
当视频只剩下一串关键词：AI生成技术如何重塑视频通信？

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/42665.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

当视频只剩下一串关键词：AI生成技术如何重塑视频通信？。上海交通大学张文军教授团队联合帝国理工学院李焱教授在中国工程院院刊Engineering发表题为Generative Video Communications: Concepts, Key Technologies, and Future Research Trends（生成式视频通信：概念、关键技术与未来研究趋势）的研究性文章。该研究系统提出了生成式视频通信这一全新范式，从编码、传输、评估三个技术维度深入探讨了如何将生成式AI与传统视频通信深度融合，为突破传统视频通信在压缩效率、能耗与服务形态上的瓶颈提供了系统性的技术路径。

视频通信优化了三十年，还能走多远？

从3G时代到如今的5G，视频通信已是我们生活里最寻常的事——刷视频、开视频会议、看直播，几乎无时无刻不在发生。国际电信联盟（ITU）更是把沉浸式视频列为未来6G网络的关键应用场景之一。

但传统视频通信的基本范式几十年没怎么变过：把视频拍下来、压缩编码、传输、再解码显示出来，一路追求的是信息域和物理域的极致优化——压缩算法越来越复杂，传输技术不断精细化。

问题是，这条路正在走向边际效应递减。以视频编码为例，从H.264演进到最新的H.266/VVC，算法复杂度大幅增长，但每一代标准带来的相对压缩效率增益在逐步收窄。与此同时，基站功耗却在飙升，5G基站的耗电量比前几代明显更高，6G时代这个趋势还会继续恶化。此外，现有的视频服务大多仍停留在传统的二维、被动式收看，很难满足人们对个性化体验越来越高的期待。

问题的根源：对视频内容的理解过于简化

研究团队认为，问题的根源在于：现有系统对视频内容的处理过于简化——将一段视频仅仅理解为一组离散的像素点，通信的目标也就变成了尽可能保真地还原每一个像素值，而忽略了像素背后所承载的语义信息。这种思路对于光纤这类追求数据绝对保真的传输场景或许够用，但对视频通信来说并不完全适用——因为视频通信的根本目标不是传输像素本身，而是服务于人的认知过程。

这里有一个值得关注的神经科学发现：人脑处理信息的能力，大约只有每秒10比特，而我们感官系统接收外界信息的速率，却高达每秒 10^7 比特量级，两者相差近一亿倍。这说明人类大脑其实一直在做一件事：从海量的感官信息里，高效地提炼出真正有用的那一小部分。可传统视频通信

恰恰忽视了这个提炼的认知阶段，一直在传输大量冗余的感官数据。

而这，正是生成式AI能够发挥关键作用的方向。变分自编码器（VAE）、生成对抗网络（GAN）、扩散模型——这些近年来快速发展的生成技术，已经在图像、视频等多媒体内容生成上展现出强大的能力。它们的核心能力在于学习并建模数据背后的统计规律，而非逐像素地记忆原始数据。这意味着视频通信有可能从被动传输转向主动生成，将认知域这一长期被低估的维度真正利用起来。

生成式视频通信：判别与生成的深度融合

基于这样的洞察，研究团队正式提出了生成式视频通信这一新范式。

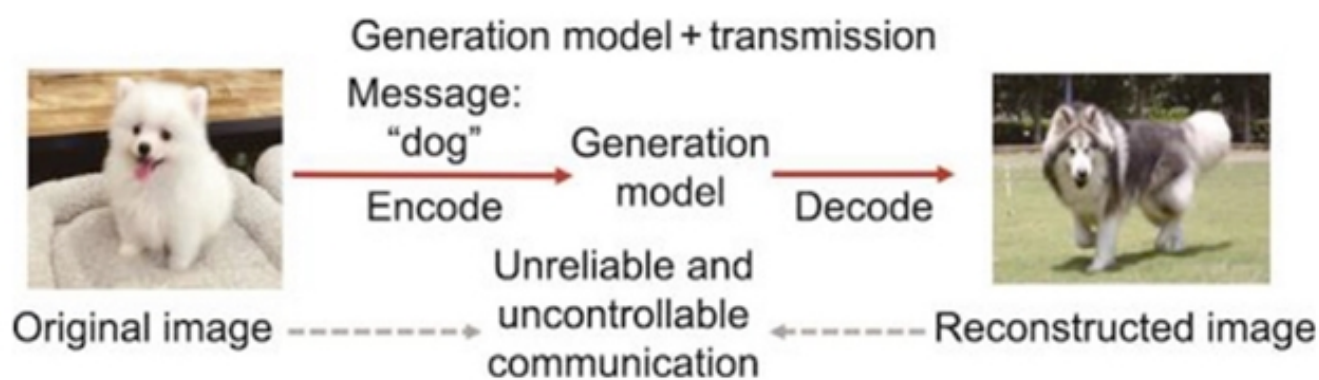


图1.生成模型与通信的基本融合示意

要理解这个概念，先看一张图（图1）：假如接收端只拿到几个关键词，比如狗，AI生成模型就能生成出一张狗的图片来——传输量小到极致，但生成的图片和原图可能相差很大，毕竟生成模型并不知道原始图像的具体内容。这正是生成式AI用于通信的核心矛盾：传输效率极高，但保真度难以保证。

因此，研究团队给出了一个关键定义：生成式视频通信，不是用生成式AI完全取代传统的判别式视频通信，而是让两者深度融合，在保证一定可靠性的前提下，尽可能发挥生成式AI的表达力优势。这个新范式有三个核心特征：

判别技术打底：传统的判别式技术依然是基础，负责保证兼容性、质量控制和可靠性，是系统稳定运行的基础。

生成技术提升表现力：生成式AI通过学习视频内容的内在规律，大幅拓展了视频通信的表达力，有望降低传输成本，让通信范式本身发生进化。

渐进式革新：新范式充分考虑计算能力、生成质量和带宽约束这些现实条件，是在传统可靠性基础上的一次渐进式革新，而非推倒重来。

更进一步，研究团队从信息论角度，把生成式视频通信的能力划分成了三个递进的层次：语法层

面的重建能力（能不能把像素级别的原始画面精确复原）、语义层面的效用能力（生成的内容能不能在下游任务里达到和原视频相似的效果）、语用层面的理解能力（生成的视频能不能在人类认知层面产生和原视频一样的感受）——这三者由具体到抽象层层递进，代表了生成式视频通信不断逼近人类认知本质的过程。

根据判别信息和生成信息的比例不同，研究团队又进一步把生成式视频通信分成了三档，并给出了一套基于比特率（bpp，每像素比特数）的选择参考标准：

判别信息主导（全参考生成）：适用于比特率高于0.05bpp的场景，比如对输入视频做语义增强，兼顾像素级和语义级的保真度；

判别与生成平衡（半参考生成）：适用于0.005—0.05bpp的中等比特率区间，典型应用是把2D视频扩展成3D视频；

生成信息主导（无参考生成）：适用于低于0.005bpp的极限低比特率场景，比如把语言或脑电信号直接转化成视频信号，只依靠极少的控制信号（如文字）驱动生成。

三条技术路径：弹性编码、协同传输与可信评估

围绕这套新范式，研究团队梳理出了三条核心技术路径（图2）：弹性编码、协同传输、可信评估。

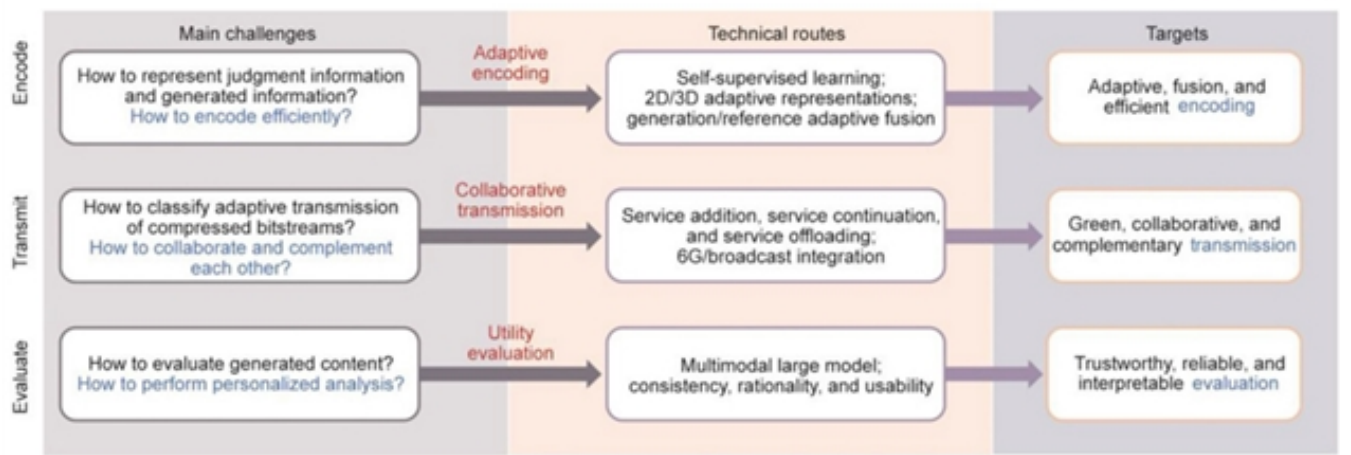


图2.生成式视频通信技术框架

第一条——弹性编码：判别信息与生成信息的灵活配比。

全参考生成：适合视频语义增强场景。研究团队设计了一套框架——先用传统视频编解码器（如H.265）压缩视频，再引入一路基于神经网络的语义信息流，在接收端通过跨码流特征融合，重建出兼具画质与语义精度的视频，能有效支撑目标跟踪、识别等下游机器视觉任务。

半参考生成：主要用于打造更沉浸的3D场景，比如把单视角视频扩展为多视角渲染，特别适合VR等需要高自由度视角切换的应用。研究团队提出的弹性3D联合编码方案，结合了显式表示（传统2D编解码）和隐式神经表示（INR）——用一路编码器保证参考信息可靠传输，另一路用INR

捕捉多视角之间的潜在关联，再通过加权融合策略输出高质量重建画面。

无参考生成：这是最具颠覆性的一档，核心思路是视觉信号→文本信号→视觉信号，只用极少的控制信号就能重建出高质量的视频内容，在极低带宽场景下表现尤其亮眼。研究团队展示的一套基于多模态大模型的生成编码方案里，先用图生文模型提取输入图像的语义信息，再用文生图扩散模型生成对应画面，同时引入高度压缩的原图信息作为控制信号，保证生成内容和原图的空间一致性。这意味着，即使带宽被压缩到几乎不可能传输画面本身，只靠一段文字描述，也能让接收端重建出一段视频。

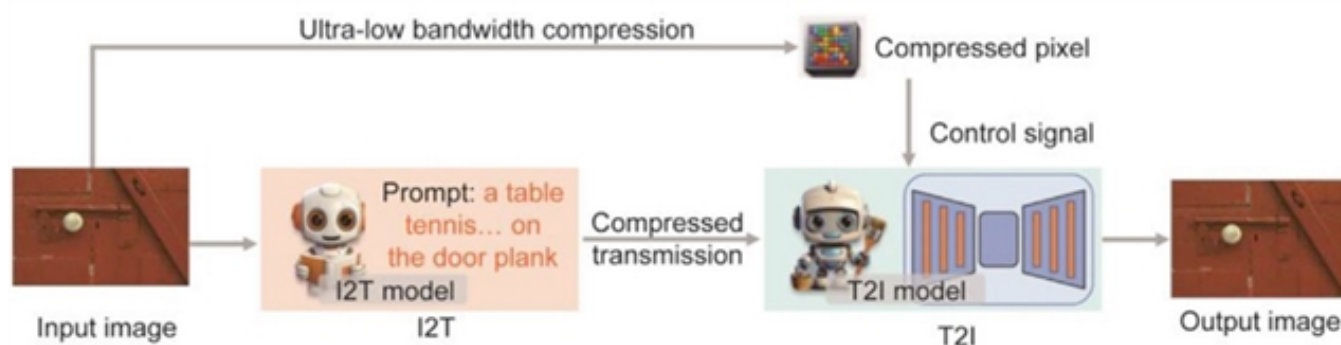


图3.多模态大模型驱动的生成式编码方案

第二条——协同传输：广播与蜂窝网络的协同。

研究团队提出的协同传输策略，充分调动了广播塔和蜂窝基站各自的优势，主要包括三种典型模式：

业务叠加：公共视频流用广播系统统一发送，个性化增强内容（比如用户专属视角）则通过蜂窝基站单播，最后在终端把两路信号融合，兼顾效率和个性化；

业务续接：当用户进入广播信号较弱的区域，核心网自动把业务转发给蜂窝基站接力，保障服务连续性；

业务卸载：当蜂窝网络遭遇突发流量高峰（比如大规模直播），核心网把部分流量卸载给广播系统分担压力，避免拥塞。

第三条——协同评估：面向生成内容的质量评估

生成式视频通信带来一个新问题：传统的PSNR、SSIM这类像素级质量指标，已经不足以全面评价生成式方法重建的视频质量了。研究团队提出，评估应该从三个维度展开：一致性（生成内容是否准确传达了应有的语义信息）、合理性（是否符合物理规律与伦理规范）、可用性（是否满足特定任务的个性化需求）。

为此，团队构建了一套基于多模态大模型的评估框架：先用视觉编码器提取参考图像和生成图像的特征，计算余弦相似度，再把一系列问题模板转化成文本特征，与视觉特征对齐后一并送入大语言模型，由它给出综合评分。这套框架还借助LoRA微调技术，只调整视觉编码器和视觉映射器的参数，最大程度保留预训练模型的先验知识。团队的实验显示，经过针对性微调后，模型在

低层视觉感知任务上的表现有明显提升——比如在LLVisionQA测试集上的整体感知准确率从60.07%提升到了69.30%。

优势与挑战并存

任何新范式都伴随着机遇与挑战。生成式视频通信的核心优势，是能用更紧凑的信号重建出高保真内容，实现低比特率下的高效传输，还能支持个性化和任务导向的内容生成，协同传输策略也能灵活适应不同设备能力和网络波动。

但挑战同样真实：生成模型天生存在不确定性，如何保证可靠性和可控性是需要解决的核心问题；实时生成和解码对计算资源要求很高，对算力有限的接收端是不小的负担；传统质量指标已经不够用，需要全新的评估体系；此外，生成式AI还带来了虚假信息、隐私泄露等新的伦理与安全风险，需要在编码、传输、评估各环节都同步跟进治理。

结果与讨论

这项研究勾勒的，是一幅视频通信范式转变的清晰路线图——从忠实传输每一个像素，走向理解内容、按需生成。

研究揭示的核心逻辑是：传统视频通信长期在信息域和物理域中持续优化，边际收益已趋饱和。而认知域，也就是视频内容真正承载的语义和语用价值，才是被严重低估的下一个增长点。生成式AI的加入，不是要取代传统技术，而是在以判别技术为基础、保证可靠性的前提下，将这一被忽视的认知域潜力释放出来。

当然，从理论到落地，前路依然漫长：生成式视频通信目前还缺乏像香农信息论那样成熟的数学基础；6G网络需要在带宽调度、低时延传输、隐私保护上做出适配；端到端的生成模型也还需要在轻量化和实时性上持续打磨。

但可以确定的是，随着生成式AI技术的持续演进，未来我们看视频的方式，很可能不再是被动地接收像素，而是以最少的信号驱动生成，重建出接近真实的视觉体验。（来源：EngineeringJournals微信公众号）

相关论文信息：<https://www.sciencedirect.com/science/article/pii/S2095809925003157>

作者：张文军等 来源：《工程》

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发