中国科学院大学

University of Chinese Academy of Sciences

学士学位论文

基于注意力注入的视频编辑方法研究

作者姓名：余正则

指导教师：钱胜胜

中国科学院自动化研究所

学位类别：工学学士

专　　业：人工智能

学院（系）：中国科学院大学人工智能学院

2025 年 6 月

Research on Video Editing Based on Attention Injection

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Bachelor of Engineering

in Artificial Intelligence

By

Yu Zhengze

Supervisor: Professor Qian Shengsheng

School of Artificial Intelligence

June, 2025

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

目前视频编辑领域中，梯度-潜变量净化（Gradient-Latent Purification, GLP）高效改善了潜变量优化过程中难免引入梯度噪声, 优化过程不可控的问题。通过局部坐标分解进行周期性梯度净化，以及最后阶段对潜变量进行稳定性重构，GLP 可以获得更准确地更新方向，同时，GLP 加入了时序平滑，提升了帧间一致性。经过多项视频编辑任务，GLP 在语义一致性、时序一致性、用户体验评估中均达到当前最优性能，并通过消融实验证明了 GP 与 LP 的协调。这些卓越的表现让我们看到了 GLP 在其他基于潜变量优化的框架中推广应用的潜力。在完整架构中，反向编辑与前向编辑相结合，以兼顾全局语义匹配与局部可控性。本文则从另一个角度出发，致力于发掘前向注意力注入的更多可能性。注意力注入机制在采样阶段不依赖梯度反传，在不改变潜变量的情况下，直接干预生成过程，对局部区域实现灵活控制。在 transformer 模型的注意力层中注入来自参考语义的信息，通过替换或融合查询向量、键向量、值向量或注意力图来引导生成结果。本文使用来自 V2VBench 对视频编辑质量全面科学评估，将典型注意力注入方式与 GLP 对比其在语义一致、对象一致、运动对齐、视频文本对齐等多方面的表现。P2P 将重建分支中的交叉注意力图注入到编辑分支中。PnP 注入自注意力图。实验结果表明，前向注意力注入优化可使 GLP 在局部区域取得更好效果。其中，PnP 与 GLP 结合在帧选择、文本对齐方面表现最优，P2P 在视频质量方面更好。因此，通过组合多种注意力注入方法，可从不同维度改进前向编辑的效果，全面提升视频编辑的水平。

关键词：视频编辑，扩散模型，梯度潜在纯化，V2V 基准，注意力注入

Abstract

In the field of video editing, Gradient-Latent Purification (GLP) has effectively addressed common challenges associated with latent optimization, including the inevitable introduction of gradient noise and the uncontrollability of the optimization process. By periodically purifying gradients through local coordinate decomposition and reconstructing the stability of latent variables in the final stage, GLP enables more accurate update directions. Meanwhile, GLP incorporates temporal smoothing to enhance inter-frame consistency. Across multiple video editing tasks, GLP achieves state-of-the-art performance in semantic consistency, temporal consistency, and user experience evaluation, with ablation studies further validating the synergy between gradient purification (GP) and latent purification (LP). These outstanding results suggest the strong potential of GLP for broader application within other latent optimization-based frameworks. In the complete architecture, GLP combines backward editing and forward editing to balance global semantic alignment with local controllability. This work, however, explores another perspective, focusing on uncovering the further potential of forward attention injection. Attention injection intervenes directly in the generation process during sampling without relying on backpropagation of gradients or altering latent variables, enabling flexible control over local regions. Specifically, reference semantic information is injected into the attention layers of the transformer model by replacing or fusing query vectors, key vectors, value vectors, or attention maps to guide the generation. We conduct a comprehensive and scientific evaluation of video editing quality using V2VBench, comparing typical attention injection methods with GLP across multiple aspects including semantic consistency, object consistency, motion alignment, and video-text alignment. In particular, P2P injects cross-attention maps from the reconstruction branch into the editing branch, while PnP injects self-attention maps. Experimental results show that forward attention injection further enhances GLP's performance in local regions. Among them, the combination of PnP and GLP achieves the best results in frame selection and text alignment, while P2P performs better in terms of overall video quality. Therefore, by integrating multiple attention injection methods, it is possible to improve forward editing performance from different perspectives and comprehensively enhance video editing quality.

Key Words: Video Editing, Diffusion model, Gradient Latent Purification, V2VBench, Attention Injection

目 录

图目录

表目录

英文缩略词表

AIGC	Artificial Intelligence Generated Content	人工智能生成内容
DDIM	Denoising Diffusion Implicit Models	去噪扩散隐式模型
DDPM	Denoising Diffusion Probabilistic Models	去噪扩散概率模型
DDS	Delta Denoising Score	增量去噪得分
GLP	Gradient-Latent Purification	梯度-潜空间净化
I2V	Image-to-Video	图像生成视频
PnP	Plug-and-Play	插件式编辑
SVD	Singular Value Decomposition	奇异值分解
T2I	Text-to-Image	文本生成图像
VAE	Variational Autoencoder	变分自编码器

第 1 章　绪论

1.1　视频编辑介绍

近年 AIGC 取得了显著的发展[4][5][6]。在计算机视觉领域，人工智能模型在生成高维数据方面表现出了优秀的能力，并对工业产品和日常生活产生深远的影响。其中，扩散模型[7]在图像生成任务中被广泛使用，包括文本生成图像[8][9][10]和图像生成图像[11][12][13]。大量方法使用扩散模型，可以根据文本描述生成高质量、语义准确的图像，或者根据指定条件修改输入图像。而视频由连续的图像序列组成。因此，将扩散模型从 2D 图像生成扩展到 3D 视频生成和编辑是视觉领域可能会是进一步的方向[14][15]。然而，将架构扩展到视频面临一些问题，例如，将对静态图像的设计适配到视频的动态和时间特性、高质量视频数据集的匮乏。研究人员需要从数据集、模型优化、硬件优化等方面去解决这些问题。解决问题需要循序渐进，视频生成任务中，编辑现有视频不需要昂贵的视频预训练，并能够对源视频进行细粒度的控制。所以可以先从视频编辑入手，对现有视频探索高效、准确的编辑方法，再将这些方法扩展到信息量更少的，自发生成的，前文所述需要大量数据集、更高效的模型、更高效硬件的昂贵训练任务。因此，从视频编辑到视频生成会是一个好的路径。

视频编辑过程[16]中，有监督学习通常依赖于明确的输入-输出映射，但在许多场景下，获取大规模标注数据成本极高甚至不可行。因此，自监督方法（不依赖外部标注，而是通过数据本身的结构或内在关系来构造监督信号，指导模型学习。）得到了广泛应用。在扩散模型[17]或 VAE[18]等框架中，视频或图像数据通常会被压缩到一个低维的“潜在空间”。这个潜在空间中的表示是一种高维特征，它比像素级数据更紧凑，并且能够捕捉视频的关键信息，如物体的形状、颜色、纹理以及运动模式。给定一个视频 I，我们可以通过编码器（Encoder）将其转换为潜在表示 $z = \text{Encoder}(I)$。在进行视频编辑时，我们的目标是找到一个优化后的潜在表示 $\hat{z}$，使得其解码后的结果 $\text{Decoder}(\hat{z})$ 具有期望的编辑效果（例如，改变物体颜色、形状或背景）。这个优化过程通常通过梯度下降或其他优化方法，基于某种损失函数来逐步调整 z，使其向目标属性收敛：$\hat{z} = z - \eta \nabla L(z)$。其中：$\eta$ 是学习率，$\nabla L(z)$ 是损失函数 L 关于 z 的梯度。目标是通过迭代优化 z，让其在潜在空间中向目标状态靠近。在基于潜在优化的视频编辑方法中，通常没有显式的“地面真实值”可用。也就是说，我们无法直接获取“理想编辑后的视频”的潜在表示 $\hat{z}$，而只能通过构造一种自监督目标来计算梯度，并引导优化。例如 DDS[19]：$L_{\text{DDS}}(z) = \left\| \epsilon_\phi(z_t, t, y) - \epsilon_\phi(z_t, t, \varnothing) \right\|^2$。其中：$z_t$ 是添加噪声后的潜在表示（扩散模型的一个中间状态），ϵ_ϕ 是扩散模型的去噪网络，y 是目标文本提示，$\varnothing$ 表示去除文本条件的信息。该损失度量了带有目标文本提示的去噪结果与不带文本提示的去噪结果之间的差异，目标是让 z 的更新方向朝着符合目标文本描述的方向前进。梯度是通过模型自身计算的，而非依赖外部标注数据，因此不

可避免地会引入梯度噪声。包括：扩散模型去噪过程的误差导致梯度计算不准确。模型训练的不完美（去噪网络 ϵ_ϕ 并不完美，它可能会错误地预测某些像素的噪声，从而导致梯度方向出现偏差。）、训练数据的偏差、采样过程中随机性的影响。随机噪声的累积使得优化过程中内容变得模糊或过度修改。每个迭代步骤都基于前一轮的结果进行更新。由于梯度是通过自监督计算的，并不是完美的，误差会逐步累积，导致模糊效应，梯度中的随机噪声会导致视频中的目标物体边缘变得模糊。过饱和，梯度方向不准确，可能会导致颜色或纹理的过度变化，使得编辑内容失真。不稳定的优化，由于梯度波动，优化过程可能会出现编辑过度或编辑不足，最终的潜在表示 $\hat{z}$ 可能并不是最优解。不同优化步骤的梯度方向可能不一致，导致优化结果不稳定或难以控制。

使用文本提示词的视频编辑方法可分为前向编辑方法和反向优化方法[20]，前向方法在模型的前向传播过程中修改和更新目标视频的潜在表示，常采用注意力特征注入技术。反向编辑方法通过设定一个优化目标来引导编辑过程，通常通过直接微调网络或潜在表示来实现所需的编辑效果。其中，微调网络的方式一般涉及调整模型中的注意力层或适配器，或结合大语言模型的能力，采用基于指令的微调策略进行监督调整。在这种反向与前向编辑相结合的方法面临一些挑战：首先是梯度方向不稳定，每一步优化的梯度可能会受到随机噪声影响，优化方向不一定始终朝向最佳结果。其次是优化步数不可控，无法预先知道多少步的优化是最合适的，可能会导致编辑不足或过度编辑。还有时间一致性差异，在视频编辑过程中，每一帧的优化是独立进行的，如果梯度方向不稳定，不同帧之间可能会出现颜色变化、形状变形等问题，导致视频的时序一致性降低。而后出现了 GLP 模型，高效改善了这两个问题[1]，GLP 收集多个优化步骤的信息，避免单步梯度的噪声影响。构造局部坐标系，通过 SVD 分解找到最稳定的优化方向。添加正则化项，让优化方向更加合理，避免过拟合。时间平滑策略，确保视频在时间维度上的一致性，减少闪烁和形变。

定性实验表明，GLP 在物体外观变换、纹理替换及背景编辑任务中，能够精准完成局部编辑，并在保持原有运动轨迹与背景结构的同时，有效避免了伪影、细节丢失或无关区域污染的问题。尤其在处理快速运动物体时，GLP 相较于传统方法表现出更强的鲁棒性和结构保持能力。定量评估方面，GLP 在 V2VBench[16] 基准数据集上，于语义一致性（Text-Align、Frame-Pick、Frame-Acc）、时序一致性（Motion-TC、Semantic-TC、Object-TC）及视频整体质量（V-Quality）指标上均优于或可比于当前最先进的方法。特别是在 Frame-Acc 和 Motion-TC 等指标上，GLP 显示出显著优势，验证了其在兼顾语义拟合与时间连贯性方面的综合能力。基于用户调查的主观评估进一步印证了 GLP 在运动保留、帧间流畅性与编辑意图匹配度等维度上的感知质量提升。本文沿此基础上，将重心聚焦在前向编辑，探索注意力注入机制在扩散采样过程中的作用与潜力。将典型的前向注意力注入方法与 GLP 对比，探究前向编辑优化的更多可能性。在评估方面，使用 V2VBench 从八个指标来给编辑效果打分，分别为：帧的质量、视频质量、语

义一致性、物体一致性、帧与文本对齐、视频与文本对齐、帧选择分数、运动对齐、内存使用和运行时间，全面地评估不同方法。

1.2　相关工作

基于注意力注入的特征编辑是视频编辑领域的一种表现出色的方法。它通过注意力机制，将预定义的特征（例如语义信息或结构属性）注入到模型中，指导图像或视频编辑的生成过程。本文主要调研与扩散模型相结合的基于注意力注入的特征编辑方法。根据注入的隐藏特征的来源，这种方法可以分为两类。一类是基于反转的特征注入，另一类是基于运动的特征注入。

1.2.1　基于反转的特征注入

从重构分支中提取的隐藏特征可以促进目标视频的生成。其次，考虑到视频中固有的时空信息复杂性，一些方法采用多分支结构来有效处理这些信息。因此，基于反转的特征注入可分为两种类型：双分支（Dual-Branch）和多分支（Multiple Branches）。

Video-P2P[3] 和 Vid2Vid-Zero[21]　将 P2P 框架扩展到视频编辑领域。两者都采用了交叉注意力图注入（Cross-attention map injection）和空文本反转（null text inversion）。交叉注意力图注入（Cross-attention map injection）：双分支：使用一个重构分支处理源图像，使用一个编辑分支处理目标图像。Prompt-to-Prompt(P2P) 强调了文本提示词（text prompts）在 T2I 扩散模型中的有效性。它将不变的 token 的交叉注意力图从重构分支注入到编辑分支，从而保留源图像的未改变区域。Plug-and-Play (PnP) 和 MasaCtrl[22] 强调自注意力层作为结构描述符的重要性。PnP 将自注意力查询、键以及每个 block 的输出从重构分支注入到编辑分支，来保持语义结构。MasaCtrl 保持查询特征不变，同时转移键和值特征。它通过阈值化编辑词的交叉注意力图，生成注入掩码，以解决前景和背景混淆的问题。特征注入方法可以通过潜在反转（latent inversion）将源图像反转，例如：

$$\mathbf{z}_t \approx \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\mathbf{z}_{t-1} + \left(\sqrt{1-\alpha_t} - \sqrt{\frac{\alpha_t}{\alpha_{t-1}} - \alpha_t}\right)\epsilon_\theta(\mathbf{z}_{t-1}, t, y).$$

$$\hat{\varnothing}_t =_{\varnothing_t} \|\mathbf{z}^u_{t-1} - \mathbf{z}_{t-1}(\mathbf{z}_t, t, y, \varnothing_t)\|_2^2,$$

随后，反转后的潜在状态被用于两个分支。

空文本反转[23]：在反向扩散步骤的早期，当噪声图像提供的信息有限时，文本条件在去噪中至关重要。假设我们想生成一个特定人物的图像，通过潜在反转技术，我们可以从一个随机噪声图像中恢复出一个具有特定特征的人物图像。如果在此过程中使用了一个带有特定描述的文本，比如“a person with a red hat”，模型会根据这个提示生成带有红色帽子的人物图像。然而，如果这个文本提示与目标图像不完全一致，模型可能会引入一些不必要的错误或偏差。在这种情况

下，我们可以使用空文本反演来消除文本提示的干扰。通过提供一个“空”的文本输入（如一个空字符串或无特定描述的文本），我们可以让模型在潜在空间中更加自由地优化和调整生成的图像，从而减少由文本提示误差引起的错误。潜在状态初始化方法擅长处理大的特征和姿势的修改。特征注入方法在风格化、背景和物体替换方面表现优秀，而文本反转方法更适用于物体个性化修改。许多正交方法可以协同使用，来实现多样的编辑任务。Video-P2P 用稀疏因果注意力替换了空间自注意力层，引入了时间注意力层，并实现了一次性调优来实现时间一致性。Vid2Vid-Zero 则是强调了全局时空注意力层的必要性，键和值的特征来自所有帧。即使没有参数调整，其膨胀的 2D 层实际证明了对双向时间建模的有效性能。

Fate-Zero[24] 和 Edit-A-Video[25] Fate-Zero 同时注入了交叉注意力和自注意力图，结合了 P2P 和 PnP 的优点。自注意力注入的融合掩码是通过二值化交叉注意力图生成的。Edit-A-Video 使用与 Fate-Zero 类似的注意力图注入方法，并使用空文本反转，来减轻与潜在反转，通过反向推理从潜在空间恢复或重构输入图像[26]。无分类器引导不依赖于传统的分类器，而是直接利用条件信息（例如文本提示）来引导生成过程。带来的误差。此外，它在第一帧和当前帧之间插值融合掩码，由自注意力图加权，从而确保编辑的时间一致性。

Make-A-Protagonist[27] Make-A-Protagonist 使用与 PnP 和 MasaCtrl 类似的方法，将重建分支中的潜在特征和自注意力映射注入到编辑分支中。与以往从注意力图中派生掩码的工作不同，它利用 Grounded SAM[28] 获得第一帧的分割掩码，并使用 Xmem[29] 来传播它，以实现精确的融合。

UniEdit[30] 和 AnyV2V[31] 通过引入两个辅助分支（重建分支和运动参考分支）来分离外观和运动注入。所有分支由预训练的视频生成模型初始化。重建分支在早期采样步骤中注入空间自注意力查询和键特征，以保持结构的连续。此外，从重建分支选择性地注入空间自注意力值特征，来保留未编辑的区域。运动参考分支通过仅注入其时间自注意力查询和键特征来促进生成所需的运动，从而避免结构不稳定性。尽管共享注入工作流程，UniEdit 和 AnyV2V 在基础模型上有区别。UniEdit 使用 T2V 模型 LaVie[32]，并有效地结合文本提示。AnyV2V 使用 I2I 专家[33][34][35][36] 编辑第一帧，随后应用基于预训练的 I2V 模型的视频编辑流程，用来提高该模型在处理不同编辑任务时的兼容性。

1.2.2 基于运动的特征注入

特征注入也可以跨帧进行。源视频中的内在运动信息为指示时间对应关系和指导帧间特征注入提供了宝贵的先验知识。

TokenFlow[37]　TokenFlow 首先通过替换空间自注意力为全局时空自注意力机制来处理一些采样的关键帧。然后，TokenFlow 为每个 token 识别来自相邻关键帧的最近邻。对于在帧 f 中具有空间坐标 (x_f, y_f) 的 token，其在相邻过去关键帧 $f-$ 和相邻未来关键帧 $f+$ 中的最近邻 token 分别表示为 (x_{f-}, y_{f-}) 和 (x_{f+}, y_{f+})。随后，每一帧都通过帧间特征注入进行独立去噪：

$$\mathbf{h}^{f}_{[x_f,y_f]} \leftarrow \alpha_f \mathbf{h}^{f-}_{[x_{f-},y_{f-}]} + (1-\alpha_f)\mathbf{h}^{f+}_{[x_{f+},y_{f+}]}.$$

其中 $\alpha_f \in (0,1) \in (0,1)$ 是一个标量，与帧 f 及其相邻关键帧 $j\pm$ 之间的距离成比例。$\mathbf{h}$ 表示自注意力层输出的中间隐藏特征。

FLATTEN[38]　FLATTEN 根据光流来选择注入的特征。对于一个 token $\mathbf{x}^{s1}_{(x_1,y_1)}$：$\mathbf{x}^{\mathbf{s}1}$ 表示在源视频的第一个帧,(x_1, y_1) 表示其空间坐标，光流估计方法可以追踪其在剩余的 $f-1$ 帧中的空间对应关系，将其表示为一条轨迹：

$$\mathcal{T}(\mathbf{x}^{\mathbf{s}}, x_1, y_1) = \{\mathbf{x}^{\mathbf{s}1}_{[x_1,y_1]}, \cdots, \mathbf{x}^{\mathbf{s}f}_{[x_f,y_f]}\}.$$

在编辑过程中，其时间引导的注意力机制将 Qtoken 与来自其他帧中同一轨迹的 Ktoken 和 Vtoken 配对：

$$\mathbf{Q}^{f}_{[x_f,y_f]} = \mathbf{h}^{f}_{[x_f,y_f]},$$
$$\mathbf{K}^{f}_{[x_f,y_f]} = \mathbf{V}^{f}_{[x_f,y_f]} = \mathcal{T}(\mathbf{h}, x_1, y_1) \setminus \{\mathbf{h}^{f}_{[x_f,y_f]}\}.$$

时间引导的注意力机制（temporal-guided attention）不涉及位置编码、投影矩阵、前馈层，因此它能在无需额外训练的情况下实现无缝集成。

FRESCO[39]　FRESCO 采用了与 FLATTEN 相似的方法，并且增加了额外的步骤来优化值特征。它对源帧 $\mathbf{z}^{sf}$ 执行一个正向和反向的 DDPM 步骤，并提取其查询特征 $\mathbf{Q}^{sf} \in \mathbb{R}^{d\times N}$ 和键特征 $\mathbf{K}^{sf} \in \mathbb{R}^{d\times N}$. 。然后，FRESCO 引入了一个空间引导的注意力机制，利用这些特征来重新加权编辑视频中的查询特征：

$$\mathbf{Q}^{f} \leftarrow \mathrm{Softmax}(\frac{(\mathbf{Q}^{sf})^{T}\cdot\mathbf{K}^{sf}}{\lambda\sqrt{d}})\cdot\mathbf{Q}^{f},$$

其中，λ 是一个缩放因子。随后，它利用高效的跨帧注意力机制，用跨帧信息增强值特征：

$$\mathbf{V}^{f} \leftarrow \mathrm{Softmax}\Big(\frac{(\mathbf{Q}^{f})^{T}\cdot\mathbf{K}_{[\mathbf{p}]}}{\sqrt{d}}\Big)\cdot\mathbf{V}_{[\mathbf{p}]}.$$

在这里，$\mathbf{p}$ 表示一组被关注的坐标：第一帧中的所有坐标都被包括在内，而后续帧只包括以前未观察到的区域。这些经过优化的值特征被用于一个时间引导的

注意力机制中，类似于 FLATTEN 的机制。此外，它还对每个解码器层的输入特征进行了微调。源视频和编辑后的视频经历了一个 DDPM 正向和逆向步骤。在逆向过程中，源视频和目标视频的第 f 帧的输入特征分别表示为 $\mathbf{h}^{\mathbf{s}f} \in \mathbb{R}^{d\times N}$ 和 $\mathbf{h}^{f} \in \mathbb{R}^{d\times N}$。这些特征嵌入通过最小化时空对齐损失来优化：

$$\mathcal{L} = \mathcal{L}_{\text{temporal}} + \lambda \mathcal{L}_{\text{spatial}},$$

$$\mathcal{L}_{\text{temporal}} = \sum_{f=1}^{F-1} \|\mathcal{O}^{f\to f+1}(\mathbf{h}^{f+1} - \mathcal{F}^{f\to f+1}(\mathbf{h}^{f}))\|_1,$$

$$\mathcal{L}_{\text{spatial}} = \lambda \sum_{f=1}^{F} \|\mathbf{h}^{\mathbf{s}f^T} \cdot \mathbf{h}^{\mathbf{s}f} - \mathbf{h}^{f^T} \cdot \mathbf{h}^{f}\|_2^2$$

其中，λ 是一个标量，用于平衡。

第 2 章　方法

2.1　基本框架

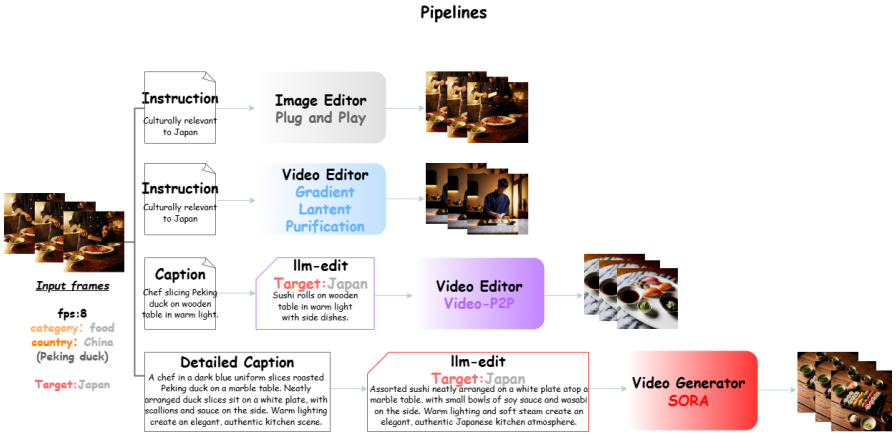

图 2-1 框架图

在编辑过程中，提出了一套用于跨文化视频内容转化的多阶段处理流程，旨在将特定文化背景下的视频帧（如中国的北京烤鸭）转换为符合目标文化（如日本寿司）的表达形式。整个系统以原始视频帧为输入，帧率设置为 8 fps，视频内容类别为“食物”，原始国家标签为“中国”，目标文化标签为“日本”。在预处理阶段，系统首先生成多种层级的文本描述，包括基本字幕（caption）和详细描述（detailed caption），并辅以目标文化提示（instruction），指导后续的图像或视频生成模块。

在图像层面，输入帧与文化指令一同送入基于插件式编辑的图像编辑器，以生成具有日本文化视觉特征的静态图像序列。在视频层面，系统引入梯度纯度净化（GLP）方法，通过潜空间优化实现跨帧一致性的视频编辑，生成风格连贯、符合文化目标的视频片段。此外，系统还集成了语言引导的编辑路径。通过大语言模型对字幕进行目标化重写后，文本与原始帧结合输入 Video-P2P 编辑器，实现语义级别的视频风格迁移。

在更高层次的生成路径中，系统使用详细描述作为输入文本，经 llm-edit 重构后送入 SORA 视频生成器，直接合成具有完整文化风格转换效果的高质量视频内容。整体流程支持从文本、图像、视频三个层次进行文化适配，形成一个集成式、多模态的视频编辑与生成框架。

2.2　梯度-纯度净化[1]

构造局部坐标轴过程中采用了奇异值分解（SVD）方法，将一个 $m \times n$ 的矩阵 A 分解为三个矩阵的乘积：$A = U\Sigma V^T$ 其中，U 为左奇异向量矩阵，大小为

$m \times m$，包含矩阵 A 的正交基（输出空间的旋转）；Σ 为奇异值对角矩阵，大小为 $m \times n$，对角线上是奇异值，表示不同方向上的缩放因子；V^T 为右奇异向量矩阵的转置，大小为 $n \times n$，对应输入空间的正交基（先将输入旋转到新的坐标系）。可以将 A 理解为三个线性操作的组合：

$$\text{输入} \xrightarrow{\text{旋转}V^T} \xrightarrow{\text{缩放}\Sigma} \xrightarrow{\text{旋转}U} \text{输出}$$

也就是说，A 首先将输入向量旋转（变换坐标轴），然后按每个方向缩放（奇异值控制），最后再次旋转得到最终输出。构造局部坐标系的步骤为：首先收集最近 J 步优化的梯度：$G = \{g^{(i)}, g^{(i+1)}, g^{(i+2)}, \ldots, g^{(i+J)}\}$。其中，每个 $g^{(i)}$ 是第 i 步优化计算得到的梯度。再计算这些梯度的协方差矩阵：$\text{Cov}(G, G) = G^T G$[40]。这有助于分析梯度在不同方向上的变化情况。而后对协方差矩阵执行 SVD 分解，找到最稳定的方向：$\lambda_k, u_k = \text{SVD}(\text{Cov}(G, G))$。其中 u_k 是新的坐标轴方向（梯度的主要变化方向），λ_k 是对应的特征值，表示变化的幅度。最后选择最稳定的方向进行优化：$\nabla z = \sum_k a_k u_k$ 其中 a_k 是优化权重。这样，就得到了一个新的梯度方向，它是多个优化步骤中最一致的方向，可以减少梯度噪声的影响。在计算梯度方向时，还需要一个正则化项，用于确保新的梯度方向：符合扩散模型的分布特性（保持高斯性）；不会过度依赖单个梯度方向（防止过拟合）；不会导致梯度方向过大或过小（控制优化步长）。数学上，正则化项的定义如下：$L(a_k) = \left(\beta_1(\nabla z)\right)^2 + \left(\beta_2(\nabla z) - 3\right)^2$[41]。其中，$\beta_1$ 用于约束梯度方向的偏斜程度；β_2 用于约束梯度方向的峰度，使其接近高斯分布的标准值 3。这保证了优化方向更加平稳，不会受到随机梯度噪声的影响。

同时，还有时间平滑轴扩展用于保证不同帧的优化方向保持一致，使得视频的时间一致性更强。在视频编辑任务中，每一帧的优化是独立进行的，如果不同帧的梯度方向不一致，可能会导致：颜色抖动（例如前一帧是红色，后一帧变回白色）；形状不稳定（例如物体在不同帧中发生形变）。方法如下：在局部坐标系构造过程中，引入相邻帧的信息：$\lambda_f \leftarrow \{\lambda_{f-1}, \lambda_f, \lambda_{f+1}\}$，$u_f \leftarrow \{u_{f-1}, u_f, u_{f+1}\}$。这样，每一帧的梯度方向都会考虑前后帧的信息，确保它们不会发生剧烈变化。在最终潜在表示优化时，引入时间平滑项：$z_f^{\text{final}} = \sum_k \frac{1}{\eta_k} \left\langle z_f^{\text{avg}}, u_f^k \right\rangle \cdot u_f^k$。其中，$z_f^{\text{avg}}$ 是过去几帧的平均潜在表示，用于确保最终的优化结果在时间上平滑过渡。

2.3 注意力注入

2.3.1 PnP 图像编辑[2]

设 x_T^G 是初始噪声，通过使用 DDIM 反向扩散对源图像 I_G 进行反演（inversion）获得。在给定目标提示词 P 的情况下，生成翻译后图像 I 时，使用相同的初始噪声，即 $x_T = x_T^G$。在反向扩散过程的每一个时间步 t，从去噪步骤中提取引导特征 f_t^l，即：$z_{t-1}^G = \epsilon(x_t^G, t)$。其中，$\epsilon(\cdot)$ 表示扩散模型的去噪函数。然后，这些特征 f_t^l 被注入到图像 I 的生成过程中：即在对 x_t 的去噪步骤中，我们用提

取的特征 f_t^l 覆盖当前层的中间特征。该操作可表示为：$z_{t-1} = \epsilon(x_t, P_t; f_t^l)$, 其中，$\epsilon(x_t, P_t; f_t^l)$ 表示在注入特征 f_t^l 的条件下进行的修改版去噪过程。若不注入特征，则简化为普通的去噪过程：$\epsilon(x_t, P_t; \cdot) = \epsilon(x_t, P_t)$。根据不同深度的层 l 注入空间特征 f_t^l 的效果，仅在浅层（如 $l = 4$）注入特征，无法很好地保持源图像的结构。随着注入层数加深，结构保持效果逐渐增强，但同时也会泄露源图像的外观信息。为了在保留结构与避免外观泄漏之间取得更好的平衡，在深层（deeper layers）不再修改空间特征，而是利用自注意力（self-attention）机制进一步控制生成过程。自注意力模块在将空间特征线性投影到查询（query）和键（key）后，计算它们之间的相关性矩阵 A_t^l。通过注入自注意力矩阵 A_t^l，实现对生成内容的细粒度控制。在不同层级下的自注意力矩阵 A_t^l，在浅层，自注意力体现了图像的语义布局（semantic layout），例如将不同语义区域归类到一起。在深层，模型捕捉到了越来越多的高频细节信息。在实践中，注入自注意力矩阵的方法是：在去噪过程中，直接用参考的 A_t^l 替换当前步的自注意力矩阵。直观上，这种操作会根据 A_t^l 中编码的亲和度信息，将特征在空间上拉近或重新组织。修改后的去噪过程表达为：$z_{t-1} = \epsilon(x_t, P_t; f_t^l, A_t^l)$, 自注意力注入的最大层数，直接控制了生成图像对于原始结构的保真程度。同时，有效缓解了外观信息泄漏的问题。

2.3.2　P2P 图像编辑[3]

每个扩散步 t 包括从带噪图像 z_t 和文本嵌入 P 预测噪声的过程，该过程通过一个 U 型网络实现。在最终步，生成的图像即为：$I = z_0$. 更重要的是，像素与文本之间的交互发生在噪声预测阶段：文本特征和视觉特征通过交叉注意力层[42] 融合, 产生针对每个文本 token 的空间注意力图。噪声图像 z_t 的深层空间特征首先被投影到查询矩阵 $Q = Q(\phi(z_t))$；文本嵌入被投影到键矩阵 $K = K(\phi(P))$ 和值矩阵 $V = V(\phi(P))$；这些是通过学习到的线性映射 Q、K、V 得到的。随后，注意力图 M 由以下公式计算：$M = \text{Softmax}\left(\frac{QK^T}{d}\right)$, 其中，$M_{ij}$ 表示第 i 个像素与第 j 个 token 之间的注意力权重，d 是键与查询向量的投影维度。最终的交叉注意力输出为：$\phi(z_t) = MV$, 并用来更新空间特征 $\phi(z_t)$。直观来看，交叉注意力输出 MV 是值向量 V 的加权平均，其中权重来源于注意力图 M，且权重反映了查询 Q 和键 K 的相似程度。使用多头注意力（Multi-head Attention）机制，并将各头输出拼接后经过线性层，得到最终结果。Imagen 与 GLIDE 类似，在每个扩散步的噪声预测过程中引入了文本条件，包括：交叉注意力；混合注意力（Hybrid attention），即将文本嵌入序列拼接到自注意力的键和值。将原始提示词 P 生成过程中的注意力图 M 注入到修改后的提示词 P' 生成中。合成出既符合修改后提示词、保留输入图像 I 结构的编辑后图像 I'。设 $DM(z_t, P_t, s)$ 表示扩散过程中的一个单步操作，它输出带噪图像 z_{t-1} 和注意力图 M_t。定义：

$\text{DM}(z_t, P_t, s)_M^M$ 表示在扩散时，覆盖使用给定注意力图 M，而仍使用目标提示 P_t 的值向量 V；

M_t' 表示使用编辑后提示词 P' 生成的注意力图；

$\text{Edit}(M_t, M_t', t)$ 是一个一般编辑函数，输入原始和编辑后的第 t 步注意力图。

总体流程是：对原始提示和修改后提示同步进行迭代扩散过程，并在每步应用基于注意力的编辑操作。注意，为了确保生成过程可控，固定扩散过程中的随机性。例如，将提示词从“a big red bicycle”修改为“a big red car”。注入源提示词对应的注意力图。但如果在所有扩散步都注入，可能会过度限制几何变化（如将车变为自行车时结构变形）。柔性注意力注入策略：

$$\text{Edit}(M_t, M_t', t) = \begin{cases} M_t, & t < \tau \\ M_t', & \text{otherwise} \end{cases}$$

其中，τ 是一个时间戳参数，控制注入源注意力的步数。如果替换的单词数量不一致（如多 token 变单 token），则可以使用对齐函数（alignment function）进行补齐或平均处理。例如，将提示词从“a castle next to a river”修改为“children drawing of a castle next to a river”。为保留公共部分细节，只对原提示和新提示共有 token 的注意力图进行注入。具体定义对齐函数 $\mathcal{A}$，它将目标提示词 P' 中的 token 索引映射到原提示词 P 中对应 token 索引（若无对应返回 None）。编辑函数为：

$$(\text{Edit}(M_t, M_t', t))_{ij} = \begin{cases} (M_t')_{ij}, & \mathcal{A}(j) = \text{None} \\ (M_t)_{i\mathcal{A}(j)}, & \text{otherwise} \end{cases}$$

其中，i 为像素索引，j 为文本 token 索引。注意力重新加权（Attention Re-weighting）用户可以调整某个 token 对生成图像的影响强弱。例如，在提示词”a fluffy red ball”中，希望让球看起来更或更少“fluffy”。做法是将对应 token 的注意力图乘以缩放参数 c：

$$(\text{Edit}(M_t, M_t', t))_{ij} = \begin{cases} c \cdot (M_t)_{ij}, & j = j^* \\ (M_t)_{ij}, & \text{otherwise} \end{cases}$$

其中，j^* 是需要调整的 token 索引，c 为缩放系数，可精细、直观地控制局部属性变化。

第 3 章　结果

3.1　数据集

我们从 V2VBench、Video-P2P、的数据集以及自己制作组成了了 20 个高质量视频作为数据源。每个视频挑选了包含主要动作、表达清晰的 8 帧组成长为 1 秒，帧率为 8 帧每秒（FPS）的 gif，分辨率均为 512×512 像素。这些视频被归为五个与实际应用场景相关的常见类别：动物、运动、人类、风景和交通运输，从每个视频片段中采样 8 帧。采样间隔（stride）根据具体情况手动调整，以保持动态范围的平衡，确保覆盖充分。将所有采样到的视频帧下采样并裁剪到 512×512 的空间分辨率，确保在大多数帧中主体相对居中。除了源视频输入外，每个视频包含了一条描述性文本提示和一条目标文本提示，以此组成了文本视频结合的编辑任务。为了涵盖视频编辑的多样化应用场景，从每个视频片段的三种不同的目标文本提示，分别对应前景对象编辑、背景编辑和风格化中随机选择一条进行编辑。不同的编辑目标组合，使编辑任务更丰富。

3.2　V2VBench 评价标准

V2VBench 从 8 个维度评估视频编辑质量：帧质量（Frames Quality）。在整体考虑所有帧之前，每一帧的质量构成了确定视频整体质量的基础。使用基于人工排名校准的 LAION[43] 美学预测器评估编辑后视频帧的质量。该预测器衡量了主观方面，包括布局、丰富性、艺术性和视觉吸引力。随后，我们计算平均美学得分，以推导出编辑后视频的整体质量得分。视频质量（Video Quality）。除了逐帧评估外，使用 DOVER[44] 得分进行视频级别评估。DOVER 是当前最先进的视频质量评估方法，基于大规模人工排名的视频数据集训练而成。除了美学考量，DOVER 得分还评估了技术指标，包括伪影、失真、模糊和内容的有意义性。语义一致性（Semantic Consistency）。CLIP[45] 视觉嵌入被广泛用于捕捉图像语义信息。相邻帧之间 CLIP 嵌入的余弦相似度是评估帧间一致性以及编辑后视频整体平滑性的标准指标。对象一致性（Object Consistency）。除了评估语义一致性外，还检查编辑后视频中对象层面的外观是否保持一致。使用 DINO[44][46] 这一自监督预训练图像嵌入模型，计算对象级的帧间相似性。帧与文本对齐（Frames Text Alignment）。除了评估视频的视觉质量外，与目标文本提示的一致性也是可控视频编辑的重要指标。CLIP 得分是评估视觉与文本对齐最广泛使用的指标。按照既有方法，逐帧计算 CLIP 得分，并在所有帧上取平均。视频与文本对齐（Video Text Alignment）。单独的帧无法充分代表运动，且传统的 CLIP 得分可能无法准确反映视频层次的对齐情况。计算 ViCLIP[47] 得分。ViCLIP 得分基于视频-文本对进行训练，用于评估时空内容与目标提示之间的对齐程度。帧选择[48] 得分（Frames Pick Score）。帧选择得分是一种基于 CLIP 的打分函数，训练于大规模人

工偏好标注数据集，综合考虑图像和文本输入，以反映内容质量和对齐度。逐帧计算帧选择得分，并取平均作为指标。运动对齐（Motion Alignment）。保持与源视频一致的运动模式至关重要。为此，利用 GMFlow[49] 为每对源-编辑视频计算光流，并测量其差异，称为端点误差（EPE）。为了与其他指标保持一致（即得分越高越好），对最终的 EPE 取相反数。

3.3 实验结果

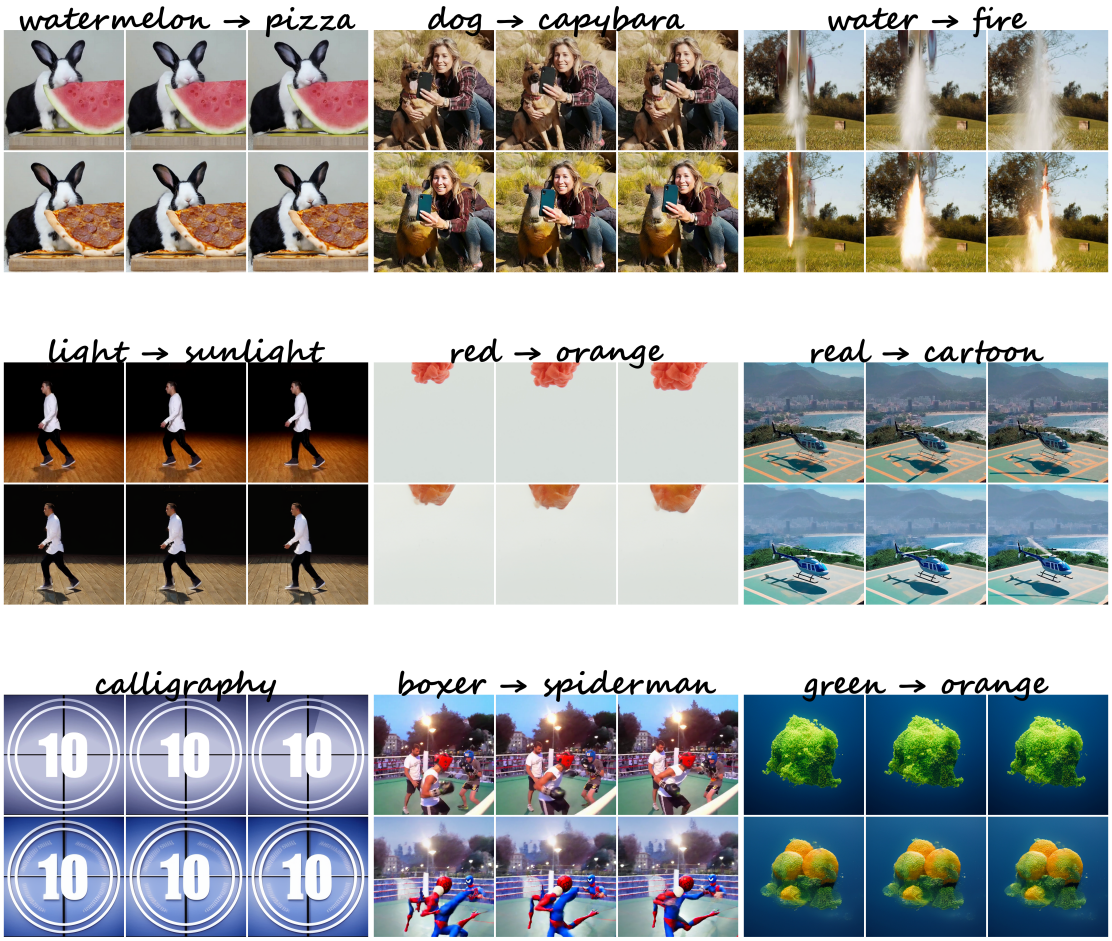

图 3-1 GLP 编辑结果展示

实验采用基于 DDIM[50] 的反演与重建框架对视频进行处理，并设计了详细的参数配置以平衡编辑效果与计算效率。整体流程包括两个阶段：反演阶段（DDIM inversion）和重建阶段（DDIM reconstruction），两者都基于相同的视频输入配置。设定了图像分辨率为 512×512，视频帧数为 8 帧，并通过 $stride = 3$ 控制帧采样间隔，即每隔三帧选取一帧，以降低计算负担。在反演与重建过程中，均设置了 500 步的扩散步数，确保潜变量能够充分学习到视频帧的结构与语义特征。此外，为了实现文本与图像的双重控制，我们同时设置了文本引导权重 $cfg_txt = 1.0$ 和图像引导权重 $cfg_img = 1.0$，使编辑既符合文本语义，又保

留原始图像的结构。基于 Plug-and-Play（PnP）特征注入框架，结合 DDIM 潜变量优化策略，对视频内容进行定向编辑。配置中详细设置了输入视频、扩散步骤、引导强度、结构注入方式以及 DDS（Delta Denoising Score）优化模块等关键参数，以保证编辑过程的可控性和可解释性。在输入部分，视频尺寸设定为 512×512，帧数为 8，通过 stride=3 控制帧间隔抽样以节省计算资源。编辑引导通过设置 $cfg_txt = 7.5$ 强化对文本提示的响应，同时通过 $editingnegativeprompt$ 明确禁止生成畸形、模糊、静态或不连贯的内容，从而提升输出质量。扩散步数为 500，初始潜变量步长索引 $ddiminitlatentstidx = 0$，确保从完整的噪声扩散路径中开始生成过程。在 PnP 注入机制中，通过 t_0, t_1, t_2 三个参数分别控制内容、运动、结构三类注意力的注入时刻。本实验中设置全部为 0，表示统一在最初阶段注入特征。同时，通过 $skiplayercontent = 0$、$skiplayermotion = 0$ 和 $skiplayerstructure = 25$ 来控制不同类型注意力跨层的注入深度。其中结构层仅保留 25 层以上的内容，细化结构控制粒度。优化迭代次数为 500 次，学习率设为 0.07，同时损失加权因子 w1=2、w2=2，被设置用于平衡结构与语义一致性目标。

P2P 实验：实验基于 Stable Diffusion v1-5 预训练模型进行微调，采用 8 帧采样，帧尺寸为 512×512 像素，采样间隔为 1 帧/秒，且从第 0 帧开始。生成过程中使用 50 步扩散推理，guidance scale 设置为 12.5，以强化文本引导效果，并启用了逆向潜变量推理，逆向步数同样为 50 步。

在训练配置方面，采用了较小的学习率（3e-5），每批次处理 1 个样本，最大训练步数设置为 500 步，未启用中间权重保存。可训练模块限定在局部的注意力权重，以期通过轻量微调增强模型编辑能力。训练过程中启用了 fp16 混合精度、xFormers 内存优化注意力机制及梯度检查点策略，同时未启用 8bit Adam 优化器。实验设定了随机种子 33，以保证结果的可复现性。

从实验结果来看，在语义一致性方面，各视频的 Pick Score 平均为 0.270，ViCLIP 文本对齐得分为 22.94，CLIP 文本对齐得分为 29.52。整体而言，模型能够较好地理解并遵循新的文本提示，特别是在 “rabbit” 视频上表现优异，展现了在复杂结构变化下的良好文本适配能力。在时序一致性方面，DINO Consistency 均值为 0.6595，表明物体形态在帧间变化平稳。Motion Alignment 得分为-1.4367，虽然在绝大多数样本中保持较好，但在 “man” 视频（Spider-Man 驾驶摩托）上存在较大的运动不稳定，提示在复杂动态建模上仍有改进空间。CLIP Consistency 均值达到 0.9445，验证了语义连续性整体良好。

在生成质量方面，Aesthetic Score 平均为 5.24，Dover Score 平均为 0.6197，均表现出稳定且较高的画面美学质量。其中，“man” 视频在 Dover Score 上达到 0.8077，显示了模型在复杂编辑任务中保持较高整体画面一致性的能力。综合来看，本次实验在语义理解与画面美学上取得了理想效果，但在动态连贯性上仍有进一步优化的空间。

表 3-1 GLP 语义一致性

video id	pick score	viclip text alignment	clip text alignment
blue-texture	0.2308	21.2133	24.8943
boxing-fisheye	0.2965	20.9596	35.0996
cat	0.2602	20.7265	25.7198
count	0.2358	21.1166	24.5689
cross-city	0.2119	21.9587	21.7187
dog-agility	0.2406	19.9359	26.0769
helicopter	0.2209	22.5143	28.2553
hike	0.2359	20.2994	30.0854
horsejump-high	0.2097	19.8040	25.0147
ink	0.1417	19.5265	21.4354
jellyfish	0.2533	21.6096	28.5388
kite-surf	0.2201	19.5570	25.2513
makeup	0.2562	20.4835	24.0115
moonwalk	0.2853	21.0387	29.0100
parkour	0.3340	20.7103	30.8088
rabbit-watermelon	0.2978	22.8497	31.9646
tennis	0.2284	20.6977	24.6512
toy-rocket	0.3190	22.6156	29.1425
volcanic-eruption	0.1809	20.9885	28.9149
woman-dog	0.3061	23.3965	34.0035
Average	0.2482	21.1001	27.4583

表 3-2 GLP 时序一致性

video id	motion alignment	clip consistency	dino consistency
blue-texture	-0.5052	0.9729	0.8130
boxing-fisheye	-4.2389	0.9520	0.3661
cat	-2.4069	0.9467	0.8555
count	-0.0988	0.9949	0.7655
cross-city	-1.1804	0.9598	0.5555
dog-agility	-2.9530	0.9100	0.5429
helicopter	-0.3542	0.9687	0.7922
hike	-0.7333	0.9606	0.7267
horsejump-high	-2.8672	0.9346	0.6070
ink	-0.8276	0.8932	0.3463
jellyfish	-2.1359	0.9835	0.7600
kite-surf	-4.6525	0.9534	0.7714
makeup	-1.7485	0.9592	0.7539
moonwalk	-0.6966	0.9056	0.6538
parkour	-2.5170	0.9106	0.6534
rabbit-watermelon	-0.2680	0.9950	0.7965
tennis	-2.8132	0.9107	0.5736
toy-rocket	-12.1154	0.9337	0.5365
volcanic-eruption	-0.4191	0.9930	0.8985
woman-dog	-1.5709	0.9531	0.8816
Average	-2.2551	0.9496	0.6825

表 3-3 GLP 视频质量

video id	aesthetic score	dover score
blue-texture	5.6281	0.9723
boxing-fisheye	4.5609	0.8922
cat	5.0145	0.9680
count	4.8534	0.9977
cross-city	5.2381	0.9615
dog-agility	4.3220	0.8687
helicopter	5.4832	0.9656
hike	5.2933	0.9784
horsejump-high	4.9420	0.8711
ink	4.3220	0.8292
jellyfish	5.0385	0.9811
kite-surf	4.9255	0.9728
makeup	4.9686	0.9301
moonwalk	4.5806	0.9596
parkour	4.4208	0.8509
rabbit-watermelon	5.4528	0.9921
tennis	4.5213	0.9253
toy-rocket	4.9952	0.8248
volcanic-eruption	5.9094	0.9909
woman-dog	5.0070	0.9170
Average	4.9739	0.9325

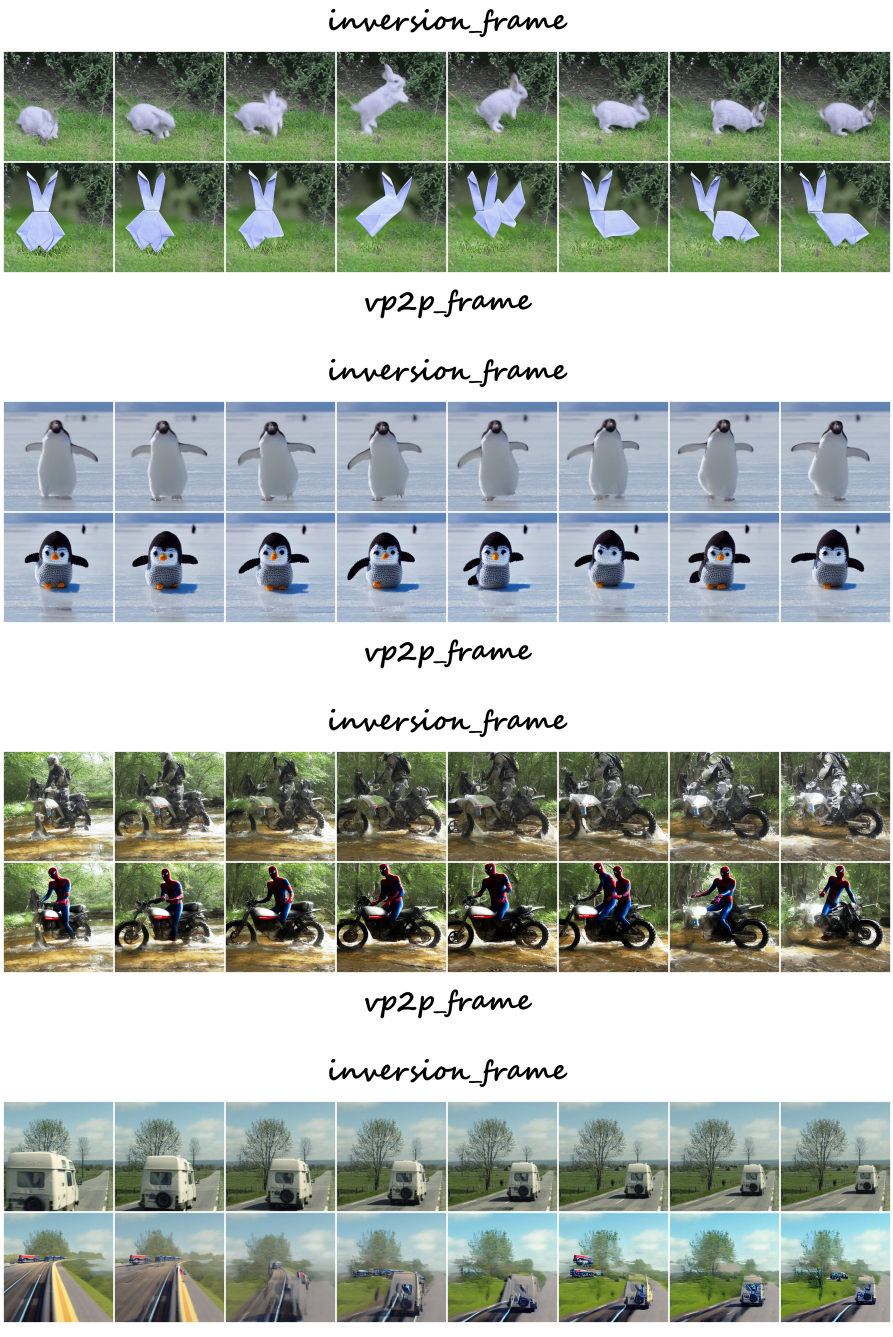

图 3-2 Video-P2P 编辑结果展示

表 3-4 Video-P2P 语义一致性

video id	pick score	viclip text alignment	clip text alignment
blue-texture	0.2822	18.7806	26.6764
boxing-fisheye	0.2284	20.3981	32.5473
cat	0.3000	20.9953	26.7531
count	0.3076	21.5035	30.5813
cross-city	0.1920	21.6535	24.7323
dog-agility	0.2179	21.7520	27.2810
helicopter	0.2751	20.3815	24.6749
hike	0.2574	20.2452	26.8134
horsejump-high	0.2693	21.8296	30.9411
ink	0.2359	20.5704	21.9544
jellyfish	0.3052	19.2692	29.5205
kite-surf	0.3011	19.4401	28.8321
makeup	0.2726	19.6967	25.9205
moonwalk	0.3151	21.3966	22.9388
parkour	0.2887	20.9709	28.2525
rabbit-watermelon	0.2356	21.6286	26.1468
tennis	0.2848	20.2873	28.8144
toy-rocket	0.2316	20.9902	28.0765
volcanic-eruption	0.3051	21.5507	33.6062
woman-dog	0.2680	20.4288	27.1623
Average	0.2687	20.6884	27.6113

表 3-5 Video-P2P 时序一致性

video id	motion alignment	clip consistency	dino consistency
blue-texture	-3.3748	0.8929	0.7315
boxing-fisheye	-1.8174	0.8920	0.3978
cat	-1.9568	0.9070	0.7986
count	-1.7517	0.8833	0.7277
cross-city	-1.0303	0.9141	0.4929
dog-agility	-3.2379	0.9304	0.5888
helicopter	-2.3547	0.9047	0.6991
hike	-1.6455	0.9106	0.6674
horsejump-high	-2.9490	0.9016	0.6157
ink	-1.1144	0.9120	0.6440
jellyfish	-1.7790	0.8784	0.6217
kite-surf	-1.4920	0.9095	0.6824
makeup	-2.3426	0.8971	0.8802
moonwalk	-1.5117	0.9628	0.2745
parkour	-1.3589	0.9472	0.6240
rabbit-watermelon	-1.8603	0.9116	0.6860
tennis	-1.8594	0.8771	0.7039
toy-rocket	-1.8378	0.8666	0.6037
volcanic-eruption	-2.5180	0.9357	0.3960
woman-dog	-2.0885	0.9079	0.7087
Average	-1.9940	0.9065	0.6272

表 3-6 Video-P2P 视频质量

video id	aesthetic score	dover score
blue-texture	4.9796	0.9274
boxing-fisheye	5.1940	0.9875
cat	5.4405	0.8419
count	4.6172	0.8594
cross-city	4.9906	0.8419
dog-agility	5.6332	0.8823
helicopter	4.7432	0.9440
hike	4.7117	0.9482
horsejump-high	5.1215	0.9358
ink	5.3784	0.8514
jellyfish	5.0452	0.8065
kite-surf	5.5722	0.9247
makeup	5.3833	0.8984
moonwalk	5.3773	0.9450
parkour	5.2622	0.9571
rabbit-watermelon	5.9717	0.9289
tennis	4.9585	0.9600
toy-rocket	4.2392	0.9335
volcanic-eruption	5.6823	0.9485
woman-dog	4.8575	0.8868
Average	5.1580	0.9105

第 4 章　讨论

在实验中，我们从三个维度对编辑后的视频进行了系统性的量化评估，分别是语义一致性（semantic consistency）、时序一致性（temporal consistency）与生成质量（generation quality）。这三类指标共同构成了我们视频编辑效果的评价体系，能够从不同角度揭示模型在处理各类编辑任务时的表现优劣。

表 4-1 GLP 与 Video-P2P 编辑效果对比

指标类别	指标名称	GLP	Video-P2P
语义一致性	Pick Score	0.2482	0.2687
	ViCLIP Text Alignment	21.1001	20.6884
	CLIP Text Alignment	27.4583	27.6113
时序一致性	Motion Alignment	-2.2551	-1.9940
	CLIP Consistency	0.9496	0.9065
	DINO Consistency	0.6825	0.6272
生成质量	Aesthetic Score	4.9739	5.1580
	Dover Score	0.9325	0.9105

GLP 与 Video-P2P 编辑效果对比中，我们比较了两种主流编辑方法的整体性能。从语义一致性来看，Video-P2P 的 Pick Score（0.2687）和 CLIP Text Alignment（27.6113）均略高于 GLP 的对应指标（0.2482 和 27.4583），说明 Video-P2P 在帧选择准确性以及帧级语义匹配方面具有更强的表达能力。而 GLP 在 ViCLIP Text Alignment 指标上略优（21.1001 vs. 20.6884），表明其在保持视频整体语义与文本描述一致性方面具备一定优势。在时序一致性方面，GLP 在 Motion Alignment（-2.2551）和 CLIP Consistency（0.9496）两个指标上均高于 Video-P2P，显示出其在保留时序结构与运动连贯性方面表现更稳定。相比之下，Video-P2P 的生成质量更具优势，其 Aesthetic Score 高达 5.1580，Dover Score 亦达到 0.9105，说明该方法在视觉美感与整体生成质量上更具优势。

LLM-edit 编辑效果：在 Video-P2P 框架中引入语言大模型（LLM）进行 prompt 重写对生成结果的影响。对比 None（无编辑）与 LLM-edit 两种设定可以发现，LLM-edit 显著提升了 ViCLIP Text Alignment（由 20.6884 提高至 21.2039）与 CLIP Text Alignment（由 27.6113 提高至 27.9552），表明重写后的 prompt 更有助于模型理解与生成符合目标语义的视频内容。尽管 Pick Score 在 LLM-edit 中略有下降（0.2436），这可能是由于生成的视频帧覆盖语义更广，从而削弱了帧选择的明确性。然而，在时序一致性方面，LLM-edit 在 Motion Alignment、CLIP Consistency 和 DINO Consistency 三项指标上均实现了提升，尤其是运动一致性从 -1.9940 明

表 4-2 LLM-edit 编辑效果

指标类别	指标名称	None	LLM-edit
语义一致性	Pick Score	0.2687	0.2436
	ViCLIP Text Alignment	20.6884	21.2039
	CLIP Text Alignment	27.6113	27.9552
时序一致性	Motion Alignment	-1.9940	-2.6195
	CLIP Consistency	0.9065	0.9284
	DINO Consistency	0.6272	0.6379
生成质量	Aesthetic Score	5.1580	5.1249
	Dover Score	0.9105	0.9380

显下降至 -2.6195，表明 prompt 重写并未破坏视频结构，反而在一定程度上增强了其时序协调性。此外，Dover Score 提升至 0.9380，说明整体生成质量有所增强，尽管 Aesthetic Score 小幅下降（从 5.1580 降至 5.1249），但整体表现更加均衡。

表 4-3 P2P 不同 Values 对比

指标名称	Values = 2	Values = 4	Values = 6
Pick Score	0.2431	0.2675	0.2859
ViCLIP Text Alignment	21.4732	21.0356	20.4187
CLIP Text Alignment	26.9124	27.6450	28.3846

P2P 不同 Values 对比在 Prompt-to-Prompt 编辑中不同 value 设置对语义一致性的影响趋势。随着 value 从 2 增加至 6，Pick Score 呈持续上升趋势（0.2431 → 0.2675 → 0.2859），CLIP Text Alignment 亦同步提升（26.9124 → 27.6450 → 28.3846），表明语义替换强度的增强有效提升了帧级图文匹配能力。相反，ViCLIP Text Alignment 出现下滑（由 21.4732 降至 20.4187），反映出语义注入过强可能破坏视频整体的一致性，导致模型对全局语义描述的匹配能力下降。该现象表明，在 Prompt-to-Prompt 编辑中，value 的设置需在“结构保持”与“语义增强”之间权衡。

综合来看，Video-P2P 相较于 GLP 拥有更高的图像质量与局部语义表达能力，而引入 LLM-edit 机制可进一步提升全局语义连贯性与结构一致性。Prompt-to-Prompt 的 value 控制实验表明：适度增强语义注入强度有助于提升语义匹配，但 value 设置过高可能削弱视频整体一致性。因此，在实际应用中，建议根据任务需求灵活选择编辑方法与参数组合，以实现语义精度与结构保真的双重目标。

第 5 章　研究结论与展望

根据上述实验结果与分析，本文提出的基于 Prompt-to-Prompt 的编辑方法在视频生成任务中展现出较强的语义控制能力与结构保持能力，尤其在语义一致性和生成质量方面表现稳定。通过对比 GLP 与 Video-P2P、引入 LLM-edit 编辑策略以及不同 values 参数下的变化趋势，实验明确验证了：适度增强语义注入强度有助于提升帧级语义表达准确性，而过高的 value 则可能破坏视频整体语义一致性；此外，LLM-edit 策略能够有效提升全局语义匹配能力并带来一定的结构稳定性增益。

Video-P2P 在 Aesthetic Score 与 CLIP Text Alignment 上的领先，说明其在视觉质量和局部语义精度方面更优；而 GLP 则在 ViCLIP Text Alignment 与 Motion Alignment 上更具优势，体现了其在保持视频整体结构和全局语义上的稳健性。LLM-edit 策略引入后，ViCLIP Text Alignment 显著提升，CLIP Consistency 与 Dover Score 均同步上升，表明语言重写可有效优化语义驱动编辑过程。同时，Motion Alignment 的下降显示结构保持能力增强，验证了语言建模与潜变量编辑的协同潜力。

Prompt-to-Prompt 方法具备良好的编辑扩展性和控制灵活性，能够在保持生成结构稳定性的基础上，引导模型实现多样化的语义变化。这种方法不仅适用于静态图像编辑，也已在本研究中成功迁移至视频领域，并通过少量参数调整实现了高质量的编辑效果。

未来的研究可进一步拓展：如引入多模态联合引导，如融合音频或动作线索以增强生成时序逻辑；二是研究跨帧一致性更强的 attention 控制策略，缓解高 value 设定下的结构扰动问题；三是结合更强的语言理解模型（如 GPT-4）实现高层语义概括能力，提升编辑可控性与通用性。通过持续优化 prompt 结构、扩散机制与注意力注入方式，Prompt-to-Prompt 视频编辑将在影视后期、虚拟角色生成与教育可视化等领域发挥更广泛的价值。

参考文献

[1] Junyu Gao X Y Y H, Kunlin Yang. Unity in diversity: Video editing via gradient-latent purification [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2025.

[2] Tumanyan N, Geyer M, Bagon S, et al. Plug-and-play diffusion features for text-driven image-to-image translation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 1921-1930.

[3] Liu S, Zhang Y, Li W, et al. Video-p2p: Video editing with cross-attention control [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8599-8608.

[4] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [J]. Advances in neural information processing systems, 2014, 27.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[6] Sun R, Zhang Y, Shah T, et al. From sora what we can see: A survey of text-to-video generation [J]. arXiv preprint arXiv:2405.10674, 2024.

[7] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[8] Li W, Xu X, Xiao X, et al. Upainting: Unified text-to-image diffusion generation with cross-modal guidance [J]. arXiv preprint arXiv:2210.16031, 2022.

[9] Balaji Y, Nah S, Huang X, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers [J]. arXiv preprint arXiv:2211.01324, 2022.

[10] Feng Z, Zhang Z, Yu X, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10135-10145.

[11] Kim G, Kwon T, Ye J C. Diffusionclip: Text-guided diffusion models for robust image manipulation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 2426-2435.

[12] Saharia C, Chan W, Chang H, et al. Palette: Image-to-image diffusion models [C]//ACM SIGGRAPH 2022 conference proceedings. 2022: 1-10.

[13] Brack M, Friedrich F, Kornmeier K, et al. Ledits++: Limitless image editing using text-to-image models [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 8861-8870.

[14] Ho J, Salimans T, Gritsenko A, et al. Video diffusion models [J]. Advances in Neural Information Processing Systems, 2022, 35: 8633-8646.

[15] Blattmann A, Rombach R, Ling H, et al. Align your latents: High-resolution video synthesis with latent diffusion models [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 22563-22575.

[16] Sun W, Tu R C, Liao J, et al. Diffusion model-based video editing: A survey [J]. arXiv preprint arXiv:2407.07111, 2024.

[17] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using

nonequilibrium thermodynamics [C]//International conference on machine learning. pmlr, 2015: 2256-2265.

[18] Kingma D P, Welling M, et al. Auto-encoding variational bayes [M]. Banff, Canada, 2013.

[19] Hertz A, Aberman K, Cohen-Or D. Delta denoising score [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 2328-2337.

[20] Chai W, Guo X, Wang G, et al. Stablevideo: Text-driven consistency-aware diffusion video editing [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 23040-23050.

[21] Wang W, Jiang Y, Xie K, et al. Zero-shot video editing using off-the-shelf image diffusion models [J]. arXiv preprint arXiv:2303.17599, 2023.

[22] Cao M, Wang X, Qi Z, et al. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 22560-22570.

[23] Mokady R, Hertz A, Aberman K, et al. Null-text inversion for editing real images using guided diffusion models [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6038-6047.

[24] Qi C, Cun X, Zhang Y, et al. Fatezero: Fusing attentions for zero-shot text-based video editing [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 15932-15942.

[25] Shin C, Kim H, Lee C H, et al. Edit-a-video: Single video editing with object-aware consistency [C]//Asian Conference on Machine Learning. PMLR, 2024: 1215-1230.

[26] Ho J, Salimans T. Classifier-free diffusion guidance [J]. arXiv preprint arXiv:2207.12598, 2022.

[27] Zhao Y, Xie E, Hong L, et al. Make-a-protagonist: Generic video editing with an ensemble of experts [J]. arXiv preprint arXiv:2305.08850, 2023.

[28] Ren T, Liu S, Zeng A, et al. Grounded sam: Assembling open-world models for diverse visual tasks [J]. arXiv preprint arXiv:2401.14159, 2024.

[29] Cheng H K, Schwing A G. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model [C]//European Conference on Computer Vision. Springer, 2022: 640-658.

[30] Bai J, He T, Wang Y, et al. Uniedit: A unified tuning-free framework for video motion and appearance editing [J]. arXiv preprint arXiv:2402.13185, 2024.

[31] Ku M, Wei C, Ren W, et al. Anyv2v: A plug-and-play framework for any video-to-video editing tasks [J]. arXiv preprint arXiv:2403.14468, 2024.

[32] Wang Y, Chen X, Ma X, et al. Lavie: High-quality video generation with cascaded latent diffusion models [J]. International Journal of Computer Vision, 2024: 1-20.

[33] Brooks T, Holynski A, Efros A A. Instructpix2pix: Learning to follow image editing instructions [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 18392-18402.

[34] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style [J]. arXiv preprint arXiv:1508.06576, 2015.

[35] Wang Q, Bai X, Wang H, et al. Instantid: Zero-shot identity-preserving generation in seconds [J]. arXiv preprint arXiv:2401.07519, 2024.

[36] Chen X, Huang L, Liu Y, et al. Anydoor: Zero-shot object-level image customization [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 6593-6602.

[37] Geyer M, Bar-Tal O, Bagon S, et al. Tokenflow: Consistent diffusion features for consistent video editing [J]. arXiv preprint arXiv:2307.10373, 2023.

[38] Cong Y, Xu M, Simon C, et al. Flatten: optical flow-guided attention for consistent text-to-video editing [J]. arXiv preprint arXiv:2310.05922, 2023.

[39] Yang S, Zhou Y, Liu Z, et al. Fresco: Spatial-temporal correspondence for zero-shot video translation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8703-8712.

[40] Oja E. Principal components, minor components, and linear neural networks [J]. Neural networks, 1992, 5(6): 927-935.

[41] Hatem G, Zeidan J, Goossens M, et al. Normality testing methods and the importance of skewness and kurtosis in statistical analysis [J]. BAU Journal-Science and Technology, 2022, 3(2): 7.

[42] Hertz A, Mokady R, Tenenbaum J, et al. Prompt-to-prompt image editing with cross attention control [J]. arXiv preprint arXiv:2208.01626, 2022.

[43] Schuhmann C, Vencu R, Beaumont R, et al. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs [J]. arXiv preprint arXiv:2111.02114, 2021.

[44] Wu H, Zhang E, Liao L, et al. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 20144-20154.

[45] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[46] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.

[47] Wang Y, He Y, Li Y, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation [J]. arXiv preprint arXiv:2307.06942, 2023.

[48] Kirstain Y, Polyak A, Singer U, et al. Pick-a-pic: An open dataset of user preferences for text-to-image generation [J]. Advances in Neural Information Processing Systems, 2023, 36: 36652-36663.

[49] Xu H, Zhang J, Cai J, et al. Gmflow: Learning optical flow via global matching [C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 8121-8130.

[50] Song J, Meng C, Ermon S. Denoising diffusion implicit models [J]. arXiv preprint arXiv:2010.02502, 2020.

致 谢

时间飞逝，本科四年的学习生活即将结束。在玉泉路的四年，我也在不断探索、思考，在四季流转中有所收获。

在此，衷心感谢中国科学院大学在本科四年中给予我的培养。无论是课堂上的严谨治学，还是课题实践中的悉心指导，都让我受益匪浅。这四年不仅仅是知识的积累，更是思考能力、科研素养的培养，让我有机会不断挑战自我，拓展眼界。

感谢高君宇老师在我科研学习上的指导，从文献调研到后续科研实践进展，引我在科研入门。同时，非常感谢我的学业导师徐常胜老师、钱胜胜老师，有幸得到老师们的教导，和老师们的交集是莫大的缘分，老师们榜样的力量，给予我前行的动力。

感谢国科大的同学们。同窗的时光十分宝贵，这些日子将成为我人生中最值得珍藏的回忆。同时，我也要感谢一路上给予我支持与陪伴的朋友们。你们在我迷茫时给予鼓励，在我困顿时给予理解，让我在不断追寻梦想的路上，始终感受到温暖与力量。

感谢我的父母。感谢你们的支持和鼓励，给了我追求理想的勇气。你们的理解、包容和付出，是我前行的动力。

在本次毕业设计中，我围绕特征注入与扩散模型技术，深入研究了视频编辑的关键问题。这一过程不仅提升了我的科研技能，也锻炼了我独立解决问题的能力。尤其是在调试实验细节、分析数据结果、撰写论文报告的过程中，我对科研的严谨性有了更深刻的理解，也更加体会到理论与实践结合的重要性。

未来，我将走向下一场相遇，下一次人生的交汇。在国科大的学习，老师们的教诲与同学们的友谊，将成为我人生路上宝贵的财富。我将继续努力，迎接挑战，在自己的专业领域中不断探索。

谨以此文，向所有在我成长道路上给予帮助、鼓励与支持的人，致以最诚挚的感谢。

2025 年 6 月