往往需要数秒才能生成仅仅一帧，而流畅视频至少需要 30 帧每秒。

能否将生成模型卓越的感知质量与传统编解码器极快的解码速度结合起来？我们提出*HybridPrompt*，一种旨在兼得两者优势的新型混合视频编码策略。我们的核心洞见是，仅对 I帧使用计算量大的生成模型，而对基于运动的 P帧和 B帧使用轻量级传统编解码器。这形成了一种"码率套利"：由于生成模型能够将 I帧压缩到几乎可忽略的大小，同时保持较高的视觉质量，因此我们节省了大量数据。随后，我们将节省下来的码率预算重新分配给传统编码的 P帧和 B帧。因此，P帧和 B帧能获得足够的数据以保持清晰度；而且由于它们由传统硬件编解码器解码，整体播放依然流畅且实时。

然而，简单地将这两种技术拼接起来会产生冲突：AI"幻觉"出的细节会干扰传统编解码器。生成模型旨在让图像在人眼看来更美观（感知质量），因此常会凭空生成逼真的纹理，例如草地或头发，但这些纹理可能无法与原始视频实现像素匹配。然而，H.265 等传统编解码器在数学上十分严格：它们需要精确的像素对齐来进行预测。如果生成模型生成的纹理略有偏移，传统编解码器就会将其视为巨大误差，并浪费数据试图"修正"它，从而抵消比特率节省。为解决这一问题，我们利用传统解码过程本质上由一系列可微操作构成这一特点，从而将迭代反演扩展为端到端优化循环。我们不再仅针对 I 帧重建来优化潜在编码，而是引入额外监督，迫使生成的 I 帧与传统编解码器的运动矢量和残差对齐。这样可以确保所有帧都具有较高的感知质量。

我们在一款商用智能手机上实现了 *HybridPrompt*，并证明它是首个实现实时1080p解码的生成式系统。我们的贡献主要有三点：首先，我们提出了一种通用的混合架构，利用生成式 I 帧为传统



(a) 真值　(b) 1200 千比特每秒　(c) 600 千比特每秒　(d) 200 千比特每秒

图 2：我们以不同的比特率对一段 1080p@30fps 视频进行压缩（H.265），图像组长度为 120 帧。可以观察到，比特率较低时，块效应和模糊会更加明显，导致画质显著下降。

表 1：在若干移动设备上使用 Stable Diffusion 生成单张 1080p 图像所需的平均时间。目前的硬件仍远不能实现实时逐帧生成。

设备	每帧平均时间（秒）
iPhone 12 Pro Max	9.19
iPhone 13 Pro Max	8.69
iPhone 14 Pro Max	4.11
iPhone 16 Pro Max	2.61

P 帧的码率预算提供补偿。其次，我们引入了一种可微桥接机制，使神经生成与传统预测这两个不同领域能够进行端到端优化，从而解决像素错位问题。最后，我们集成了两阶段生成策略，使其在 iPhone 16 Pro Max 上的解码速度超过 150 帧每秒，比纯神经网络基线快几个数量级。同时，在 200 千比特每秒下，与传统编解码器相比，我们的 LPIPS 平均提升了 8%。

2 动机与关键洞见

2.1 困境：移动视频点播中的质量与效率

移动设备上的视频点播（VoD）面临带宽限制与计算能力之间的根本冲突。H.264 [1],H.265 [2], VP8 [4], 和 VP 9 [10] 等传统算法计算开销较低，但在低带宽下难以保持质量。如图 2 所示，当比特率降低（例如降至 200 千比特每秒）时，I帧的质量会急剧下降，并出现严重的块效应。由于 I帧是后续 P帧和 B帧的参考帧，这种质量下降会在整个图像组（GOP）中传播，形成预测编码范式固有的质量瓶颈。

相反，新兴的神经编解码器 [3, 14] 利用生成先验，可提供更出色的感知质量。然而，它们的计算开销过大，难以在移动设备上运行。如表 1 所示，即使在 i Phone

表2：不同目标比特率下神经编码与传统编码的I帧感知质量（LPIPS ↓，越低越好）和帧大小（千字节）对比。与传统帧内帧相比，神经关键帧始终能实现更优的质量与大小权衡。

比特率（千比特每秒）	200	400	600	1200
传统（H.265）				
LPIPS ↓	0.543	0.510	0.494	0.454
大小（千字节）	7.4	11.2	14.8	27.7
神经网络（Promptus）				
LPIPS ↓	0.463		0.552	
大小（千字节）	8.8		4.4	

16 Pro Max 上，使用 Stable Diffusion 生成一帧 1080p 画面也需要 2 秒以上，比实时播放（30 帧每秒）所要求的 33 毫秒慢了好几个数量级。这就造成了一个两难局面：我们既需要神经网络模型的视觉质量，也需要传统编解码器的解码速度。

2.2 混合架构的可行性

我们的第一个洞见是，通过混合架构可以实现“鱼和熊掌兼得”。核心观察在于，神经编解码器对I帧具有非对称的高效率。如表2所示，与传统编解码器（H.265）相比，神经生成的I帧（Promptus）以更小的文件大小（8.8千字节对11.2千字节）实现了显著更高的感知质量（LPIPS 0.463对0.51）。

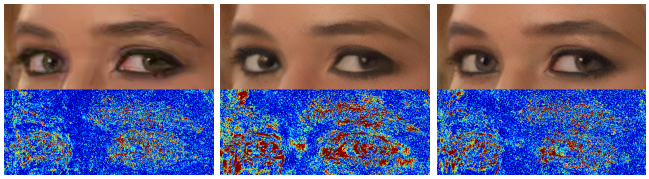
这表明了一种“码率套利”策略：仅使用神经编解码器对I帧进行编码，以节省大量比特率，然后将这部分预算重新分配给后续由传统编解码器编码的P帧（其计算成本较低）。从理论上讲，这应能提升整个视频流的总体质量。

2.3 挑战：感知-精度失配

然而，将神经帧与传统帧进行简单组合在实践中并不可行。这是由于**感知真实感**与**逐像素精度**之间存在根本性冲突。

生成式神经编解码器旨在最大化感知质量，重建在肉眼看来逼真的高频细节（例如草地纹理、发丝）。尽管这些细节在视觉上质量很高，但它们可能并不严格符合真实值的像素网格。相比之下，传统编解码器（如H.265）会严格针对均方误差（MSE）等客观指标进行优化。它们要求严格的**逐像素精度**，并将逼真纹理即使轻微的空间错位也视为严重错误。

我们在图3中展示了这一现象，其中比较了传统I帧与未经端到端优化、由模型直接生成的“朴素神经”I帧，两者



（a）传统编解码器（H.265） （b）朴素神经（c）我们的方法（优化后） 图3：感知-精度失配。三个 I 帧均压缩至相同大小 8.8 千字节。每个子图展示 RGB 帧（上）及其对应的均方误差热力图（下）。

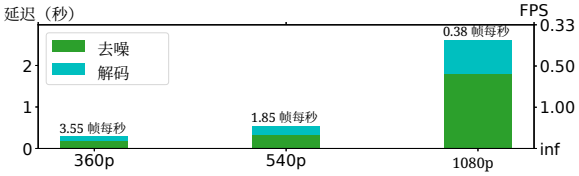


图4：不同分辨率下基于扩散的生成运行时间分析，测试设备为 iPhone 16 Pro Max。较低分辨率对应显著降低的延迟。

均压缩至相同的8.8千字节。传统I帧（a）画面模糊，但在数学上“精确”（均方误差较低，以蓝色表示）。朴素神经I帧（b）视觉上清晰逼真，但在数学上“未对齐”（均方误差较高，以红色表示）。当神经生成帧作为参考帧时，传统编解码器会将这些视觉上合理的细节视为显著误差。因此，它不得不耗费过多比特率来尝试“修正”像素级不一致。这种适得其反的修正会导致**双重劣化**：它消耗了比特率节省，同时破坏了清晰的神经纹理（降低感知质量），且未能完全恢复像素对齐（导致均方误差增大）。

2.4 洞见：通过可微端到端优化弥合差距

为解决这一冲突，我们需要让生成模型适应传统编解码器。我们的关键洞见是，整个传统解码序列由可微分的数学运算组成。

具体而言，获得运动矢量和残差后，重建过程包含三个不同的步骤。首先，根据运动矢量对参考帧进行**扭曲**，这相当于一种可微采样操作。其次，将解码后的**残差相加**到扭曲后的帧上，这是一种可微加法操作。第三，帧会经过**去块滤波器**，这些滤波器可以建模为可微卷积。

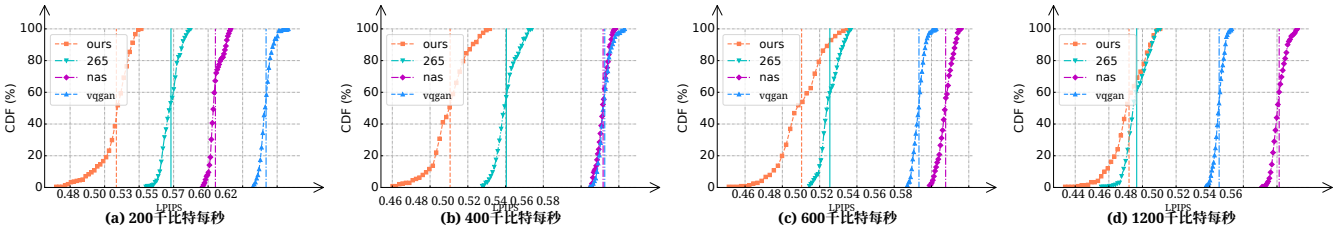


图5：在1080p分辨率下，不同比特率设置的感知质量（LPIPS，越低越好）对比。HybridPrompt在所有带宽条件下始终取得最低的LPIPS（最佳质量）。

由于这一链条中的每个步骤都是可微分的，我们可以构建一个**端到端优化循环**。我们不再只针对 I 帧自身的视觉质量进行优化，而是优化其潜在表示，以尽量降低所有后续 P 帧和 B 帧的重建误差。通过将这些帧的损失经由扭曲、相加和滤波层反向传播至 I 帧的潜在表示，我们促使生成模型生成的细节不仅在视觉上逼真，而且**有利于传统编解码器**。

如图3(c)所示，我们优化后的I帧保留了神经模型的清晰纹理。关键在于，与朴素神经方法相比，它显著抑制了像素级误差。虽然均方误差并未严格超越H.265，但这种改进后的对齐避免了传统编解码器将比特率浪费在不必要的校正上，从而恢复了混合流的效率。

2.5 速度优化：解耦分辨率以实现实时解码

在质量和兼容性问题得到解决后，最后的障碍是速度。我们在iPhone 16 Pro Max上进行了实验，以测量扩散模型在不同分辨率下的生成延迟。如图4所示，我们的结果表明，推理时间大致与像素总数成正比，这意味着较低的分辨率能显著加快生成速度。

基于这一观察，我们采用**两阶段生成**策略以实现实时性能。我们并非直接生成完整的1080p帧，而是先生成540p低分辨率图像。随后，该图像通过轻量级超分辨率模块上采样至1080p。关键在于，由于该上采样操作完全可微，我们将其纳入端到端训练流程。这种联合优化最大限度地减少了通常与超分辨率相关的质量下降，确保最终的高分辨率输出保持高保真度。

3 系统设计

3.1 架构概述

如图 1 所示，整个生命周期如下：

服务端转码。 服务器以原始视频为输入，并将其分解为多个图像组（GOP）。对于每个图像组，第一帧会通过我们的神经网络流水线反演为紧凑的生成式潜在编码。后续帧则使用标准的传统 H.265 编码器进行编码，并以神经网络生成的 I 帧作为参考帧。

码流复用。 我们设计了一种自定义传输格式。我们不传输体积庞大的纯 H.265 码流，而是丢弃 I 帧，并将紧凑的潜在编码注入容器。最终的混合码流由两条交错的轨道组成：包含 I 帧的神经流，以及包含 P-/B 帧运动矢量和残差的传统流。

客户端双重解码。客户端收到流后，会对数据进行解复用。神经引擎对潜在编码进行解码，以重建 I 帧；随后，这些帧被放入播放缓冲区，作为硬件解码器重建 P/B 帧时的参考帧。

3.2 服务器端：转码流水线

如图 6 所示，我们方法的核心贡献是服务器端优化策略，它确保生成式 I 帧不仅在视觉上令人满意，而且在数学上与 H.265 预测逻辑保持一致。该过程包含以下三个连续步骤：

步骤 I：生成式反演。首先，对于每个原始 I 帧I_{gt}，我们旨在找到能够重建该帧的潜在编码 z。为确保移动客户端的实时性能，我们采用两阶段生成策略。编码 z 首先驱动扩散模型合成一张低分辨率（540p）图像，然后通过可微分的超分辨率模块将其上采样至 1080p。我们通过迭代优化 z，尽可能减小生成帧 I_{gen} 与I_{gt}之间的感知差异，从而为图像组初始化视觉“锚点”。

步骤 II：混合运动编码与码率控制。生成 $\hat{I}_{gen}$ 后，我们使用传统H.265编码器对图像组中的其余帧进行编码。一个关键挑战是确保编解码器使用 I_{gen} 的确切结果进行运动估计。为此，我们采用参考帧注入与$\hat{}$量化参数=0策略，将 I_{gen} 输入 H.265 编码器，并强制将量化参数（QP）设为

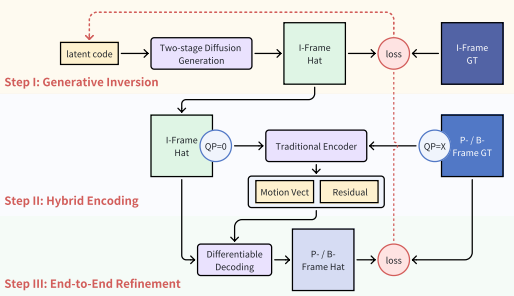


图 6：混合转码流水线概览。

0（无损模式），用于 I帧。这样可以确保编码器根据我们的生成结果精确计算运动矢量（MVs）和残差，而不会引入额外的压缩噪声。请注意，生成的超大 H.265 I帧比特流会被丢弃，仅保留 P/B 帧数据。为管理带宽，我们同时采用二分搜索分配，动态调整 P/B 帧的量化参数，直至总比特率（代码大小 z + P/B 流大小）达到目标带宽预算。

步骤 III：端到端联合优化。 步骤 I 得到的初始编码 z 仅针对 I帧进行优化，因此可能无法为图像组中的其余帧提供理想的预测。在最后阶段，我们固定步骤 II 得到的运动矢量和残差，并进行端到端微调。我们构建了一条可微解码链，通过使用固定运动矢量进行扭曲 I_{gen}，重建预测帧^。我们计算图像组所有 帧的重建损失，并反向传播梯度以更新潜在编码 z。这一过程迫使生成式 I帧与视频的运动轨迹对齐，确保"幻觉"细节与编解码器的预测逻辑相兼容。

我们确实注意到，解码链中包含一些不平滑或会削弱梯度的算子（例如截断、舍入或类似阈值的步骤）。但它们并不会在所有位置完全阻断梯度。如图 3(c) 所示，流经大多数可微分组件的梯度足以支持端到端精细化，并使 I_{gen}与 H.265 预测逻辑相兼容。

3.3 客户端：并行解码

为了在真实手机上实现实时播放，我们实现了一条并行解码流水线。该过程以异步方式进行：视频处理器（VPU）解码图像组 N 中的 P 帧和 B帧时，神经处理单元（NPU）则在后台为图像组 $N+1$ 生成 I帧。

即时流拼接。 为确保在无需驱动级修改的情况下与商用硬件解码器无缝集成，我们引入了一个轻量级重编码步骤。一旦神经网络处理器生成I帧，我们

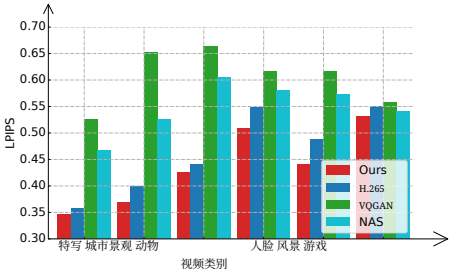


图 7：不同视频类别在 UVG数据集上的 LPIPS 对比（越低越好）。

立即以**量化参数=0**（无损）执行软件编码，并将该本地I帧头与接收到的P/B帧比特流拼接。这种即时拼接构建出符合标准的H.265码流，使视频处理器能够将混合流视为普通视频文件。我们的实验表明，这一中间重编码极为轻量，开销可忽略不计，并严格保持实时解码速度。

4 评估

我们设计评估来回答三个关键问题：（1）**质量与泛化能力：**在不同带宽和多样化内容场景下，混合架构能否始终优于基线方法？（2）**机制：**"码率套利"策略如何有效地重新分配带宽，从而提升整体保真度？（3）**效率：**系统能否在移动硬件上实现实时性能？各架构组件对速度-质量权衡有何贡献?

4.1 实验设置

测试平台。 服务端转码在一台配备 NVIDIA RTX 4090 GPU（24 GB 显存）的工作站上进行。在客户端，我们将解码流水线部署在 iPhone 16 Pro Max（A18 Pro 芯片）上。我们利用神经引擎（神经网络处理器）进行生成式推理，并使用硬件视频处理单元（VPU）进行 H.265 解码。**数据集与指标。**我们使用**UVG数据集** [9],，将其标准化为 1080p、30 帧每秒。该数据集涵盖多种类别，包括城市场景、人脸和*3D*游戏。为评估感知质量，我们报告 LPIPS [17]（越低越好），因为对于生成式内容而言，它比峰值信噪比更能贴合人类视觉。我们还测量每秒帧数（帧每秒）以量化吞吐量。

基线。 我们将本方法与三类不同的方法进行比较。首先，为代表**传统编解码器**，我们评估 H.265 [2],这一当前的行业标准。其次，对于**神经增强流媒体**（超分辨率），我们评估 NAS [16],，它近似于 NEMO [15],的移动端实时版本，作为对比基准

内容感知超分辨率方法 [11, 18]。第三，为代表**轻量级生成式神经编解码器**，我们选择 VQGAN [6, 8]，这是一种通过变分自编码器 [7]量化潜变量 [13] 的基于令牌的方法。

表 3：P/B帧大小对比。与标准 H.265 相比，我们的方法为 P/B帧分配了显著更多的带宽，从而实现了更好的质量。

目标比特率	我们的方法（混合）		标准 H.265	
	LPIPS ↓	帧大小	LPIPS ↓	帧大小
200 千比特每秒 (P)	**0.475**	**2.2 千字节**	0.549	1.6 千字节
200 千比特每秒 (B)	**0.474**	**0.4 千字节**	0.536	0.25 千字节
600 千比特每秒 (P)	**0.451**	**8.6 千字节**	0.494	4.4 千字节
600 千比特每秒 (B)	**0.450**	**1.7 千字节**	0.494	0.9 千字节
1200 千比特每秒 (P)	**0.462**	**19.9 千字节**	0.467	9.6 千字节
1200 千比特每秒 (B)	**0.440**	**4.1 千字节**	0.460	1.7 千字节

4.2 感知质量分析

整体质量与泛化能力。 我们评估了我们的混合架构能否在不同网络条件和内容类型下提供更优的感知质量。图 5 展示了四个比特率级别下所有帧的 LPIPS 分数累积分布函数（CDF）。与 H.265、NAS 和 VQGAN 相比，我们的方法始终实现了最佳的质量分布。在 200 千比特每秒下，这一优势尤为明显，因为传统编解码器会出现严重的质量下降。此外，我们还通过图 7 验证了方法在多种场景下的鲁棒性，包括特写镜头、城市场景和高动态游戏。无论场景复杂度如何，我们的方法都始终优于基线，展现出广泛的适用性。

机制："码率套利"效应。尽管使用的是相同的 H.265 后端，我们的 P/B 帧仍具备更高的质量，这可以通过资源重新分配来解释。如表 3 所分析，我们的生成式 I 帧表示极为紧凑，使我们能够从 I 帧"窃取"带宽，并将其重新分配给 P 帧和 B 帧。例如，在目标比特率为 600 千比特每秒时，我们的方法为 P 帧分配的平均大小为 **8.6 千字节**，几乎是标准 H.265 所分配的 **4.4 千字节**的两倍。这种"码率套利"确保运动矢量和残差能够以显著更高的保真度进行编码，从而带来更清晰的整体视频流。

4.3 系统效率与消融研究

为量化各系统组件的贡献，我们在 iPhone 16 Pro Max 上进行了消融研究，结果汇总于表 4。

实时性能。完整的 HybridPrompt 系统实现了高达 **153.7 帧每秒**的解码速度。这证实了我们的并行化架构能够有效消除生成式瓶颈：该架构将 I 帧卸载至神经网络处理器，并由视频处理器处理 P帧。

组件分析。我们进一步分析了特定模块的影响。首先，**两阶段生成**策略对于效率至关重要；将其替换为直接生成高分辨率内容会使吞吐量降至 41.8 帧每秒，并显著增加功耗。其次，**联合优化**对保真度至关重要；禁用端到端细化会使 LPIPS 从 0.458 降至 0.463。这证实，仅仅将生成器与编解码器"拼接"起来是不够的，还必须确保梯度流经编解码器，才能严格地使生成先验与视频流对齐。最后，尽管对每一帧进行纯神经编码能带来更优的质量，但其延迟过高，因此完全不适用于移动端部署。

表 4：iPhone 16 Pro Max 上的消融研究。

配置	LPIPS（↓）	帧每秒（↑）
完整 HybridPrompt（我们的方法）	**0.458**	**153.7**
无联合帧+比特率优化。	0.463	153.7
无两阶段生成	0.454	41.8
无混合策略（仅提示词）	0.384	0.38

5 结论与未来工作

在本研究中，我们提出了 HybridPrompt，这是一种适用于移动设备的视频编码框架，利用生成先验以较低比特率实现出色的感知质量。但仍有若干方向值得进一步探索：

自适应比特率分配。 我们的系统目前采用启发式策略，在神经流与传统流之间分配带宽。未来的工作将探索更出色的码率控制方法，根据实时信道波动和场景复杂度，动态优化语义 I 帧与纹理丰富的 P 帧之间的权衡。

下一代混合编码。 尽管我们已证明该混合范式在 H.265 上的有效性，但它与编解码器无关。未来版本将集成更先进的传统编解码器（如 AV1 或 VVC），以及更强大的扩散模型，进一步提升压缩效率。此外，我们还计划探索量化技术，以进一步降低移动端神经处理器的能耗。

本工作获得国家自然科学基金项目（62431017）和字节跳动项目（CT20241126107484）的资助。我们衷心感谢媒体融合生产技术与系统国家重点实验室、智能媒体技术重点实验室提供的支持。通讯作者为 Xinggong Zhang（zhangxg@pku.edu.cn）。

参考文献

[1] 2024。H.264。https://www.itu.int/rec/T-REC-H.264。[2] 2024。H.265。https://www.itu.int/rec/T-REC-H.265。[3] 2024。Stable Diffusion。https://stability.ai/。[4] Jim Bankoski、Paul Wilkins 和 Yaowu Xu。2011 年。VP8（一种面向 Web 的开源视频编解码器）技术概述。载于 *2011 IEEE*国际多媒体与博览会会议。IEEE，1–6。[5] Joshua P Ebenezer、Yixu Chen、Yongjun Wu、Hai Wei 和 Sriram Sethuraman。2022 年。低比特率下高速运动体育视频的主观和客观质量评估。载于 *2022 IEEE*图像处理国际会议（*ICIP*）。IEEE，521–525。[6] Patrick Esser、Robin Rombach 和 Bjorn Ommer。2021 年。驾驭变换器进行高分辨率图像合成。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集。12873–12883。[7] Diederik P Kingma 和 Max Welling。2014 年。自编码变分贝叶斯。机器学习研究期刊（*JMLR*）15，1（2014），1929–1958。[8] Tianhong Li、Vibhaalakshmi Sivaraman、Lijie Fan、Mohammad Alizadeh 和 Dina Katabi。2023 年。Reparo：用于视频会议的抗丢失生成式编解码器。*arXiv*预印本 *arXiv:2305.14135*（2023）。[9] Alexandre Mercat、Marko Viitanen 和 Jarno Vanne。2020 年。UVG数据集：用于视频编解码器分析和开发的 50/120帧每秒 4K 序列。载于 第*11*届*ACM*多媒体系统会议论文集。297–302。[10] Debargha Mukherjee、Jingning Han、Jim Bankoski、Ronald Bultje、Adrian Grange、John Koleszar、Paul Wilkins 和 Yaowu Xu。2015 年。VP9——最新开源视频编解码器的技术概述。*SMPTE*运动成像期刊 124，1（2015），44–54。[11] Seonghoon Park、Yeonwoo Cho、Hyungchol Jun、Jeho Lee 和 Hojung Cha。2023。Omnilive：面向移动设备的超分辨率增强型 360 度视频流媒体直播。在 第 *21* 届年度国际移动系统、应用与服务会议论文集 中。261–274。 [12] Sandvine。2024。2024年全球互联网现象报告。https: // www.sandvine.com/global-internet-phenomena-report-2024。[13] Aaron Van Den Oord、Oriol Vinyals 等。2017。神经离散表示学习。神经信息处理系统进展 30（2017）。[14] Jiangkai Wu、Liming Liu、Yunpeng Tan、Junlin Hao、Liang Zhang 和 Xinggong Zhang。2026。Promptus：提示词流媒体能否取代视频流媒体。在 *AAAI*人工智能大会论文集 中。[15] Hyunho Yeo、Chan Ju Chong、Youngmok Jung、Juncheol Ye 和 Dongsu Han。2020。NEMO：在普通移动设备上实现神经增强视频流媒体。在 第 *26* 届年度国际移动计算与联网会议论文集 中。1–14。[16] Hyunho Yeo、Youngmok Jung、Jaehong Kim、Jinwoo Shin 和 Dongsu Han。2018。神经自适应内容感知互联网视频传输。在第 *13* 届 *USENIX* 操作系统设计与实现研讨会（*OSDI 18*） 中。645–661。[17] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。2018。深层特征作为感知度量的非凡有效性。在 *IEEE*计算机视觉与模式识别会议论文集 中。586–595。[18] Qihua Zhou、Ruibin Li、Song Guo、Peiran Dong、Yi Liu、Jingcai Guo 和 Zhenda Xu。2022。CADM：面向神经增强视频流媒体的编解码器感知扩散建模。*arXiv* 预印本 *arXiv:2211.08428* （2022）。