上海大学
SHANGHAI UNIVERSITY

本科毕业论文（设计）

UNDERGRADUATE THESIS (PROJECT)

题　　目：基于语言大模型微调的短视频隐喻研究

学　　院：通信与信息工程学院

专　　业：电子信息工程

学　　号：21122164

学生姓名：谢雨梦

指导教师：吴汉舟

起讫日期：2025.02-2025.06

姓　　名：谢雨梦　　学号：21122164

论文题目：基于语言大模型微调的短视频隐喻研究

原 创 性 声 明

本人声明：所呈交的论文是本人在指导教师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名：谢雨梦　日 期：2025.5.14

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密的论文在解密后应遵守此规定）

签 名：谢雨梦　指导教师签名：　日期：2025.5.14

摘　要

隐喻是一种常见的表现手法，指通过将一个概念域映射到另一个概念域来理解抽象或复杂概念。近年来，虽然文本形式隐喻的识别与生成已得到广泛研究，其他形式隐喻仍处于初步探索阶段。如今短视频凭借其时长短、信息丰富的特点，成为人们社交娱乐和表达观点的主流媒介。由于短视频通常融合网络流行梗、视觉隐喻等多种隐喻性表达，研究短视频中的隐喻具有重要现实意义。

尽管当前主流的大语言模型在通用任务中表现出色，但在处理隐喻，尤其是融合视觉隐喻的短视频等复杂内容时仍存在一定局限。为此，本文聚焦于短视频隐喻生成这一新兴任务，构建短视频隐喻数据集对大模型进行微调，以提升其对短视频隐喻的理解和生成能力。本文的主要工作如下：

为了促进短视频隐喻生成任务的发展，本文构建了一个中文短视频隐喻数据集，包括 898 个短视频及其对应的隐喻文本标签，涵盖五种常见视频内容类型，为后续研究提供了高质量的研究语料。基于该短视频隐喻数据集，本文在最新的多模态大语言模型上展开七种微调方法的对比实验，最终成功获得一个能够有效理解并生成短视频隐喻内容的聊天助手。实验证明，该模型在多个文本生成评价指标上表现出较强的隐喻生成能力，能够提升大模型对短视频隐喻的捕捉和理解能力。

关键词：隐喻；短视频；大语言模型；微调

ABSTRACT

Metaphor is a common representation that refers to the understanding of abstract or complex concepts by mapping one conceptual domain to another. In recent years, while the identification and generation of textual metaphors have been widely studied, other forms of metaphors are still in the preliminary exploration stage. Nowadays, short videos have become the mainstream medium for people to entertain and express their opinions due to their short duration and rich information. Since short videos usually integrate various metaphorical expressions such as popular Internet stems and visual metaphors, it is of great practical significance to study metaphors in short videos.

Although mainstream large language models perform well in general tasks, they still have limitations when dealing with metaphors, especially in visually rich short videos. To this end, this thesis focuses on the emerging task of short-video metaphor generation, constructs a short-video metaphor dataset, and performs instruction tuning on the large language model to improve its ability to understand and generate short-video metaphors. The main work of this thesis is as follows:

In order to facilitate the development of short video metaphor generation task, this thesis constructs a Chinese short video metaphor dataset, including 898 short videos and their corresponding metaphor text labels, covering five common video content types, which provides a high-quality research corpus for the subsequent research. Based on this short video metaphor dataset, this thesis conducts comparison experiments of seven fine-tuning methods on the latest multimodal large language model, and finally succeeds in obtaining a chat assistant that can effectively understand and generate short video metaphorical content. The experiments prove that the model exhibits strong metaphor generation ability on multiple text generation evaluation metrics, which can improve the ability of the big model to capture and understand short video metaphors.

Keywords: Metaphor; Short video; Large Language model; Fine-tuning

目　录

第一章　绪论

1.1　研究背景及意义

隐喻是文学中最常见、最富表现力的手法之一[1]，通过将已有的、熟悉的概念映射至陌生的、抽象的概念[2]，实现了对复杂事物的理解与体验。随着对隐喻的深入研究，这种跨概念域映射的表达方式也逐渐从文字延伸至视觉媒介，通过图像、视频等形式传达复杂信息的视觉隐喻越来越多地被应用到实际生活中。在广告创意领域中[3]，视觉隐喻被证实能够增强受众对品牌的积极认知[4]。在教育传播领域，合理运用动画、图像中的视觉隐喻有助于提升学习效果[5]。此外，视觉隐喻在政治宣传和社会运动中也具有重要作用，能够通过隐晦而富有象征意义的视觉表达引导公众的情绪和认知[6]。

近年来，随着互联网的飞速发展，短视频已成为全球主流的媒介形式，抖音、快手、TikTok 等平台迅速兴起。用户通过视觉、字幕、音频等多种载体快速获取和传递信息。短视频创作者也通过视觉符号、网络“梗”等形式隐喻增强视频内容的感染力。在此背景下，短视频隐喻作为一种融合图像、音频与文本等多种符号系统的隐喻表达形式，逐渐成为值得关注的新兴研究领域。

与此同时，大语言模型及人工智能助手的快速发展，使得人机交互方式正在发生深刻改变[7]。新一代人工智能助手，如 ChatGPT、DeepSeek 等，具备接收图像、文本、音频等多模态输入的能力，能够综合处理不同模态的信息，在跨模态推理、自然语言生成、上下文理解等方面表现出色。依托庞大的预训练数据和先进的指令微调技术，这些模型在通用任务与标准数据集上表现出色，展现出强大的语义理解能力和推理能力，具备较强的泛化与迁移能力。然而，尽管新一代人工智能助手在标准任务上表现卓越，其对复杂隐喻现象的解析能力仍然有限，特别是在需要理解图像、视频乃至短视频等内容中隐晦、间接或文化语境相关表达时，模型的认知与推理能力仍存在明显不足。

在此背景下，构建一个短视频隐喻数据集来填补当前短视频隐喻研究资源匮乏的空白具有重要意义。基于该数据集，本文进一步探索利用大语言模型进行微调的方法，训练出能够生成短视频隐喻内容的人工智能助手。通过结合视觉、文本、音频

等多模态特征，微调后的模型将具备更高水平的跨模态推理能力与复杂语义解析能力，从而有效提升对短视频隐喻现象的处理性能。

本研究不仅有助于推动短视频隐喻在人工智能领域的应用研究，还为大语言模型在复杂认知任务中的扩展应用提供了新的范例。此外，训练得到的人工智能助手可广泛应用于短视频内容理解和创作、文化传播与教育等多个实际场景，具有重要的理论价值与实践意义。

1.2 国内外研究现状

1.2.1 视觉隐喻

随着自然语言处理与认知计算的快速发展，隐喻检测与生成等任务受到越来越多研究者的关注。大量句子级和词级数据集的发布[8] [9] [10]极大推动了文本隐喻理解与建模的进展。然而，相较于已经相对成熟的文本隐喻研究，视觉隐喻等其他模态的隐喻研究[11]起步较晚。

2019 年 Petridis 和 Chilton 等人[12]评估了多种视觉隐喻解释理论。结果表明，在缺乏解释性文字的情况下，人类对视觉隐喻的正确理解率仅为 41.3%，凸显了视觉隐喻本身的复杂性和模糊性。为了推进视觉隐喻的建模与理解，Indurkhya 和 Ojha 等人[13]强调了源和目标图像在颜色、形状等方面之间的感知相似性在隐喻理解和创造性解释中的重要作用，为视觉隐喻建模提供了理论支撑。

2021 年 Achlioptas 等人[14]提出了 ARTEMIS 数据集，包含了来自 WikiArt 的约 80,000 幅艺术作品，标注了丰富的情感归因与解释性文本，其中蕴含了多种视觉隐喻与比喻表达。该数据集被用于训练图像字幕生成系统，使其能够表达视觉刺激所传达的情绪与内涵。Chakrabarty 等人[15]进一步提出了从文本隐喻生成视觉隐喻图像的任务，并构建了名为 HAIVMet 的数据集，包含由 DALL-E2 生成的 6476 个视觉隐喻图像[16]。此外，越来越多研究开始关注真实用户生成内容中的隐喻实践。张等人[17]构建了 MultiMet 数据集，系统性地收集并标注了来自 Twitter 平台的大量英语隐喻句子，涵盖多个主题领域，为研究语言模型在复杂社交语境中识别和理解隐喻提供了重要资源。Hwang 等人[18]发布了 MemeCap 数据集，专注于模因 (meme) 中的多模态隐喻表达。模因是一种社交媒体平台上流行的交流形式，通过文本和图像的结合传达隐喻，从而表达出大众的情感和想法[19]。MemeCap 数据集包含图像与文本之间隐喻关联的配对信息，揭示了图文结合在隐喻传播过程中的独特机制。这些数

据集不仅丰富了隐喻研究的语料资源，也推动了语言模型在社交媒体、多模态语境中的应用探索，为更广泛的社会认知研究和情绪传播分析提供了有力支撑。

在系统性任务建构方面，2023 年 Akula 等人[20]首次提出了用于评估视觉隐喻理解能力的一组基准任务，并指出主流视觉模型在处理此类任务时性能明显不足。为进一步推动视觉隐喻的研究，同时期，Yosef 等人[21]发布了一个包含隐喻、比喻与成语图像配对的多模态数据集，并构建了两个新颖的基准任务：多模态修辞语言检测与检索，旨在测试模型对象征性语言表达的感知与理解能力。实验结果显示，现有模型在此类任务中的表现远低于人类基准，强调了视觉隐喻建模在多模态学习中的挑战性与研究空间。

进一步地，2024 年 Kalarani 等人[22]则将研究焦点延伸至视频语境下的隐喻表达，提出了一项新颖的视频隐喻字幕生成任务，旨在通过生成具隐喻性的描述来提升机器对视频中深层语义的建模能力。为此，文章构建了一个手工标注的视频隐喻数据集 VMC Dataset，包含 705 个广告类视频及 2115 条隐喻字幕，并设计了一个固定句式模板来统一生成格式。除此之外，为提升模型的隐喻理解与生成能力，文章提出了一个低资源的模型架构 GIT-LLaVA：使用冻结的视频基础模型提取视频表示，通过一个多层感知器映射网络将表示送入大语言模型，实现隐喻字幕的自动生成。

综上所述，近年来针对视觉隐喻的研究不断扩展，从图像隐喻识别到模因、广告等多模态语境下的理解与生成，相关数据集和基准任务逐渐丰富，极大推动了视觉隐喻建模的发展。然而，现有研究仍主要集中于静态图像与配套文本之间的隐喻表达，尚缺乏对动态视频内容中隐喻的系统生成方法，尤其在中文语境下的研究更为稀缺。

1.2.2 大语言模型

自 2017 年 Transformer 架构[23]被提出以来，基于注意力机制的预训练语言模型逐渐成为自然语言处理领域的主流范式。通过引入多头自注意力机制，显著提升了模型对长距离依赖的建模能力，为后续大规模语言模型的快速发展奠定了基础。之后两年内，涌现出 BERT[24]（3 亿参数）和 GPT-2（15 亿参数）等重要模型，验证了预训练-微调范式的巨大潜力。更多策略如基于人类反馈的强化学习、代码预训练、指令微调等开始出现，被用于进一步提高推理能力和任务泛化。2020 年至今大模型呈现出爆发式增长态势，微软 Turing-NLG（170 亿参数）和 Google T5（120 亿参数）在

2020 年率先突破百亿参数规模，2021 年更迎来技术突破的集中涌现，包括 NVIDIA MT-NLG（5300 亿参数）、Google Switch Transformer（16000 亿参数）等语言模型，以及 DALL-E 2（6.4 亿参数）、Florence（120 亿参数）等多模态模型。2022 年产业界全面跟进，OpenAI GPT-3（1750 亿参数）、百度文心（2600 亿参数）和华为盘古（10850 亿参数）等大模型相继发布，展现出强大的语言理解和生成能力。2023 年，多模态能力取得重大突破，GPT-4（预估 10000 亿参数）和阿里巴巴 M6（100000 亿参数）等模型实现了跨模态理解与内容生成。1950-2023 年间大模型发展历程如图1.1所示。

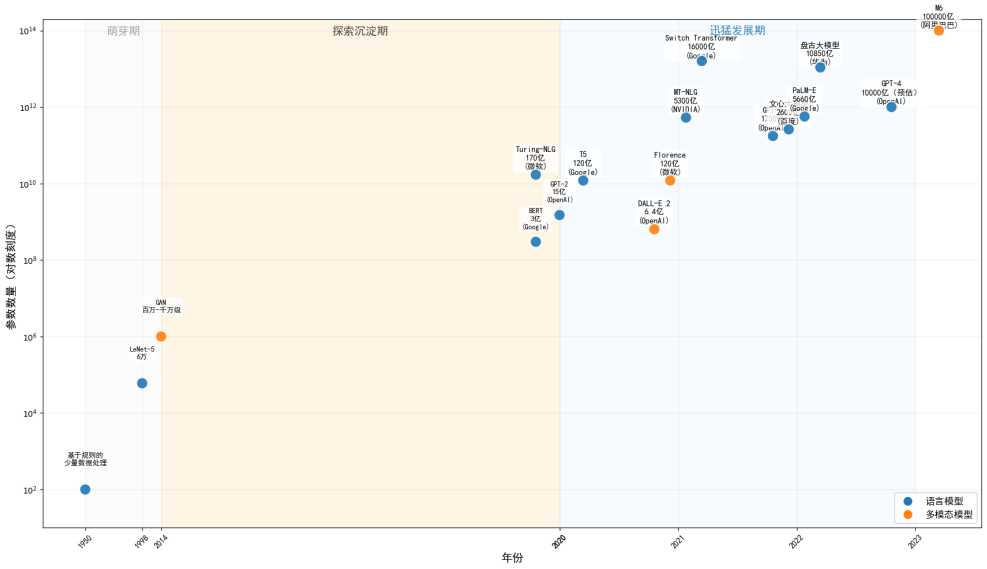


图 1.1 大模型参数规模发展历程

近两年，在大模型不断演进的过程中出现了如 DeepSeek 等新一代开源大模型。这些模型在开源性、推理能力、多语言支持等方面都有显著提升。DeepSeek-V2 通过引入更大规模的训练数据、更高效的模型架构以及多种推理优化技术，使得模型在开放式问答、长文本推理等任务上达到了新的水平。同时，DeepSeek 家族在兼顾推理准确性和响应速度方面，也体现出较强的工程落地潜力，为多种实际应用场景提供了新的选择。

在此基础上，大模型如何在具体行业与任务场景中实现有效部署，成为近年来的重要研究议题，具体部署思路可分为三种。一是基于高质量领域语料的预训练。通过收集特定领域的大规模高质量语料，对语言模型进行预训练，可以提升其在特定任务上的表现。例如，Li 等人[25]针对中文名词性隐喻生成问题，设计了基于隐喻识别机制的多任务学习框架，并使用大规模类比和诗歌语料进行预训练，从而在隐

喻生成中提升了句子的隐喻性与连贯性。二是利用知识库检索增强模型。通过知识图谱、词典等外部知识库与模型的动态交互，补充模型对低频概念、专业术语和事实性知识的理解。张鹤译、王鑫等人[26]通过收集中医领域数据形成知识库，并将其与大语言模型结合，从而实现了大模型在中医垂直领域中的有效部署应用。Youn 和 Tagkopoulos 等人[27]提出的知识图谱语言模型（KGLM）架构，在语言模型中引入了新的实体和关系嵌入层，使模型能够学习知识图谱的结构信息。通过对语言模型进行进一步的预训练，并在标准微调阶段使用从知识图谱中提取的三元组，KGLM 在链接预测任务上达到了新的性能水平。Wang 等人[28] 针对隐喻理解任务，提出将 FrameNet[29]和 ConceptNet[30]等语义知识库结合进 Prompt 工程，显著提升了隐喻解释的合理性。三是微调通用模型以适应特定任务。在通用预训练模型的基础上，利用少量优质的特定任务数据进行微调使模型适应特定的应用场景。Stove 等人[31] 提出了隐喻性释义生成模型，通过在 BART 模型[32]上进行微调，实现了从字面句子生成隐喻句子，且该模型生成效果优于人工生成。Singh 等人[33]通过结合情感标签与语义约束对 GPT-2 进行微调，训练出能够根据上下文生成具有特定情感色彩与修辞风格的对话模型。王婷、王娜等人[34]通过微调 ChatGLM 等 4 种大语言模型，训练出了具备语义分析的农业知识问答大模型，用于提升果蔬农技知识服务的智能化水平。

越来越多的研究开始关注大语言模型在特定应用场景中的实际落地与成效，反映出其在社会环境中的高度可塑性。以教育领域为例，已有学者尝试利用 ChatGPT 等生成式预训练模型辅助大学生进行论文写作。通过多轮互动式对话，学生能够获得针对文章结构、逻辑衔接与语言表达等方面的个性化建议，从而有效提升写作质量与表达能力。潘雪峰、王超等人[35]探讨了 ChatGPT 在健康谣言鉴别任务中的应用效果。隐喻识别与生成方面，Xu 等人[36] 针对心理健康语料对 T5 模型进行了微调，使其具备识别语言中隐含隐喻风险信号的能力，为智能心理健康评估系统提供技术支持。

1.3 本文研究内容

由于隐喻本身具有抽象性和语境依赖性，其在视频环境中的识别与建模任务面临更为复杂的挑战。在短视频这一以信息密度高且语义碎片化为特征的媒介中，隐喻往往需要视觉线索、语言表达以及背景知识的深度融合，进一步加剧了语义理解与建模的难度。近年来，针对视频隐喻，特别是短视频隐喻的系统性研究仍较为稀

缺，相关隐喻数据集严重不足，且尚缺乏具备广泛适用性与可拓展性的模型框架与实际应用系统。

受 Kalarani 等人提出的视频隐喻字幕生成框架[22] 启发，本文以大语言模型为技术核心，聚焦于短视频语境下隐喻生成任务与对话式表达机制，旨在探索基于预训练语言模型的视频隐喻理解与生成系统的可行性与有效性。具体而言，本文的研究工作包括以下三个方面：

（1）构建高质量中文短视频隐喻数据集。为填补中文短视频语境下隐喻语料的空白，本文从多个中文短视频平台中筛选具有潜在隐喻表达的短视频，由 13 个具备良好语言表达能力和隐喻理解能力的标注者结合视觉信息与文本字幕进行人工标注，并将视频分为讽刺类、诙谐类、生活态度类、硬核类、自嘲幽默类五种类型。之后将标注好的隐喻标签经过大模型格式统一和人工校审，以增强数据的有效性。由此，构建了一个面向中文短视频的隐喻数据集，其具备良好的通用性与可扩展性。

（2）引入大语言模型微调方法构建短视频隐喻模型。针对多模态输入与隐喻语言建模之间的融合挑战，本文设计了一个基于预训练大语言模型进行指令微调的短视频隐喻生成模型。该模型结合视频信息与输入提示，通过少量高质量数据对模型进行微调，使其具备理解视频隐喻语境、生成具象隐喻文本以及提供解释性对话内容的能力。实验表明，该方法在隐喻准确性、语言流畅性与交互合理性方面均优于通用生成模型。

（3）部署短视频隐喻聊天助手系统。为了验证模型在实际场景中的应用潜力，本文进一步开发并部署了一个可交互的短视频隐喻聊天助手系统。该系统支持用户上传视频，并通过模型自动生成隐喻式解读文本与对话回应，具有切换视频隐喻微调模型，查看历史聊天记录等多项功能。该助手系统不仅为用户提供可参考性的视频隐喻内容理解，也为内容创作者提供了创意激发。

1.4 本文组织结构

本文的研究内容共分为五章，具体内容安排如下：

第一章为绪论部分。本章确立了论文的研究基调，通过系统阐述视频隐喻生成任务的研究背景和现实价值，揭示了该任务的独特挑战性与前沿性，并指出将大语言模型引入该任务的可行性与创新潜力。在此基础上，回顾并梳理了视觉隐喻识别与生成的相关研究进展，以及近年来大语言模型在生成任务中的典型应用与发展趋

势，从而为本研究提供了坚实的理论支撑与技术参考。

第二章为相关理论与技术的综述与分析。本章从技术层面出发，聚焦于本研究所依赖的关键理论与核心方法，主要涵盖对新兴多模态大语言模型、大模型参数微调策略，以及文本生成质量评价指标的介绍。通过对上述技术理论的深入分析与比较，本章不仅为后续的模型设计与实验实现提供理论依据，也为方案选择与实验优化提供参考路径。

第三章围绕短视频隐喻数据集的构建与评估展开。本章详细阐述了面向本研究任务构建高质量语料的整体流程，内容包括数据收集、大模型统一格式和人工校审三大阶段。在数据收集过程中，明确了语料选取的语义覆盖范围，确保所构建数据集具有多样性与代表性。然后通过多位具备较强语言理解的标注人员进行人工标注隐喻标签。在数据增强方面，包括大数据统一格式和人工校审两种增强策略，以提升数据集的语义丰富度与任务适配性。此外，本章还对所构建数据集在数据分布均衡性、语义多样性、标注准确性与一致性以及格式规范性四个维度进行了分析与评估，为后续模型训练提供可靠数据基础。

第四章讲述了大语言模型微调实验与性能分析。本章主要围绕不同参数微调策略在短视频隐喻生成任务中的性能效果展开研究。首先，介绍了实验的整体设计方案，包括基座模型的选取依据，微调方法的实施流程及实验结果分析。随后，通过对比多种微调策略下模型在多种文本生成指标上的实验结果，评估其在该任务中的性能表现。此外，本章还进一步探讨了模型在实际应用场景中的落地能力，展示了将微调后模型部署至云端环境所构建的短视频隐喻生成对话系统，验证了模型方案的实用性与技术可行性。

最后为结论与展望。本章总结了本文在数据集构建、模型设计与系统实现等方面的主要贡献。在此基础上，本文也指出了当前工作的局限性，包括数据规模和语境覆盖范围有限、大模型在当前任务下应用能力不足等等。为此，未来可从引入知识增强机制以及探索新型生成范式等方面入手，进一步提升模型在隐喻生成任务中的表现。

最后，本文的整体研究框架如图 1.2 所示。

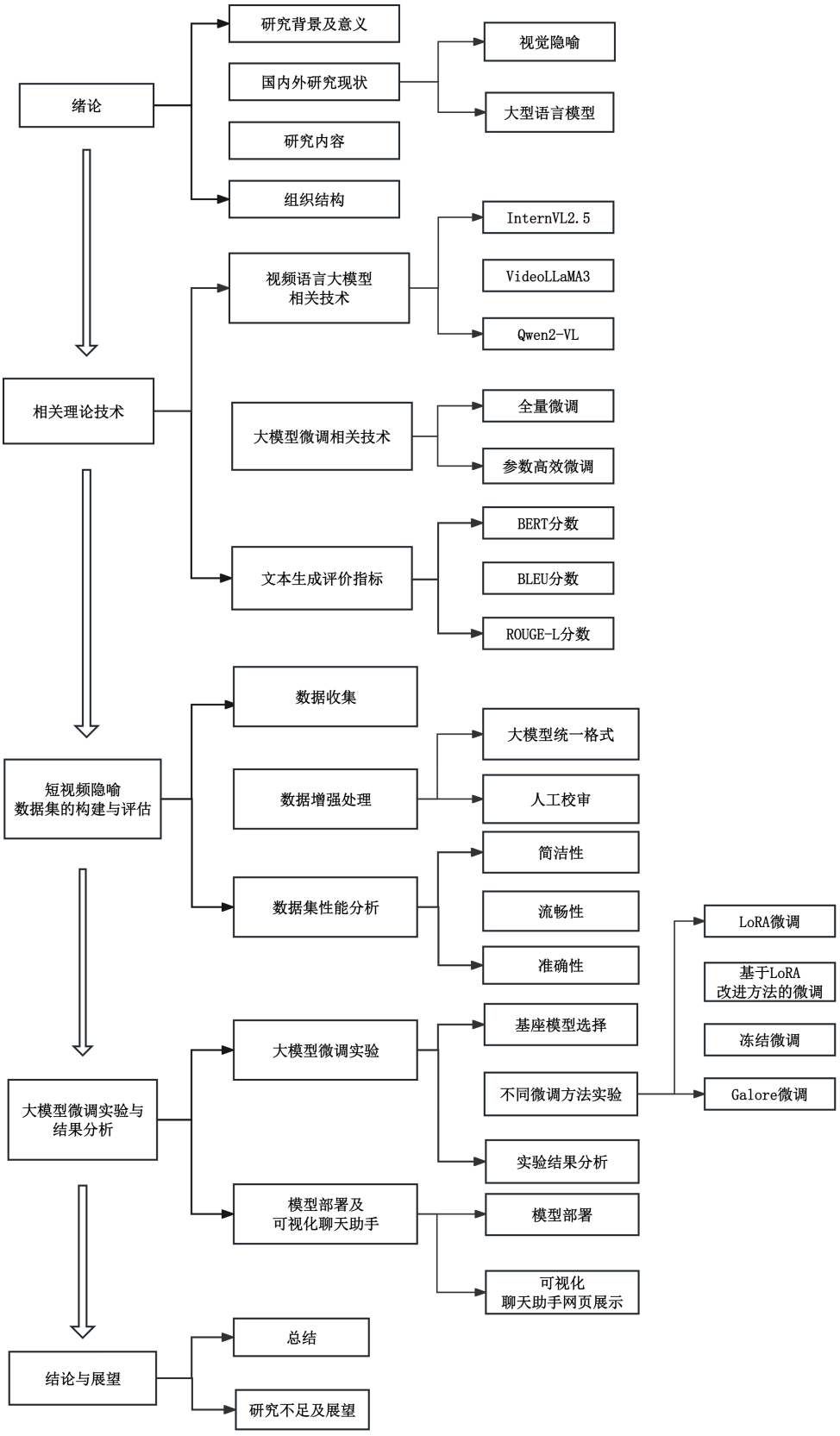


图 1.2 本文组织结构框架图

第二章　相关理论与技术

2.1　多模态大语言模型相关技术

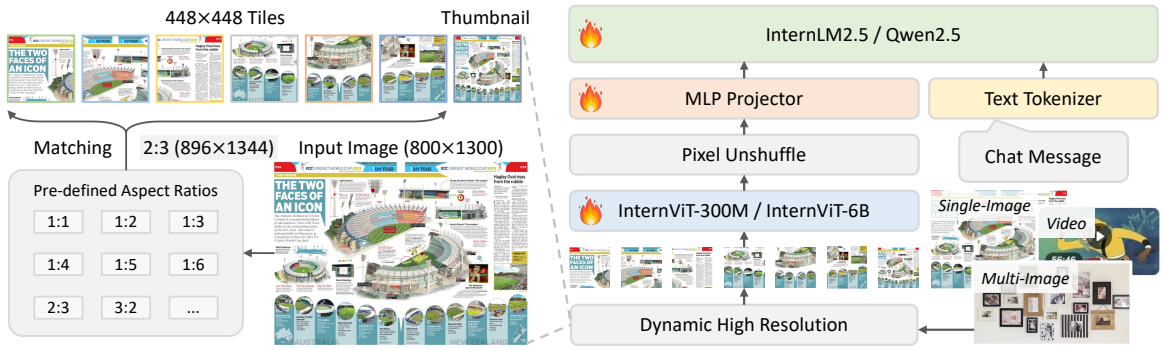
多模态大语言模型（Multimodal Large Language Models, MLLMs）在大语言模型（Large Language Models，LLMs）强大语言处理能力的基础上，进一步支持对图像、音频乃至视频等多种模态的信息处理与理解。MLLMs 通常由预训练模态编码器、预训练 LLM 和模态融合与对齐模块组成。模态编码器将原始信息压缩成 LLM 能够理解的紧凑表示，例如图像或视频输入通过视觉编码器（如 Vision Transformer）提取视觉特征。不同模态的信息经过融合与对齐模块实现统一的建模表示，并输入至 LLM 完成理解和推理处理。下面介绍三种最新的、相对成熟的多模态大语言模型。

2.1.1　InternVL2.5

InternVL2.5 模型是由上海人工智能实验室团队 2024 年在文献[37]中提出的开源多模态大语言模型。在继承 InternVL 2.0 的基础架构的同时，InternVL2.5 模型针对多模态数据处理、训练策略以及模型扩展性等方面进行了系统性的优化和创新，显著提升了模型在图像、视频、文本等多模态任务中的表现。

为了显著提升多模态模型对高分辨率图像的表征能力与推理效率，InternVL 2.5 引入了一种结构化的动态分辨率切片策略，以缓解传统视觉编码模型在处理超大尺寸图像时因统一缩放所导致的空间信息损失问题。该策略的核心思想是依据输入图像的原始尺寸与长宽比，对图像进行分块处理，将其裁切为若干个固定大小的图像块（tiles），每块分辨率为 448×448。同时，模型保留并处理一个缩放后的整图（thumbnail），作为全局上下文补充，从而实现局部细节与整体语义的双重捕捉。

在模型架构设计方面，InternVL 2.5 采用 Vision Transformer（ViT）作为视觉编码器，通过多层感知器（Multilayer Perceptron, MLP）进行模态对齐，最终与大语言模型融合。为了适应不同规模场景下的多模态任务需求，InternVL 2.5 提供了两个不同规格的视觉编码器配置：InternViT-6B 和 InternViT-300M。其中，InternViT-6B 拥有 45 层，约 5.5 亿参数，适用于大规模 LLM，例如 26B、38B、78B 等。InternViT-300M 则为 InternViT-6B 的蒸馏版本，参数量为 3 亿，适用于中小规模 LLM，例如 1B、2B、4B、8B 等。其数据处理和框架结构如图2.1所示。



(a) Data Preprocessing　　(b) Model Architecture

图 2.1 InternVL2.5 **数据处理和模型框架**[37]

InternVL 2.5 使用了一种策略渐进性的 2.5 阶段训练范式，以实现模型在多模态对齐、特定任务适配以及泛化能力提升等多个维度的协同优化。具体而言，第一阶段为 MLP 预热阶段。在此阶段，冻结视觉编码器和语言模型，仅训练 MLP 投影层，以实现初步的模态对齐。该阶段使用较高的学习率，加速模型收敛。随后 1.5 阶段为 ViT 增量训练阶段，该阶段为可选模块。此阶段解冻视觉编码器和 MLP，针对特定领域的数据进行增量训练，以增强模型在特定任务中的表现。为防止灾难性遗忘，该阶段采用数据重放策略，将前一阶段的数据与新数据混合训练。阶段二为完整模型指令微调阶段。在此阶段解冻所有模型参数，在高质量的多模态指令数据集上进行训练，以提升模型的多模态理解和生成能力。该阶段强调数据质量控制，避免低质量数据对模型性能的负面影响。此外，InternVL 2.5 引入了具备高度工程实效性的渐进式扩展策略，即先使用较小规模的 LLM（如 20B）进行训练，优化视觉能力和模态对齐，然后将训练好的视觉编码器迁移到更大规模的 LLM（如 72B）中，从而大幅缩短大模型训练时间，提升迁移后的初始性能表现。因此，在训练更大的模型时，可以跳过阶段 1.5，因为早期阶段优化的 InternViT 模块被重复使用。这一策略避免了重复性的高资源训练流程，同时保证了大模型在引入高维视觉信息时的稳定性与一致性。

此外，InternVL 2.5 还扩展了对多图像和视频数据的支持。对于多图像数据集，模型将总图块数分配到每幅图像上，并标注辅助标签。对于视频数据，模型对每帧进行 448×448 的调整，并使用帧标签进行标注，从而实现对视频序列的有效建模。InternVL 2.5 在多个多模态任务如图像和视频分析、视觉问答、文档理解和多语言处理等中表现出色，特别是在多模态理解基准上取得了超过 70 的得分，超越了 ChatGPT-4o 和 Claude-3.5-Sonnet 等商业模型。

2.1.2 VideoLLaMA3

VideoLLaMA3 模型在 2025 年由文献[38]提出，专为图像和视频理解任务打造。模型基于 Qwen 2.5 架构，融合先进的视觉编码器（如 SigLip）与强大的语言生成能力，具备高效处理长视频序列、支持多语言的视频内容分析和视觉问答任务的能力。

VideoLLaMA3 在结构上引入了任意分辨率视觉标记化（Arbitrary Visual Tokenization, AVT）和差分帧剪枝器（Differential Frame Pruner, DiffFP）两项关键技术。AVT 使视觉编码器能够根据图像尺寸生成相应数量的视觉标记，而非固定数量，从而更好地捕捉图像中的细粒度细节。DiffFP 则通过比较相邻帧像素空间的差异，修剪冗余视频标记，提高视频处理效率，减少计算需求。

在训练方面，VideoLLaMA3 特别采用了以视觉为中心的训练范式，强调高质量图像-文本数据的重要性。为了支持高质量的训练数据，团队构建了包含 700 万图像-字幕对的 VL3Syn7M 数据集。该数据集通过长宽比过滤、美学评分过滤、文本-图像相似度计算、视觉特征聚类和图像重新标注等步骤，确保数据的多样性和质量，从而提升模型的学习效果。具体训练过程分为四个阶段。第一阶段是视觉编码器适配阶段，模型对视觉编码器进行调整，使其能够处理动态分辨率的图像输入。第二阶段是视觉语言预训练阶段，通过使用大规模图像-文本数据，包括场景图像、文档、图表，以及纯文本数据等对视觉编码器、投影器和大语言模型联合微调。第三阶段是多任务微调，通过图像-文本问答数据和视频-文本数据进行微调，以提升其在多种任务中的表现。最后是以视频为中心的微调，专注于提升模型在视频理解任务上的性能。训练流程如图2.2所示。

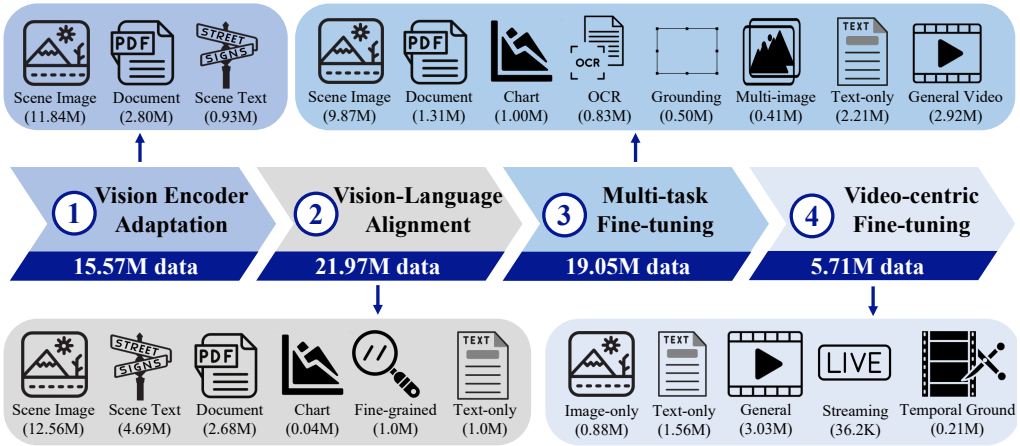


图 2.2 VideoLLaMA3 **训练流程**[38]

得益于上述设计，VideoLLaMA3 在多个图像和视频理解基准测试中取得了优异的表现，展示了其在多模态理解任务中的强大能力。此外，模型支持多语言生成，适用于视频内容分析、视觉问答、多模态应用等多种场景，具备广泛的应用潜力。

2.1.3 Qwen2-VL

Qwen2-VL 是阿里巴巴达摩院在文献[39]中提出的下一代多模态大语言模型，其核心目标是打破传统多模态系统在视觉处理方面的分辨率瓶颈，构建具备更强感知能力与推理能力的统一视觉-语言系统。，Qwen2-VL 延续之前的中文语言优势，在预训练过程中引入了大量中文图文指令数据，在中文图文理解与生成方面具备强大的泛化能力。

模型采用了“视觉-语言直接耦合”架构，以 ViT 作为视觉编码器，与大规模预训练语言模型 Qwen2 深度融合，不再依赖传统的中间投影头或轻量解码器。这种耦合结构不仅提升了视觉信息在语言生成过程中的直接参与度，也为视觉语言的对齐和共同建模提供了更大的灵活性和容量。同时，通过改进的 ViT 结构，模型不再依赖统一图像尺寸输入，而是可直接处理任意尺寸图像输入。这一特性依托于二维旋转位置编码（2D-RoPE）的引入，使视觉 token 之间的空间相对关系在不同分辨率条件下仍保持一致性，进而增强模型在图像尺度变换下的稳健性。同时，在视觉 token 生成和压缩过程中引入 MLP 层，对相邻 2×2 的 token 进行压缩处理，既保留了原始空间结构信息，又降低了下游自注意力模块的计算成本。模型框架如图2.3所示。

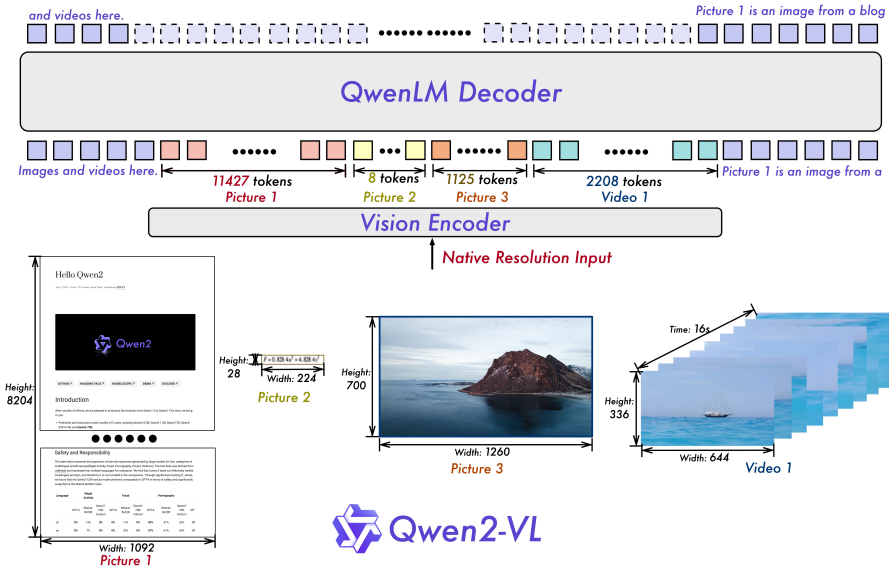


图 2.3 Qwen2-VL **模型框架**[39]

Qwen2-VL 在位置编码机制上进行了关键性创新，提出了多模态旋转位置编码（Multimodal Rotary Position Embedding, M-RoPE），通过对传统 RoPE 进行扩展，使其同时适用于文本、图像及视频等不同模态。在图像处理时，模型将二维位置信息（高度与宽度）分别嵌入，并进行旋转操作。在视频处理中，加入时间维度的旋转编码，使得模型能理解事件的时序关系，实现了对多帧之间事件依赖关系的捕捉，显著增强了对长视频中动态目标的建模能力。

如今 Qwen2-VL 训练策略采用了三阶段渐进式。第一阶段为视觉编码器预训练，选用大规模图文对数据（约 6000 亿 token），采用对比学习与跨模态对齐目标联合优化 ViT 表示能力。第二阶段为多任务联合训练，引入 8000 亿 token 的混合模态数据，涵盖图像问答、视频理解、图文推理等任务，构建统一的多模态推理框架。此阶段的训练目标包括语言生成、目标定位、多选分类等结构化任务，从而显著增强模型在泛化和零样本任务下的表现。第三阶段为指令微调阶段，模型在自然语言指令范式下进行优化，使其在多轮对话、多模态交互等任务中具有更高的实用性和可控性。模型训练流程如图2.4所示。

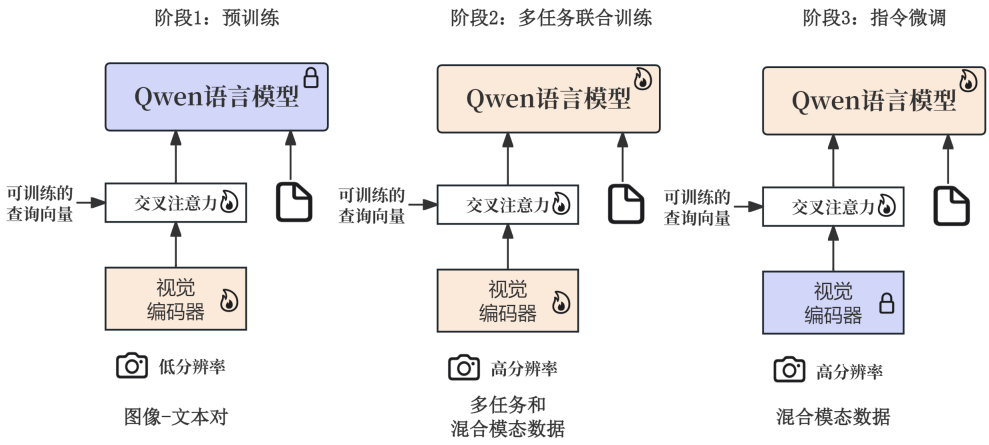


图 2.4 Qwen2-VL **训练流程**

Qwen2-VL 已在文档智能解析、工业质检、高阶内容审核、视频安防分析等多个场景中展开落地部署，并通过开源形式向学术界和开发者开放了基础模型权重与训练代码。该模型不仅填补了中文语境下的开源多模态大模型空白，也为多模态智能体、边缘视觉理解等下游技术提供了稳固的基础设施。总的来看，Qwen2-VL 凭借其动态分辨率支持、跨模态位置对齐能力与多阶段优化机制，在大模型时代的多模态

融合路径上提供了范式性的技术指引与可扩展性示范。

2.2 大模型参数微调相关技术

2.2.1 全量微调

全量微调（Full Fine-tuning，FFT）是对预训练模型的所有参数进行微调，即预训练模型的所有层和参数均被会更新和优化，从而适应目标任务的需求。需要注意，与预训练一样，全量微调需要足够的内存和计算来存储和处理训练过程中的所有梯度、优化器和其它需要更新的部分。计算公式如式 (2.1) 所示。

$$W = W_0 + \Delta W \tag{2.1}$$

其中 W_0 是初始参数，ΔW 是更新参数。在这种方法中，模型的所有组成部分，包括语言模型、视觉编码器以及多模态融合模块在内的全部权重都被解冻，参与梯度计算与反向传播。全量微调一般具有较强的表达能力调整能力，能够最大程度地发挥模型在特定任务上的性能，特别适用于目标任务与预训练任务差异较大或数据量充足的情况。但是全量微调对计算资源的消耗非常高，训练成本大，并在小样本场景中存在过拟合风险，此外还可能破坏原始预训练模型所具备的泛化能力。

2.2.2 参数高效微调

参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）是一类旨在降低大模型微调过程中资源开销，同时保持较强下游任务性能的技术方法。与全参数微调不同，PEFT 方法并不更新模型的全部参数，而是通过引入结构化的参数优化策略，仅对模型的一小部分参数进行调整，从而显著减少显存消耗与训练时间。常见的 PEFT 方法可大致分为以下四类。

（1）冻结方法微调

即在微调过程中对模型的大部分参数进行冻结，仅允许部分特定层或特定类型的参数参与训练。这类方法操作简单、适应范围广，常用于 Transformer 结构中冻结所有 Transformer 层，仅微调顶部的分类器或少量的 Transformer 模块层，甚至只更新层归一化和偏置项等少数参数。冻结方法通过限制训练参数的范围，有效降低了训练成本，但在任务分布差异较大时可能存在性能瓶颈。近年来也出现了更细粒度

的冻结策略，如只微调注意力权重或特定模块，从而在控制开销的同时获得更好的效果。

（2）加性方法微调

即在原始模型架构中插入额外的轻量模块，仅对新增的可独立微调参数的模块进行训练，预训练参数保持不变。这类方法的代表包括基于适配器的方法、基于提示学习的方法等。基于适配器的方法即在每个 Transformer 层中插入小型的瓶颈网络，训练时仅更新适配器的参数。该设计允许模型在不同任务间灵活切换适配器模块，实现任务间的高效迁移。在提示学习方面,Li 等人[40]推出的 Prefix-Tuning 方法进一步扩展了提示向量的应用范围，其在每一层 Transformer 的注意力模块中引入了一组可微调的前缀向量（prefix tokens），这些向量作为额外的上下文信息参与注意力机制的计算，但不干扰原始输入和主干模型的参数结构。通过仅优化前缀参数，Prefix-Tuning 实现了更高效的任务适应能力，特别适用于大规模语言模型在多任务环境下的低开销微调。这些方法的共同特点是与原始模型参数解耦，具备模块化、轻量化、迁移性强等优势，特别适用于任务众多、资源有限的多任务场景。

（3）重参数化方法微调

作为参数高效微调的一种重要策略，重参数化方法近年来在大规模预训练模型的下游任务适配中展现出显著优势。早期研究中 Aghajanyan 等人[41] 和 Qin 等人[42]指出，在微调阶段，大模型参数空间中存在大量冗余，这使得仅需更新极少数参数即可获得与全参数微调相当的性能表现。基于这一观察，重参数化方法应运而生，其核心思想是通过对预训练模型中权重参数的结构性重构，将原始高维参数空间投影到一个维度更低的子空间中进行优化，从而大幅减少训练所需的参数数量与计算成本，同时保留模型的表示能力与泛化能力。与直接在原始参数矩阵上执行梯度更新的方法不同，重参数化技术通常借助低秩分解、张量分解或稀疏结构约束等形式，构建一个新的参数化路径，以间接方式对模型参数进行调整。

其中，最经典的代表方法就是 LoRA（Low-Rank Adaptation）[43]。LoRA 方法的核心思想是在微调过程中，冻结原始模型的权重，仅在模型中的部分线性层（如全连接层、注意力投影层等）引入一个低秩可训练分支，以替代直接更新高维参数矩阵的方式。具体而言，LoRA 将一个参数矩阵分解为两个参数矩阵的向量乘积。已知

向量乘积的计算方式表述为式（2.2）所示：

$$C = A \times B \tag{2.2}$$

其中，A 矩阵的尺寸大小为 m × q，B 矩阵的尺寸大小为 q × n，C 矩阵的尺寸大小为 m × n。例如，A 矩阵的尺寸大小为 3 × 1，B 矩阵的尺寸大小为 1 × 3, 那么可以将 3 × 3 的矩阵转换成两个 1 × 3 和 3 × 1 的矩阵，参数量减少 $\frac{1}{3}$。同理，假设原始权重更新矩阵为 $\Delta W \in \mathbb{R}^{d\times k}$ ，当 $r \ll min(d, k)$ 时，ΔW 可以被两个低秩矩阵 $B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$ 逼近，即 $\Delta W = B \times A$，从而显著压缩了训练参数的维度。在实际训练中，模型的前向传播形式被重构为式（2.3）所示：

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{2.3}$$

其中 W_0 为冻结的预训练权重矩阵，x 是输入，h 是输出。训练时仅更新 A 和 B，而保持 W_0 不变。也就是先将输入向量投影到一个较小维度（r）的向量当中，再恢复到原始的维度（d），从而即可将计算复杂度从 $O(d^2)$ 降低至 $O(rd)$。图2.5展示了 LoRA 方法示意图。

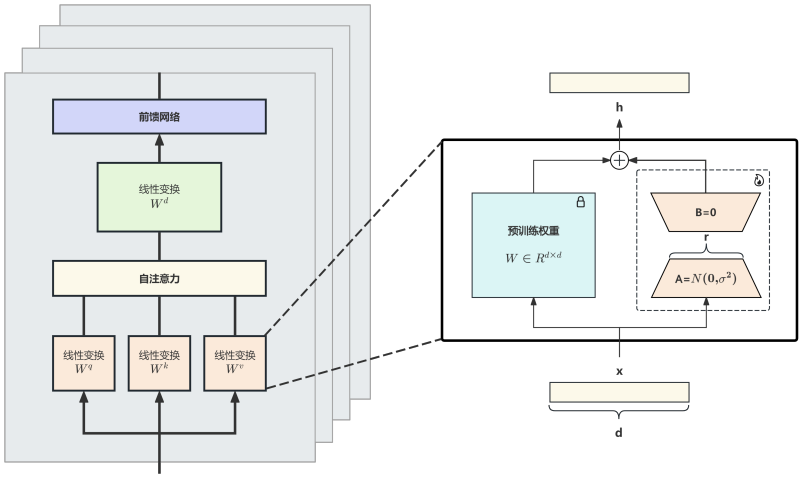


图 2.5 LoRA 方法示意图

我们对矩阵 A 使用高斯随机初始化，而将矩阵 B 初始化为全零，因此初始时 $\Delta W = BA$ 为零。在训练过程中，我们将 $\Delta W x$ 乘以一个缩放因子 $\frac{\alpha}{r}$，其中缩放因子 α 是一个在秩 r 维度上的常数。在使用优化器时，调整 α 的效果相当于调整学习率。实验中通常将 α 设为首次尝试的秩 r 的两倍，并不对其进行额外调优。这个缩放操

作有助于在调整 r 值时，减少重新调节超参数的需求。

在 LoRA 微调过程中，若设输入为 x，标签为 y，则微调时的优化目标为最小化下游任务上的监督损失，如公式 (2.4) 所示：

$$\mathcal{L} = \mathcal{L}_{\text{task}}(f(x; W + BA), y) \tag{2.4}$$

其中 $f(x;\cdot)$ 表示模型的前向传播函数。在训练过程中，原始参数 W 被冻结，反向传播仅作用于新增的可训练参数 A 和 B，从而在显著降低训练成本的同时保持模型表示能力。根据链式法则，损失函数关于 A、B 的梯度如下（2.5）所示：

$$\frac{\partial \mathcal{L}}{\partial A} = \frac{\partial \mathcal{L}}{\partial (W + BA)} \cdot B^{\top}, \quad \frac{\partial \mathcal{L}}{\partial B} = A^{\top} \cdot \frac{\partial \mathcal{L}}{\partial (W + BA)} \tag{2.5}$$

对应地，参数更新采用常规的梯度下降策略，更新公式如下（2.6）所示：

$$A_{new} = A - \eta \frac{\partial \mathcal{L}}{\partial A}, \quad B_{new} = B - \eta \frac{\partial \mathcal{L}}{\partial B} \tag{2.6}$$

其中，η 是学习率。由此可以看出，由于矩阵 A 和 B 的秩 r 远小于原始参数矩阵的维度，LoRA 显著减少了反向传播中参与计算和存储的参数数量不仅如此，LoRA 以附加模块的形式引入到线性变换层中（如注意力机制中的 QKV 权重），不改变原始模型结构，具有极强的模块化与兼容性，能够无缝集成到主流大规模预训练模型架构中。其参数更新可在不干扰主干网络的前提下独立训练，为大模型部署中的灵活性与可控性提供了有力支持。

此外，在实际应用中，秩的选择对微调效果与资源利用率具有重要影响。经验上，任务难度、基座模型能力和数据规模共同决定了最优的秩取值范围。对于语义简单、训练数据充足或使用性能强大的基础模型时，较小的秩即可满足迁移性能；而对于任务复杂、领域特化或多任务联合训练等情形，则需要适当增大秩值以增强表达能力。在实际使用中，常通过验证集性能评估或网格搜索的方式选择合适的秩，并结合缩放因子进行更新幅度调控，从而实现性能与效率的最优折中。

（4）其他方法微调，Galore

Galore[44]（Gradient Low-Rank Projection for Memory-Efficient Adaptation）是一种新颖的参数高效微调方法，旨在减少大规模模型训练过程中的显存和内存开销，同时保持或接近全参数微调的性能。与以往方法主要压缩模型参数不同，GaLore 的核

心思想是直接对梯度进行低秩投影，在保持优化路径关键方向的同时，有效降低训练中梯度存储与更新的维度。

在标准微调中，模型梯度的维度与模型参数本身相同，若模型参数规模为 m×n，则其梯度也是同样的高维矩阵。而 GaLore 方法认为，训练过程中的权重梯度矩阵 $G \in \mathbb{R}^{m \times n}$ 通常具有低秩结构。基于此，GaLore 利用这一特性，通过奇异值分解（SVD）将梯度矩阵 G 投影到一个低秩形式, 如式（4.1）所示:

$$\tilde{G} \approx P^T G Q \tag{2.7}$$

其中，$P \in \mathbb{R}^{m \times r}$,$Q \in \mathbb{R}^{n \times r}$，且秩 $r \ll min(d, k)$ 。GaLore 方法理论上证明，在一定条件下，若采用固定的投影矩阵 P 与 Q，优化过程依然能够收敛，为该策略提供了坚实的理论支撑。

在训练过程中，GaLore 在每次反向传播后，将梯度 G 投影至当前所处的低秩子空间中，仅需维护并更新投影矩阵 P 与 Q，而无需显式存储完整的高维梯度。通过周期性地更新投影矩阵，GaLore 实现了多个低秩子空间之间的转换，使模型得以在多个子空间中进行学习，而不局限于单一的投影方向。这样一来，在不更改前向结构的前提下，就可以实现对显存需求的显著压缩。

如图2.6为 GaLore 方法示意图。在训练前阶段 $t_1 \in [0, T_1 - 1]$ 时间内，梯度更新在低秩子空间 $\tilde{G_{t_1}}$ 进行。在 T_1 之后，投影空间切换至 $\tilde{G_{t_2}}$，继续后续的训练过程。通过这种分阶段的动态切换策略，GaLore 兼顾了内存效率与训练性能，为大规模模型在资源受限场景下的训练提供了一种有效解决方案。

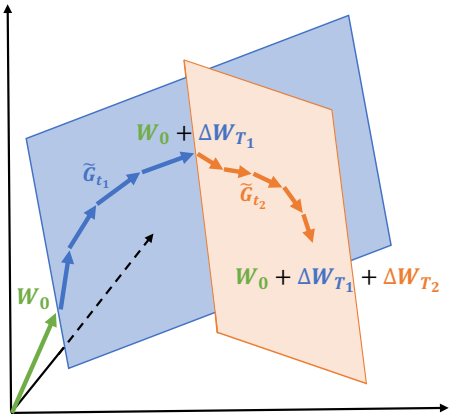


图 2.6 Galore **方法示意图**[44]

虽然与 LoRA 在名字上都含有“低秩”，但 GaLore 并不显式引入新的训练模块，

而是通过对反向传播中的梯度矩阵进行投影近似来压缩训练开销。当子空间秩等于矩阵最小维数时，GaLore 与原始模型完全一致。通过梯度低秩投影，GaLore 可以在不牺牲模型性能的前提下显著降低内存消耗，使得在消费级 GPU 上训练大语言模型成为可能。

2.3 文本生成评价指标

本文提出的任务是从视频内容中理解其潜在隐喻意义，并生成具备隐喻性、语义契合的自然语言描述。该任务既涉及视频中深层语义和象征信息的建模与理解，又要求输出具备语言创造性和认知隐喻迁移能力的描述文本，因此在隐喻文本生成评估上需同时考虑理解能力与生成质量两个维度。接下来介绍三种本文使用的评价指标。

2.3.1 BLEU 分数

BLEU（Bilingual Evaluation Understudy）[45] 是一种用于评估机器生成文本与参考文本之间相似度的自动化指标，其核心思想是将文本视作由若干“词语片段”（即 n-gram）组成，并通过这些片段的匹配情况来衡量两个句子之间的相似性。具体而言就是将生成的句子拆解成连续的词组单元（如单词、词对或词组），再将这些词组与参考答案中出现的词组进行比对。同时为了防止模型只生成非常短、但完全匹配参考短语的“投机性”句子，BLEU 在计算中引入了简短惩罚因子（Brevity Penalty，BP），以确保生成句子在长度上也具有一定合理性。BLEU 分数的计算公式如下（2.8）所示：

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.8}$$

其中，BP 的的大小由模型输出的文本长度 c 和参考文本的长度 r 共同决定。BP 值的计算公式如下（2.9）所示：

$$\text{BP} = \begin{cases} 1 & c > r \\ e^{(1-r/c)} & c \leq r \end{cases} \tag{2.9}$$

w_n 代表 n-gram 精度值在计算过程中的权重，一般设置为 1/n 表示各个 n-gram 精度值在计算过程中拥有同等的权重。在实际过程中，为了在捕捉语言表达的准确性与流畅性之间取得的平衡，BLEU 评价指标通常采用 n=4 的 4-gram。基于此，本文同样选择 BLEU-4 作为评价指标。

2.3.2 ROUGE-L 分数

ROUGE-L（Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence）[46]是一种基于最长公共子序列（Longest Common Subsequence, LCS）的方法，用于评估自动生成文本与参考文本之间的结构和内容相似性，尤其适用于语言生成任务。

ROUGE-L 的核心在于它能够同时捕捉到词汇内容匹配和语序信息保持，这是基于 LCS 的方法相较于传统 n-gram 匹配的优势所在。LCS 是指两个序列中按照顺序一致但不要求连续的最长子序列，其长度越大，说明两个序列越相似。ROUGE-L 利用 LCS 来衡量生成文本和参考文本之间共享结构的程度，通过同时计算召回率（Recall）和精确率（Precision）并基于它们的调和平均得到最终得分。

$$\begin{cases} \text{ROUGE-L} = \dfrac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}} \\ R_{LCS} = \dfrac{LCS(X,Y)}{len(Y)} \\ P_{LCS} = \dfrac{LCS(X,Y)}{len(X)} \end{cases} \tag{2.10}$$

其中，LCS（·）为求最大公共子序列的函数，len（·）为求序列长度的函数，β 为可调整的超参数。ROUGE-L 的优势在于能够考虑句子中单词的顺序关系，适合结构化语言输出任务。同时，相比于严格的 n-gram 匹配，LCS 方法更能容忍轻微的措辞变动。因此，在视频隐喻生成任务中，ROUGE-L 能够有效评估模型生成描述在语义内容与结构表达上的保真度，尤其适合那些在表达方式上可能与参考答案不完全一致但传达相似意图的生成结果。

2.3.3 BERT 分数

在自然语言生成任务中，大多数传统的自动评价指标（如 BLEU 等）主要基于 n-gram 匹配机制，仅衡量候选文本与参考文本在词级别的表层相似性，忽视了语义

层面的深度对齐。为克服这一限制，BERT 分数（BERT-Score）被提出[47]，作为一种能够捕捉文本间语义相似性的评价指标，因其更契合人类语言理解方式而在近年广泛应用于生成任务的评估中。

BERT-Score 利用预训练语言模型（如 BERT、RoBERTa）对文本进行上下文感知的表示编码，然后通过计算候选文本与参考文本的每个词向量之间的余弦相似度，获取整体的匹配得分。与传统指标不同，BERT-Score 不要求候选词与参考词完全重合，而是基于语义上最相似的词进行对齐，因此即使词汇使用不同，只要语义一致，也可获得较高的得分。此外，BERT-Score 通常会引入逆文档频率（Inverse Document Frequency, IDF）作为加权系数，对稀有词赋予更高权重，从而更好地反映信息密度与重要性。计算公式如下（2.11）所示：

$$\begin{cases} R_{BERT} = \dfrac{1}{\mid x \mid} \displaystyle\sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} x_i^T \hat{x_j} \\ P_{BERT} = \dfrac{1}{\mid \hat{x} \mid} \displaystyle\sum_{\hat{x_j} \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x_j} \\ F_{BERT} = 2 \cdot \dfrac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \end{cases} \tag{2.11}$$

由于不同的表达方式可以传达相同的信息，所以基于 N-Gram 重叠的指标常常无法捕获生成的隐喻的质量。因此，本文还使用了 Bertscore，以比较生成隐喻和参考标签的语义相似性。

2.4 本章小结

本章全面阐述支撑本文实验工作的关键技术框架，包括所采用的大语言模型、主流的大模型参数微调方法，以及用于性能评估的各项指标。第一节重点介绍了本文选用的大语言模型，涵盖当前表现突出的 InternVL2.5、VideoLLaMA3 以及 Qwen2-VL 模型，为后续实验模型的选型和应用提供理论支撑。第二节则系统梳理现有的大模型参数微调技术，为本文后续的大模型微调实验打下坚实基础。第三节详细说明了实验中采用的评价指标体系，包括 BLEU、ROUGE-L 与 BERTScore 三种常用自然语言生成评价指标，介绍各自的计算原理以及在本研究中的选用依据，以确保实验结果的客观性与可比性。

第三章　短视频隐喻数据集的构建与评估

3.1　数据集制作流程设计

为弥补当前在中文短视频隐喻研究领域中尚无公开数据集的现状，本文自主构建了一个面向短视频隐喻的中文数据集，旨在为相关研究提供基础数据支持。该数据集的构建不仅填补了该领域数据资源的空白，也为多模态隐喻识别与生成等任务提供了有力支撑。数据集制作流程如图3.1所示。

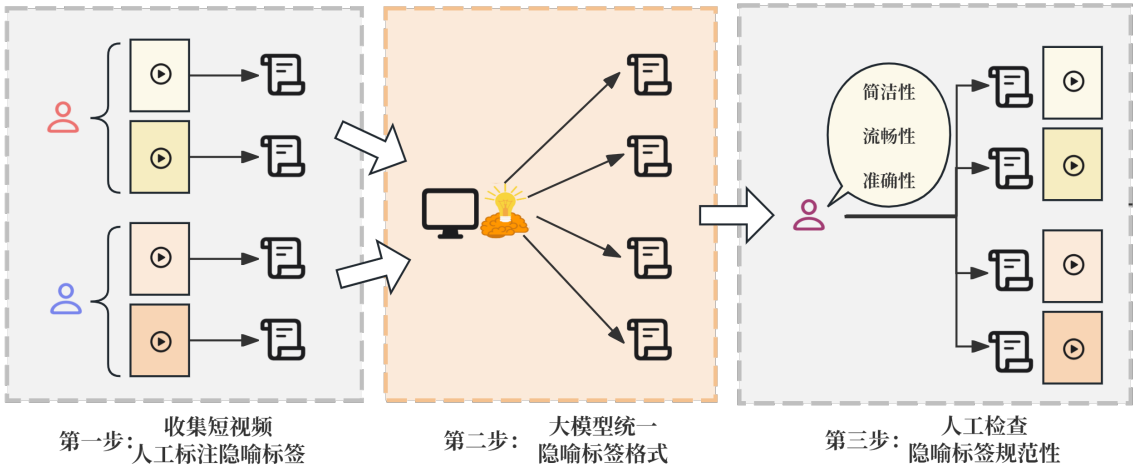


图 3.1　数据集流程设计

具体而言，该数据集制作包含人工收集数据、大模型统一格式处理和人工校审三阶段流程，以确保短视频中文隐喻数据集在准确性、一致性与语言质量等方面达到较高水准。

第一阶段为收集短视频，人工标注隐喻标签。本研究从抖音等主流中文短视频平台中人工筛选出具有代表性的 898 个视频样本，重点关注在内容表达上富有创意、具有一定语言复杂度和文化内涵的片段。随后，由 13 位具备良好语言表达能力及隐喻理解能力的标注者对筛选视频中的隐喻进行细致标注与分类。标注内容包括该视频的隐喻类别以及初步的隐喻解释。

第二阶段为大模型统一隐喻标签格式。将第一阶段中人工完成的标注信息输入经过现有的商用大语言模型，对标注文本进行统一格式化处理、语言风格优化及细

节润色，以提升标注内容的语言流畅性和表达规范性。大模型在此过程中不仅起到统一风格的作用，也能发现部分表达不清或存在歧义的标注信息，进行初步修正。

第三阶段为人工复审隐喻标签规范性。对大模型输出的结果进行逐条审阅，重点检查其在简洁性（是否冗余）、流畅性（语言是否自然通顺）、准确性（隐喻判断是否合理）等方面的表现，确保最终形成的标注结果既符合语言表达规范，又真实反映视频内容中的隐喻现象。

通过以上多阶段、系统化的数据构建流程，本文所构建的中文短视频隐喻数据集在确保数据质量的同时，强调其标注规范性，力求在表达的隐喻性、语境的多样性与标注的一致性之间取得良好平衡，为后续在隐喻识别、生成与理解等关键任务中提供坚实有力的数据支持。

3.2 数据收集

在数据收集阶段，本文采用人工筛选的方式构建高质量的中文短视频隐喻语料库。具体而言，由 13 位具备扎实语言表达能力与较强隐喻理解能力的标注人员组成的标注团队，面向主流中文短视频平台开展大规模视频筛选工作。筛选过程中，标注者重点关注那些在表达内容上具备明显隐喻特征、语言风格富有创意的视频样本，以确保所构建的数据集具有语义深度。

此外，为提升数据集在语义风格上的多样性，本文根据短视频隐喻所处的语境特征，将全部样本划分为：讽刺类、诙谐类、生活态度类、硬核类与自嘲幽默类五类，有助于对不同类型隐喻进行系统性的归纳与分析。最终收集到的数据集中，数量最多的是诙谐类，共计 425 条，反映出用户刷短视频通常以娱乐为主的大趋势。讽刺类视频主要通过讽刺和批评的方式揭示社会现实问题，共计 140 条。生活态度类侧重于表达个体的处世观念与价值取向，共计 119 条。硬核类则以强烈突出的语言或视频风格为主要特征，共计 105 条。自嘲幽默类通过戏谑自身的方式构建更具反思性的深层隐喻，共计 109 条。最终形成的隐喻视频语料库共计 898 条样本，涵盖了多样的语言风格与语义类型，具备良好的覆盖广度和分析深度，为中文多模态隐喻研究提供了坚实的数据支撑。

所采集的原始隐喻标签首先以结构化的 Excel 表格格式进行存储，每条样本均包含视频文件名、隐喻类别标签及隐喻含义的详细解释等信息，便于查看和批量处理。

下表3.1列举了原始 Excel 表中的五大类别样本数据，可以直观看出不同类别的短视频传达的隐喻截然不同。

表 3.1 原始 Excel 样本数据

短视频文件名	类别	隐喻解释
自嘲幽默类 _1	自嘲幽默类	视频中的作者在进行立定跳远，在跳跃过程中手机从口袋中掉落，恰巧掉落在作者落脚的地方，导致作者没有站稳而摔倒。作者本人给视频的标题带有“神操作”、“猎奇”的 tag，充满了自嘲的意味。
诙谐类 _47	诙谐类	哈基米是日语“蜂蜜水”的发音，因被当做 BGM 用于各类猫猫萌宠视频中而走红，后成为一切“可爱事物”的代称。
硬核类 _1	硬核类	视频中的作者是一名在读中学生，画面记录了两个场景，一个是“住宿生回宿舍的路”，画面阴暗、灯光呈现冷色调；另一个是“住宿生回家的路”，画面温馨、灯光呈现暖色调。作者配文“阴阳两隔”，从侧面硬核地表达了中学生两种不同的居住生活所体现的精神状态。
生活态度类 _21	生活态度类	松弛感指面对压力时从容应对、善待自己、不慌张、不焦虑的心理状态，倡导一种轻松、自在的生活态度。
讽刺类 _27	讽刺类	视频主题为“太感动了，以后僵尸肉只卖给大学生”，讽刺了网络上挂羊头卖狗肉，以量大好吃为买点，骗取大家的信任，贩卖僵尸肉给孩子们。

为了满足大语言模型在训练与推理阶段对输入数据结构的规范要求，本文进一步将该 Excel 格式的数据统一转换为符合 ShareGPT 规范的 JSON 格式。该格式广泛应用于对话式大模型的数据输入中，能够有效支持隐喻识别与生成任务中对复杂语义关系的建模与理解。

ShareGTP 格式包含“videos”和“messages”两个关键字段。其中，“videos”字段记录了与样本对应的短视频文件路径，确保模型能够准确关联文本内容与视觉输入。“messages”字段则以对话轮次为单位组织语义信息，其内部为一个数组结构，每个元素表示一次完整的对话交互。每条“messages”由两个字段组成“role”与“content”。

”role” 代表对话参与者的角色，通常为”user” 或”assistant”。”content” 代表该轮对话中角色所说的具体内容，可以是问题、回答或指令等文本。下图3.2展示了 ShareGTP 格式数据。之后的数据增强阶段将对该格式的数据标签进行处理。

ShareGPT格式数据


```
{
  "messages": [
    {
      "content": "<video>理解视频的隐喻内容。",
      "role": "user"
    },
    {
      "content": "视频中的作者在进行立定跳远，在跳跃过程中手机从口袋中掉落，恰巧掉落在作者落脚的地方，导致作者没有站稳而摔倒。作者本人给视频的标题带有“神操作”、“猎奇”的tag，充满了自嘲的意味。",
      "role": "assistant"
    }
  ],
  "videos": [
    "meme/meme/梗/自嘲幽默类/自嘲幽默类_1.mp4"
  ]
}
```


图 3.2 ShareGPT **格式数据**

3.3 数据增强处理

在构建短视频隐喻数据集的过程中，除初始人工筛选与标注阶段外，第二阶段和第三阶段统称为数据增强处理部分，旨在提升数据质量和增强语义一致性，统一在本节做详细介绍。

首先，针对第一阶段收集得到的原始隐喻标签数据，尽管标注者在理解与表达层面具备较高能力，但由于语言表达风格存在差异，部分描述内容在结构与语义上呈现出较强的非一致性。这种非一致性不利于模型在训练阶段对隐喻语义结构的统一建模。因此，本文在第二阶段调用商用大语言模型 DeepSeek-V3，利用其在文本重写与语言组织方面的能力，通过 API 接口对所有隐喻标签数据进行格式化重写，统一输出为如下结构模板：

本视频讲述了 [视频内容]。[核心概念] 出自 [出处]，其因为 [特征/新含义/事件]，表达出了 [情绪/态度]。

该统一格式模板具备清晰的结构，有助于提升训练数据的规范性与模型对隐喻三元关系的学习能力。其中，“视频内容”简明概述视频的基本情境。“核心概念”对应隐喻中的源域概念，是隐喻意义构建的基础。“特征/新含义/事件”反映隐喻的映射机制，揭示源域向目标域的转化路径。“情绪/态度”则体现出目标域的情感或立场，是隐喻所要传达的深层意涵。而“出处”作为背景补充信息，有助于模型建构更具上下文意识的语义理解框架。

通过这一格式化过程，原始标注数据被转化为语义结构统一、内容要素清晰的高质量输入语料，为后续模型微调与隐喻推理任务奠定了坚实的语料基础。同时，这种隐喻解析格式在语言学与认知语义学层面也具备可解释性，体现了“源域—映射机制—目标域”这一经典隐喻结构，有助于模型在隐喻识别与生成任务中实现更好的结构对齐与语义抽象能力。转换后的数据对比如下图3.3所示。

原始隐喻标签

视频中作者作为伴娘的身份与小孩子们等待新人扔红包，但是新娘作为伴娘的好闺蜜直接将一堆红包直接塞给作者，一旁的新郎仍然喜笑颜开，视频幽默风趣，激起了大家与朋友的情谊，结婚意味着新生活的开始，整个视频带给人们积极乐观的观感。

(a) **原始隐喻标签**

大模型处理后隐喻标签

本视频讲述了作者作为伴娘在婚礼上与孩子们一起等待抢红包，但新娘作为闺蜜直接将红包塞给她，而新郎依然笑容满面的场景。[核心概念]“婚礼红包”出自中国传统婚俗文化，其因为[新娘打破常规的亲密举动]与[新郎宽容幽默的态度]，表达出了[对友情的珍视]和[对新生活的乐观期待]。视频通过反差式幽默重构了“抢红包”的仪式感，以闺蜜特权隐喻深厚友谊，用新郎的豁达象征婚姻生活的包容性，最终传递出温暖积极的情绪。

(b) **大模型处理后隐喻标签**

图 3.3 第二阶段前后隐喻标签对比

经过统一格式化处理后的隐喻标注数据不仅严格遵循“本视频讲述了 [视频内容]。[核心概念] 出自 [出处]，其因为 [特征/新含义/事件]，表达出了 [情绪/态度]。”这一结构规范，还在内容上补充了隐喻分析与理解。这一处理过程显著增强了数据的结构化程度和语义透明度，为模型的语义建模和认知推理提供了有力支持。

然而，尽管第二阶段通过调用 DeepSeek-V3 大模型的 API 实现了初步的格式统一与语义提炼，显著提高了隐喻标签文本的结构化程度与语义清晰度，但由于当前

大模型在中文隐喻类内容生成方面仍可能存在一定程度的偏差与不稳定性，生成文本在内容表达、语言风格以及逻辑结构上仍可能出现冗余、压缩或轻微失真等问题。因此，本文进一步引入了人工校准阶段作为第三步处理流程，以提升文本质量的整体可读性与语义准确度。

在该阶段，本人主要围绕简洁性、流畅性与准确性三个核心维度对大模型生成结果进行细致校对。首先通过删减冗余信息、合并重复表达和重构语序，使文本更加紧凑高效。其次，重点对不自然的句式、机械拼接结构及翻译腔语言进行润色，使文本符合中文母语者的表达习惯与语用逻辑。最后，针对大模型在隐喻映射关系理解上的潜在误判，重新核对视频语境与隐喻要素之间的关联，确保核心概念、出处、特征以及情绪态度的表达准确无误。第三阶段数据前后对比如下图3.4所示。

大模型处理后隐喻标签

本视频讲述了一对情侣通过即兴舞蹈化解负面情绪的故事。\"沙发舞\"出自日常生活物品的非常规使用，其因为被赋予\"手风琴\"的象征意义（通过动作与音乐的配合实现情感宣泄），表达出了从压抑到欢快的情绪转变。 \n\n（说明：重构后严格遵循给定格式，将\"搬沙发跳舞\"提炼为\"沙发舞\"这一核心概念，并点明其通过\"物品功能转化\"和\"音乐协同\"实现隐喻升华——用荒诞的肢体语言解构了\"别不开心了\"这句日常安慰，使抽象情感抚慰获得可视化表达。）

(a) **大模型处理后隐喻标签**

人工复审后隐喻标签

本视频讲述了一对情侣男生情绪低落，女生通过即兴搬沙发跳舞化解他的负面情绪，随后两人一起跳舞心情逐渐明媚的故事。\"搬沙发跳舞\"出自日常生活物品的非常规使用，其因为被赋予\"手风琴\"的象征意义，通过动作与音乐的配合实现情感宣泄，表达出了从压抑到欢快的情绪转变。视频通过\"物品功能转化\"和\"音乐协同\"实现隐喻升华——用荒诞的肢体语言解构了\"别不开心了，让我为你唱首歌\"这句日常安慰，使抽象情感抚慰获得可视化表达。

(b) **人工复审后隐喻标签**

图 3.4 第三阶段前后隐喻标签对比

从图中对比可以观察到，在对隐喻内容进行解析与生成的过程中，大模型在寻找关键概念时往往会表现出一定的自主压缩与抽象能力，这种能力虽在一定程度上体现了语言模型对语义提炼的潜力，但也可能带来概念上的偏离。例如大模型将“搬沙发跳舞”这一具体动作概念压缩为“沙发舞”这一新造词汇，可能会导致目标隐喻

意图的模糊化，进而影响模型对语义映射的准确把握。同时，大模型将视频内容压缩为一句话“一对情侣通过即兴舞蹈化解负面情绪”，丢失了原有视频的细节，准确性降低，需要人工进行补充。

实际上，原始视频内容是一对情侣男人坐在桌子前操作手机情绪低落，过了一会儿，女人搬起家里的沙发跳起舞来，像是在拉手风琴一般。男人看到后也搬起了另一个沙发，与女人一起跳起舞来，情绪变得不再低落。视频表达的核心隐喻构建在“搬沙发跳舞”这一具体行为之上。通过借助“沙发”这一家庭生活中的具象物体，比作“手风琴”这样的艺术象征，从而传递出情侣之间的积极向上、不惧辛苦的生活态度，还在视觉层面传达了一种温柔的安抚与心理慰藉效果。

此外大模型在生成过程中常常伴随大量格式性文字、注释信息或“思考轨迹”式的推理文本，这些内容虽在一定程度上体现了模型的推理路径与生成机制，但也容易造成语义表达冗余，尤其是在隐喻类任务中可能导致核心概念被淹没在冗长的文本中，失去标签应有的凝练特性。因此，在隐喻处理任务中，如何在语言简化与语义保真之间取得平衡，避免模型在压缩语义时牺牲核心概念的完整性，是一个值得进一步探讨的问题。也从侧面说明了在大语言模型生成过程中，人类语义判断仍具有不可替代的重要性，尤其是在涉及复杂认知与社会语境解读的多模态任务中，人工干预对于提升生成质量具有关键意义。通过人工校审，可以在确保语义完整性与逻辑一致性的前提下，有效提升标签生成质量与适应性，从而增强模型在面对真实社会语境时的泛化与稳健能力。

经过第三阶段的人工校准与优化处理，最终形成的隐喻标签在语言表达上更加自然流畅，隐喻核心概念表达简洁、逻辑清晰，同时保留了关键细节与语境线索，使得每条标签不仅具备高度的信息浓缩能力，还为后续模型在情感生成、意图识别与社会语境建模等下游任务中提供了更加坚实的语义基础。

3.4 数据集性能分析

在完成短视频隐喻数据集构建后，本节将对其整体性能进行系统化分析，重点考察数据分布、语义表现、标注一致性及格式规范性等多个关键维度，旨在评估该数据集在服务下游任务中的适用性、扩展性与研究价值。

（1）数据分布均衡性

本文所构建的数据集共包含 898 条高质量样本，涵盖讽刺类、诙谐类、生活态

度类、硬核类与自嘲幽默类五种主要的短视频类型，虽在数量上存在一定差异，但整体分布与真实短视频平台上用户刷到的类型数量趋势基本保持一致。例如，诙谐类占比显著，反映出当前短视频语境下以娱乐、趣味为主的隐喻表达具有较高的流通度与用户接受度，而讽刺类、自嘲类等则在数量上相对较少但语义深度较强，充分保障了数据集在内容表达上的覆盖广度与层次多样性。这种结构不仅为后续模型训练提供了足够多样的隐喻实例，也提升了模型对不同语义风格的区分能力和泛化能力。

（2）语义多样性

数据集中隐喻语义表达的多样性也体现出其较高的语言学价值与实用性。在初始数据收集阶段，标注团队特别关注那些语言风格独特、情感倾向复杂、需要背景知识参与推理的视频隐喻内容，确保样本在隐喻机制、文化负载、修辞手段等方面具有足够的变异性与代表性。为进一步提升数据质量，第二阶段使用 DeepSeek-V3 大模型对初始标签进行了格式化增强，统一转换为“本视频讲述了 [视频内容]。[核心概念] 出自 [出处]，其因为 [特征/新含义/事件]，表达出了 [情绪/态度]”这一结构化模板，该模板不仅覆盖了源概念到目标概念的映射路径，还引入了文化语境与情绪表达两个维度，强化了隐喻意义的可解释性与可迁移性。在此基础上，第三阶段通过人工复审手段对语言流畅度、概念准确性以及表达的精炼性进行了再度优化，使得最终标签文本既简洁凝练，又不失细节丰富，适合用于模型训练与生成任务。

（3）标注准确性与一致性

本文所有数据样本均由 13 位具备隐喻理解能力与语言表达素养的标注者手工筛选完成，初步形成语义标签后，再经由大模型生成规范结构，最后通过人工校准环节，确保每条标签在隐喻映射路径、语境信息、情感表达等方面具有明确边界与清晰指向，最大限度降低主观歧义与语义偏差。此三阶段式的标签生产机制，使数据集在内容层面具备高度的一致性与可靠性，同时避免了因大模型过度生成或语义压缩导致的误判问题。例如，部分初始生成结果中存在对源概念的创新性压缩，在人工审校中均被修正为更贴近视频实际内容的表达形式，确保隐喻内涵不会因语言异化而丢失。

（4）格式规范性

本文数据集在结构设计上充分对接大语言模型的输入需求，将原始 Excel 表格转换为符合 ShareGPT 标准的 JSON 结构，其中包含视频路径与多轮对话式消息数组，

显著提升了模型对复杂语境的建模能力与上下文理解能力。在 ShareGPT 格式中，每条样本通过“user 提问，assistant 解释”的方式呈现隐喻识别过程，模拟真实交互场景，便于模型学习隐喻理解的推理逻辑与语言生成方式。同时，格式结构清晰，支持多任务扩展（如情感分类、风格识别、文化来源溯源等），具备良好的工程部署性与研究灵活性。

综上所述，本文构建的中文短视频隐喻数据集在数据分布均衡性、语义多样性、标注一致性与结构规范性方面均表现出色，既具备较高的语言学研究价值，也为大语言模型在中文语境下的隐喻识别与理解提供了坚实的数据支撑。该数据集涵盖多种语用类型，语言风格多样，表达层次丰富，能够较好反映真实语境中的隐喻使用特征。通过引入大模型辅助处理与人工校审机制，确保了标签在语义准确性与表达规范性上的双重质量，为多模态语言理解研究、模型推理能力提升以及中文语义分析等任务提供了重要参考与保障。

3.5 本章小结

本章介绍了中文短视频隐喻数据集的构建及其系统性评估 Z。第一节介绍了数据集制作流程，分为数据收集、大模型处理与人工校审三个阶段。第二节详细说明了数据收集过程。13 位具备语言理解与隐喻识别能力的标注者从主流短视频平台筛选隐喻性视频，并依据语用功能与社会语境将样本划分为讽刺、诙谐、生活态度、硬核及自嘲幽默五类。第三节介绍数据增强与处理过程，包含第二三阶段。在第二阶段本文使用商用大语言模型统一隐喻标签表达。第三阶段引入人工校审，对大模型生成内容中存在的语言冗余、概念偏差等问题进行修正与优化，确保标签简洁、准确、保留关键信息。第四节从数据分布均衡性、语义多样性、标注准确性与一致性以及格式规范性四方面对数据集性能进行评估。分析显示，该数据集在类型覆盖与语义深度方面表现良好，标签质量高，适配大模型输入需求。

第四章　大模型微调实验与结果分析

4.1　实验流程设计

本章旨在探索大模型在描述短视频中存在的隐喻任务上的应用潜力，深挖大模型对短视频语境和潜在情绪表达的理解能力。通过使用上一章构建的中文短视频微调数据集，对比不同微调方法在基座模型上的微调效果，期望得到一个性能最佳且可交互的隐喻视频助手。实验流程设计如图4.1所示。

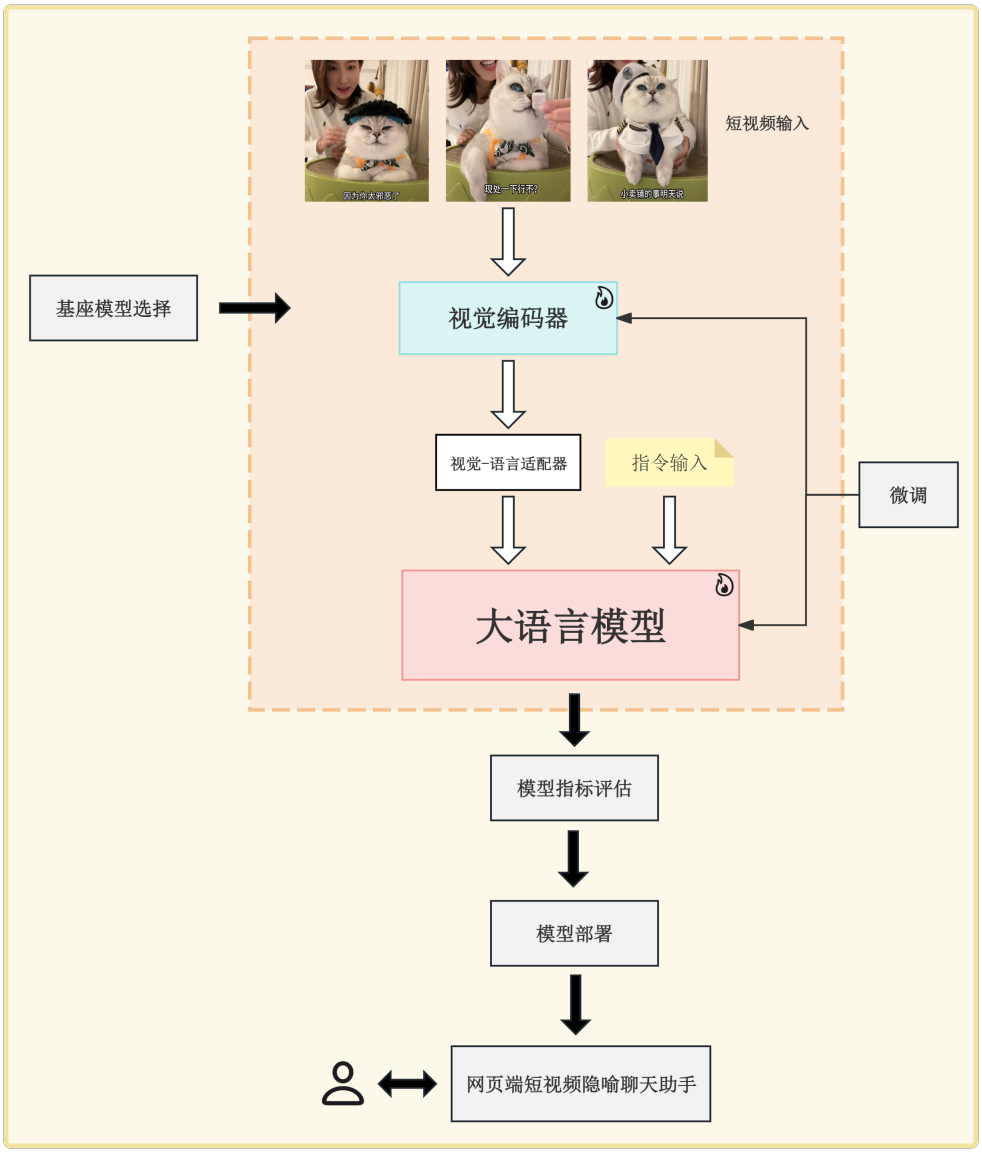


图 4.1 流程设计框架

首先，本文对当前主流的开源多模态大语言模型进行系统性评估，重点考察其在零样本条件下对中文短视频隐喻理解与生成任务的适应能力。通过对比不同模型在无监督设置下的表现以及对具有代表性的测试样本开展定性分析，评估大模型对短视频隐喻内容的理解能力与语言生成的连贯性。最后综合各项结果，选取在语义理解、生成连贯性及对比任务表现中综合能力最优的模型作为基座模型，为后续微调实验提供稳定可靠的起点。

在基座模型确定之后，本文进一步开展了一系列主流微调方法的实证对比实验，以探索不同训练策略在中文短视频隐喻生成任务中的表现差异和适配优势。所选 PEFT 方法包括最主流的 LoRA 微调方法、在 LoRA 基础上调整的改进方法、传统的 Freeze 策略以及近年来兴起的 Galore 方法。每种方法均在相同的训练集、验证集与测试集下进行实验，以探索不同方法在中文短视频隐喻任务中的表现差异。为确保评价结果的客观性与全面性，实验中采用 BLEU-4、ROUGE-L、BERT-F1 分数多项文本生成评价指标，全面验证大模型在短视频隐喻任务中的有效性、鲁棒性与泛化能力。

在实验验证基础上，本文选取性能最优的微调模型进行实际应用部署。微调后的模型被集成至云端服务框架，并通过前端交互式网页界面实现可视化访问。用户可通过浏览器本地访问该短视频隐喻聊天助手系统，进行多轮交互式问答。系统支持用户注册与登录、个性化头像更换、不同版本微调模型切换、历史对话记录查询等实用功能，极大增强了模型的可用性与可扩展性。

本文实验的软硬件开发环境如下表4.1所示。

表 4.1 实验开发环境

开发环境	描述
操作系统	Ubuntu22.04
CPU	16 vCPU Intel(R) Xeon(R) Gold 6430 120GB
GPU	NVIDIA RTX 4090 24GB
Python	3.12
Pytorch	2.6.0+cu124
transformers	4.49.0

4.2 基座模型选择

本文选取了2024年之后发布的主流多模态大语言模型，包括InternVL2.5、VideoLLaMA3与Qwen2-VL三种具备视频输入、图文理解与自然语言生成能力的模型进行深入对比与分析。这三个模型均支持端到端的视频感知与语义抽象处理，能够较好地完成视频到语言的跨模态映射，具备参与本任务评估的技术基础。重点考察三者在零样本场景下的综合性能，同时结合模型生成结果进行人工主观评估与实例效果分析，最终选出视频理解能力强、语言生成质量高、泛化能力出色的模型作为本次研究的基座模型。

为保证评估的公平性与广泛性，本文统一采用三种模型在2B参数级别（约20亿参数）下的预训练版本进行测试与对比。这样做一方面能够在保持相对较小算力需求的同时，支持较高质量的多模态语义建模。另一方面，对于时长通常不超过1分钟、语义浓缩程度较高的短视频而言，2B级别模型已能覆盖大部分隐喻识别所需的推理深度与信息整合能力。此外，从应用角度出发，中小参数量模型更具部署灵活性，能够在轻量级服务器、边缘设备甚至本地环境中实现快速响应，为隐喻识别系统的实用化提供良好基础。

实验随机选择100条短视频测试数据，对三种模型进行零样本能力测试。测试样本均匀涵盖五大短视频类别（讽刺类、诙谐类、生活态度类、硬核类与自嘲幽默类）各20条，以确保评估覆盖多样化语境和复杂情感表达。输入命令为:

理解视频的隐喻内容，由以下格式回答：本视频讲述了*[视频内容]*。*[核心概念]*出自*[出处]*，其因为*[特征/新含义/事件]*，表达出了*[情绪/态度]*。将*[]*内的内容替换成视频理解的内容。

参与实验的不同基座模型零样本泛化能力对比如表4.2所示。

表4.2 不同基座模型的零样本泛化能力

基座模型	BERT-F1	BLEU-4	ROUGE-L
InternVL2.5-2B	62.97	2.23	14.04
VideoLLaMA3-2B	58.52	**6.88**	22.98
Qwen2-VL-2B	**67.77**	6.46	**24.28**

从上表可以看出，Qwen2-VL-2B在多个文本生成评价指标上均表现出色，尤其

是在 BERT-F1 和 ROUGE-L 分数方面明显优于其他两个大模型，体现出其在语义一致性与信息覆盖能力上的显著优势。同时，在 BLEU 分数这一注重 n-gram 精准匹配的指标上，Qwen2-VL-2B 也展现出与 VideoLLaMA3-2B 相当的性能，进一步说明其在文本生成的准确性与保真度方面具有良好的表现。整体来看，Qwen2-VL-2B 在各项自动化指标中均处于领先水平，在多方面兼顾了隐喻生成任务所要求的语义深度、语言自然度与内容贴合度。

为了更全面评估三种大模型在短视频隐喻任务中的实际能力，本文进一步采用定性分析方式，选取具有代表性的测试样本并进行结果展示，以直观对比各模型在视频理解、隐喻建构、文本生成等方面的能力差异。不同基座模型的测试样本如表4.3所示。

表 4.3 不同基座模型测试样本

基座模型	测试结果
InternVL2.5-2B	这段视频由几个部分组成。首先，视频开头显示了一个计算机芯片的特写镜头，它的表面被蓝色的 LED 灯光勾勒出来，中间清晰地显示着“AI”这个词，周围是黑色的电路板。芯片和周围电路板上的每一个线路都清晰可见，整体色调偏冷，给人一种科技感。接下来是同样的场景，时间流逝通过微小的变化展示了芯片的亮度和蓝色灯光的强度，显示时间的推进。最后是黑屏上有一个搜索输入栏，提示“搜索号: yangyang_2person”，并带有搜索图标，鼓励观众搜索这个账号以获取更多内容。整个视频节奏较快，通过快速切换镜头增强了视觉效果。视频风格现代，采用了数字化和未来主义元素。
VideoLLaMA3-2B	本视频讲述了芯片中的 AI。AI 出自 [出处]，因为 [特征/新含义/事件]，表达出了 [情绪/态度]。
Qwen2-VL-2B	本视频讲述了 AI 芯片的未来科技。[核心概念] 出自 AI，其因为 AI 在计算机科学和工程学中的重要性，表达出了 AI 在未来科技中的重要性和潜力。将 AI 内的内容替换成视频理解的内容。
隐喻标签	本视频讲述了人工智能技术在多个领域的应用与发展。人工智能出自计算机科学领域，其因为能够模拟、延伸和扩展人的智能并在图像识别、语音识别、自然语言处理等方面取得重大突破，表达出了对科技进步的乐观态度与对未来发展的期待。

分析上表可以看出，InternVL2.5-2B 虽然能够较为详尽地识别视频中的多模态信

息，包括图像内容、字幕文本与背景音频等，但由于未对信息进行有效聚合与语义抽象，其生成结果往往是“逐帧复述”式的内容堆叠，对隐喻核心的理解能力有限。VideoLLaMA3-2B 的表现相较于其他两个大模型不太理想。在多个测试样本中，无法稳定输出连贯完整的长文本，生成的内容多为片段式语句或关键词堆积，难以构建出完整的语义链条或表达清晰的隐喻关系。该模型可能更擅长对静态图文类输入进行描述，对于动态、含义抽象度较高的短视频隐喻场景尚存在适应不足。而 Qwen2-VL-2B 的生成结果在内容完整性、逻辑连贯性与情感表达等方面均表现良好，并未局限于视频的表层描述，而是进一步挖掘其中的表达内涵“AI 未来发展潜力”。

整体来看，Qwen2-VL 模型在整体表现中最为突出，多模态融合能力更强、文本生成更具语义张力和隐喻适应性，因此被最终选定为本文后续微调实验与系统部署的基座模型。

4.3 微调实验与评估

由于实验硬件显存有限，且在处理视频输入时模型的显存占用会显著增加，传统的全量参数微调方法在实际训练过程中会带来极高的显存消耗与计算成本。因此，为了在有限的计算资源条件下有效提升大模型在短视频隐喻任务上的表现，本文选用了当前广泛应用且资源友好的参数高效微调作为核心微调策略。

在这里本实验采用了四类具有代表性的 PEFT 方法：LoRA 微调、改进版 LoRA 方法微调、冻结微调以及 Galore 微调，在基座模型 Qwen2-VL-2B 上对比它们在本文构建的中文短视频隐喻数据集上的表现，以找到最适合该任务的微调策略。本次实验选用上一章构建的中文隐喻短视频数据集，设置训练集 679 个样本，验证集 119 个样本，测试集 100 个样本。其中为了保证测试集的准确性，每个类别的样本分别 20 个，均匀分布。

4.3.1 LoRA 微调实验

LoRA 微调方法作为当前应用最广泛的参数高效微调技术之一，因其卓越的训练效率和良好的泛化能力，在多个大模型下游任务中展现出强劲性能。在本次实验中，LoRA 微调过程中总共引入的可训练参数仅为 9,232,384 个，相比于 Qwen2-VL-2B 模型原始参数 2,218,217,984 个，训练所需参数量仅占用 0.42% 的原始参数数量。这一结果充分验证了 LoRA 在大模型场景中的轻量级特性。训练过程显存占用控制在

19GB 左右。

本次实验对关键超参数进行了系统性的调优与对比分析，包括训练轮数、学习率以及秩和缩放系数。通过多组控制变量实验的结果对比，全面评估了各超参改变在本文构建的中文短视频隐喻数据集上 BERT-F1、BLEU-4、ROUGE-L 指标的表现。

（1）训练轮数

训练轮数决定了模型在数据集上的迭代深度，是影响模型拟合能力和过拟合风险的关键因素。在初步实验中，我们观察到 LoRA 微调在第一轮训练中指标上升较快，但在第三至第八轮时逐步趋于收敛。基于此，在本次对比实验中选择了 0、1、4、8 个训练轮数进行评估，其他超参保持变不变，学习率设置为 $5e^{-5}$, 秩设置为 8. 缩放系数设为 16。不同轮数对比结果如下表4.4所示。

表 4.4 不同轮数对比的 LoRA 微调

轮数	BERT-F1	BLEU-4	ROUGE-L
0	67.77	6.46	24.28
1	69.06	11.44	27.62
4	70.52	**12.32**	31.44
8	**70.67**	11.76	**31.49**

由上表可以看出，LoRA 微调在训练轮数提升至 4 时，模型性能趋于稳定。其中 ROUGE-L 相较于一轮微调提升了近 3.8 个点，展现了对文本覆盖性和语义复现能力的显著增强。同时 BLEU-4 和 BERT-F1 也分别实现了可观的增长。当训练轮数继续上升，训练指标出现了一定浮动，比 4 轮的结果稍好，但训练时长过长，因此，本实验选择 4 轮作为 LoRA 微调的最优设定方案。

（2）学习率

学习率作为最为关键的超参数之一，直接影响模型的收敛速度与最终性能表现。较大的学习率收敛速度快，能够使模型快速逼近最优解，但同时也可能导致梯度更新过大，损失函数震荡不稳定，影响模型训练的稳定性和最终性能。而较小的学习率则使模型在训练过程中更新更为平稳，能在后期细致地逼近最优点；但由于收敛速度较慢，可能在训练初期陷入局部最优，训练效率也随之降低。

本实验保持其他超参数不变，秩设置为 8. 缩放系数设为 16。选取四组具有代表

性的学习率设置：1×10^{-5}、3×10^{-5}、5×10^{-5} 和 1×10^{-4}，统一训练一轮，以评估不同学习率对模型在短视频隐喻数据集上的文本生成质量的影响。不同学习率对比结果如下表4.5所示。

表 4.5 不同学习率对比的 LoRA 微调

学习率	BERT-F1	BLEU-4	ROUGE-L
e^{-5}	67.99	8.96	25.96
$3e^{-5}$	**70.36**	11.29	**31.02**
$5e^{-5}$	69.06	**11.44**	27.62
e^{-4}	70.12	11.38	30.29

可以看出，不同学习率对模型性能存在明显影响。当学习率较低（1×10^{-5}）时，模型训练步幅较小，模型未能充分学习隐喻知识，整体指标处于最低水平。而当学习率提升至 1×10^{-4} 时，虽然 BERT-F1 和 ROUGE-L 分数较高，但 BLEU-4 并未持续增长，反映出较大的学习率可能使得模型出现过快震荡、泛化能力下降的风险。

当学习率设为 3×10^{-5} 时，模型在 BERT-F1 和 ROUGE-L 两项指标上均达到 70.36 和 31.02 最高，表明该设定下模型在语义理解和句法覆盖层面上具有最优表现。而在 BLEU-4 指标上，学习率为 5×10^{-5} 时表现略优，达到 11.44，虽然高于 3×10^{-5} 的 11.29，但差异并不显著，因此综合考虑，3×10^{-5} 设定为最优学习率。

（3）秩和缩放系数

秩和缩放系数是 LoRA 微调中两个核心的结构性超参，它们决定了注入到原始模型中的可学习参数的表达能力。秩的大小直接决定了低秩矩阵的维度，也影响了模型在参数插值空间中对原始权重的近似程度。而缩放系数则控制了 LoRA 模块输出对原始权重影响的强度，其大小常与秩成两倍设定，以保持训练稳定性和梯度合理放大。因此，在本实验中统一进行讨论。

本实验在固定学习率为 5×10^{-5} 的条件下，考察 LoRA 秩与缩放系数的组合设置对模型性能的影响。根据前人研究与经验表明，LoRA 的秩在 1 到 16 的范围内对模型效果影响最为显著，而当秩值进一步增大到 32 至 128 时，模型性能提升趋于饱和甚至出现稳定不变的趋势。因此，本实验选取了具有代表性的四组秩值：8、12、16 和 32，分别与其两倍的 Alpha 值（16、24、32、64）进行组合评估。不同秩和对应

缩放系数的对比如下表4.6所示。

表 4.6 不同秩和缩放系数对比的 LoRA 微调

LoRA 参数		BERT-F1	BLEU-4	ROUGE-L
Rank	Alpha			
8	16	69.06	11.44	27.62
12	24	70.19	11.25	30.40
16	32	**70.28**	**11.84**	**30.68**
32	64	70.08	11.25	30.46

随着秩和缩放系数的增大，在秩等于 16，缩放系数等于 32 时，模型性能达到了三个指标参数的最佳，说明此设定下模型兼顾了充分的表达能力和参数效率。当继续增加秩的大小时，学习率反而会略低于秩为 16 的设置，可能是由于参数增多带来的过拟合或训练不稳定所致。可见，适当增加秩与缩放系数能有效增强模型性能，本组实验中最优的配置选择为秩为 16，缩放系数为 32。

4.3.2 基于 LoRA 改进方法的微调实验

LoRA 的成功使研究者进一步探索更具表现力和稳定性的改进方法。在本研究中，为了验证 LoRA 方法在构建的中文短视频隐喻数据集上的可扩展性与优化空间，选取了四种具有代表性的 LoRA 改进技术：LoRA+、DoRA 、rsLoRA 和 PiSSA，对其在同一任务下的微调效果进行了系统性实验对比。

（1）LoRA+

在标准的 LoRA 中，适配器矩阵 A 和 B 使用相同的学习率进行训练。然而，这种统一的学习率设置可能无法充分发挥两部分矩阵的作用。于是 LoRA+ 通过引入一个可调参数 λ，允许为矩阵 A 和矩阵 B 设置不同的学习率，从而实现更细粒度的学习率控制。矩阵 A 学习率为 ηA，矩阵 B 学习率 ηB 为 $\lambda \times \eta A$。在这里实验取等于 $\lambda = 0.1$，使得训练更加稳定。

（2）DoRA

DoRA（Weight-Decomposed Low-Rank Adaptation）是一种在 LoRA 与全量微调之间寻求性能折中的改进方法。尽管标准 LoRA 能大幅降低训练与推理开销，但其

性能仍与全量参数微调存在一定差距。DoRA 提出了一种权重矩阵分解机制，将原始权重 W 分解为单位方向矩阵与尺度系数矩阵的乘积：

$$W = s \cdot \hat{W} \tag{4.1}$$

其中 $\hat{W}$ 表示方向矩阵，s 表示尺度因子。在实际训练中，DoRA 对这两个部分分别进行微调。对于方向矩阵 $\hat{W}$，进一步使用 LoRA 结构进行低秩分解。对于尺度部分 s，则直接学习一个可训练的参数。

（3）rsLoRA

rsLoRA（Rank-Stabilized LoRA）针对 LoRA 微调在高秩设置下容易出现的梯度塌陷问题进行了优化。在传统 LoRA 中，当秩设置较高时，尽管表达能力增强，但模型梯度往往容易出现收缩现象，导致训练过程不稳定，最终效果不佳。为解决这一问题，rsLoRA 在低秩适配器的训练过程中引入更稳定的缩放机制，从而稳定梯度的尺度分布，保障训练过程的数值稳定性。

（4）PiSSA

由于 LoRA 的适配器矩阵 A 使用高斯随机初始化，矩阵 B 全初始化为 0，导致在一开始训练时梯度较小，收敛较慢。基于此，PiSSA（Principal Singular Values and Singular Vectors Adaptation）通过对原权重矩阵奇异值分解进行初始化，其优势在于它可以更快更好地收敛，具体公式如 (4.2) 所示。

$$W = U\Sigma V^T. \tag{4.2}$$

之后选取前 r 个主奇异值及其对应的奇异向量，构建低秩适配器矩阵 A、B：

$$B = U_r\Sigma_r^{1/2}, \quad A = \Sigma_r^{1/2}V_r^T \tag{4.3}$$

经过以上初始化，模型在微调开始阶段就已经包含了原始权重的主要信息，避免了 LoRA 初始阶段输出不变的问题。

本文在统一实验设置下对 LoRA+、DoRA、rsLoRA 和 PiSSA 四种技术进行了对比实验。所有方法均采用相同的超参数配置：秩设为 16，学习率设置为 3×10^{-5}，并统一训练一轮。采用 PiSSA 方法进行微调时，预先生成了用于初始化的适配器矩阵，并将秩设定为 16。下表4.7为不同 LoRA 改进微调方法对比。

表 4.7 不同 LoRA 改进微调方法对比

LoRA 改进方法	BERT-F1	BLEU-4	ROUGE-L
LoRA+	69.94	11.63	**30.75**
DoRA	69.96	11.19	30.04
rsLoRA	**70.30**	**11.75**	30.38
PiSSA	68.28	9.60	27.22

其中，rsLoRA 在 BERT-F1 和 BLEU-4 两项指标上均取得 70.30 与 11.75 的最高值，显示出该方法在语义保持和生成精度方面具有较强的表现力。LoRA+ 在 ROUGE-L 指标上达到了 30.75 最优，说明其在文本结构的还原方面有一定优势。综合来看，rsLoRA 在本轮实验中综合性能最优。

4.3.3 冻结微调实验

冻结微调作为一种最基础的参数高效微调方法，其核心思想是将大部分预训练参数固定，仅对模型顶部的若干层进行训练，显著降低训练计算开销的同时得到良好的微调结果。因此，这里本文固定其他超参数，学习率保持 $5e^{-5}$ 训练 1 轮，对比了在所有模块上不同可训练层数的微调效果，实验表格如下4.8所示：

表 4.8 不同可训练层数对比的冻结微调

可训练层数	BERT-F1	BLEU-4	ROUGE-L
2	69.89	11.28	30.30
4	**70.19**	11.39	29.75
6	70.14	**11.60**	**30.43**

其中，两层可训练层数的参数为 93,595,648 个，四层为 187,191,296 个，六层为 280,786,944 个。通过上表可以看出，冻结微调在调整不同数量的可训练层数时，变化幅度相对平稳，性能提升有一定局限性。从结果上看，随着可训练层数的逐渐增加，模型在三个文本生成指标上大致呈现逐步提升的趋势，尤其是在 BLEU-4 指标上，从 11.28 提升至 11.60，表明模型在语义内容的精确匹配能力方面有所增强。整体来看 6 层训练参数设置在三个指标上具有更好的综合性能。

4.3.4 Galore 微调实验

GaLore 微调是一种结合梯度累积与低秩投影的高效训练策略，其核心思想是通过在梯度更新过程中引入低秩约束。在本次实验中，GaLore 微调 Qwen2-VL-2B 模型的可训练参数量达到 1,543,714,304 个，已接近于全参数微调的规模。训练过程显存占用控制在 23GB 左右，几乎保持最高的显存占用。

为充分挖掘 GaLore 的微调效果，实验保持其他超参不变，学习率为 3×10^{-5}，围绕 GaLore 微调的核心调节因子秩以及更新步数展开。系统评估了不同参数配置下模型在文本生成任务中的表现，旨在选出最优的参数组合，以实现更好的性能与稳定性之间的平衡。

（1）更新步数

GaLore 中采用的梯度投影矩阵按照固定步数进行周期性更新。这个周期就由更新步数控制。本次实验中，在保持其他超参数设置一致的前提下，仅调整更新步数的值，其中更新步数为 10 需更新八次，更新步数为 20 步大约更新四次，50 步更新步数只更新一次。实验固定秩为 16，以观察不同更新步数对模型性能的影响。下表4.9所示为不同更新步数对比的 Galore 微调。

表 4.9 不同更新步数对比的 Galore 微调

Galore 更新步数	BERT-F1	BLEU-4	ROUGE-L
50	70.19	**11.88**	30.39
20	69.78	11.11	29.96
10	**70.33**	11.75	**30.81**

结果显示，当更新步数为 10 时，在 BERT-F1 和 ROUGE-L 指标上取得了更优的结果。更新步数为 50 时 BLEU-4 的得分更高。三种更新步数的区别不大。但相较之下，更新间隔为 20 的设置在本实验中出现了性能下降的趋势。这说明过于频繁的更新可能会引入过拟合或扰乱训练过程，从而影响模型的泛化能力。

（2）秩

GaLore 微调时，梯度会在每一次更新前被投影到一个低秩子空间中。此时，秩控制的是这个低秩空间的维度。本次实验中，在保持其他超参数设置一致的前提下，

仅调整秩的取值，并固定更新步数为 50 进行了对比实验。具体而言，秩分别设定为 4、8、16 和 32，覆盖了从低到高的梯度投影维度配置，以观察不同低秩约束对模型训练能力与泛化性能的影响。下表4.10所示为不同秩对比的 Galore 微调。

表 4.10 不同秩对比的 Galore 微调

Galore 秩	BERT-F1	BLEU-4	ROUGE-L
4	70.30	11.75	30.77
8	**70.68**	**12.20**	**31.23**
16	70.19	11.88	30.39
32	70.24	11.88	30.38

从上表可以看出，随着秩值的增加，模型在各项指标上呈现先上升后略有回落的趋势。当秩为 8 时，模型在 BERT-F1、BLEU-4 和 ROUGE-L 三项指标上均取得了最优结果。当秩进一步增大至 16 和 32 时，指标有所下降或趋于稳定，说明过高的秩可能带来冗余参数，并未显著提升模型效果。

4.4 实验结果与分析

本次微调实验使用七种微调方法对基座模型 Qwen2-VL-2B 进行微调。经过上述实验表明，LoRA 微调在短视频隐喻生成任务上具有明显的优势，最终微调后的 BERT-F1、BLEU-4 与 ROUGE-L 性能指标如下表 4.11 所示。

表 4.11 微调结果性能指标

微调方法	BERT-F1	BLEU-4	ROUGE-L
LoRA	70.52	12.32	31.44

由上表可知，最终微调后的模型在 BERT-F1、BLEU-4 和 ROUGE-L 三项主流自动评估指标上分别达到了 70.52、12.32 和 31.44，表现出良好的语义一致性、表达覆盖度以及一定程度的字面匹配能力。除此之外，在这里选取了一个微调成功案例和一个微调失败的样例进行展示。通过表 4.12 中的具体案例可以进一步验证微调效果的有效性与局限性。

表 4.12 微调结果样例展示

样本	标签	输出
成功	预测	本视频讲述了男子在视频中模仿“避雷”动作并展示“用红酒洗头”的荒诞行为。“避雷”出自网络流行文化，其因为男子的反差式操作和自嘲式回应（“咋整啊?”），表达出了对“避雷”概念的戏谑态度和幽默感。
	真值	本视频讲述了网络用语” 避雷” 的隐喻含义及应用场景。” 避雷” 出自现实生活中的防雷概念，其因为被赋予” 避开不良事物” 的新含义（如劣质商品、糟糕体验等），表达出了网民在信息筛选时警惕、劝诫的实用态度。
失败	预测	本视频讲述了《妈妈，如果你买了 8 个瑞士卷，你怎么分啊》这一话题。” 无糖” 出自日常生活中的食物选择，其因为被赋予了” 健康” 的新含义，表达出了对长辈健康饮食习惯的赞赏与对家庭温馨氛围的温暖感。
	真值	本视频讲述了一位全职宝妈因家庭瑞士卷分配问题与丈夫发生争执的故事。” 八个瑞士卷怎么分” 这个梗出自家庭生活场景，其因为家庭成员间资源分配不均引发的矛盾（女儿、儿子、丈夫各占 1/4 而宝妈权益被忽视），表达出了女性在家庭劳动价值被贬低时的委屈与愤怒情绪。

从成功案例来看，大多数情况下，经过 LoRA 微调的 Qwen2-VL-2B 模型不仅能够准确提取视频中的关键概念，如“避雷”这一网络流行语，还能结合视频中的图像和语境信息，成功进行隐喻分析，揭示出短视频中复杂的语义关系。然而，在某些情况下，模型的表现存在一定的偏差。例如失败案例视频的真实语义核心在于“瑞士卷分配不均”所引发的家庭冲突，而模型却将“无糖”作为关键词进行分析。这一误判可能源自模型在多模态信息对齐过程中的偏差，视觉内容未能准确关联至语义重心，从而导致生成的文本虽然逻辑连贯，但显著偏离视频主题。

总的来说，相较于未经过微调的 Qwen2-VL-2B 模型，通过构建的短视频隐喻数据集进行微调后的模型在性能上有了显著提升。模型在理解和生成隐喻方面表现出了更强的能力，能够更准确地识别和抓取视频中的关键信息和核心概念，在隐喻分析过程中更好地捕捉语境中的深层次含义。

4.5 模型部署与可交互网页展示

为实现短视频隐喻聊天助手的实际应用落地，本实验在模型训练完成后，将微调得到的适配器参数与基座模型 Qwen2-VL-2B 进行合并，并对合并后的完整模型进行导出与切片操作，将每个模型文件控制在约 2GB 以内，并且支持 CPU 与 GPU 的硬件，便于灵活加载与部署。

部署过程中采用 Flask 框架构建后端服务接口，并结合前端交互界面开发，实现完整的用户交互流程。用户无需安装本地模型或配置环境，只需在本地浏览器中访问提供的网页地址，即可通过 Web 界面与部署于云端的模型进行交互。网页功能结构图如下图4.2所示。

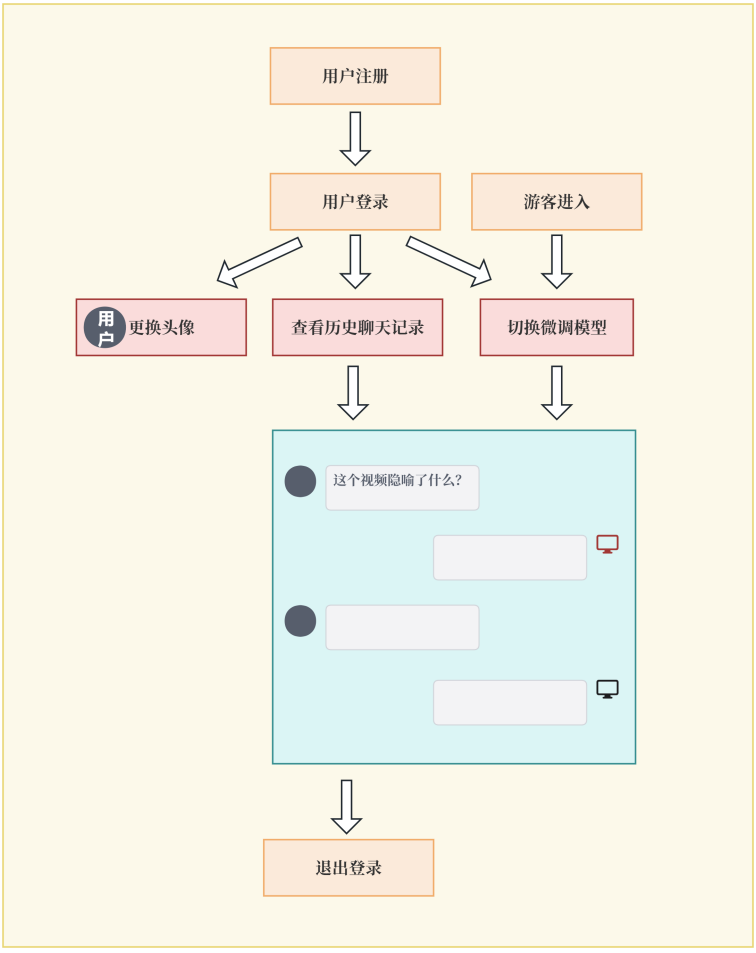


图 4.2 网页功能结构图

网页支持注册登录机制并兼容游客模式访问。用户可以选择注册新账号或使用已有账号进行登录，用户的注册信息和账户数据将被安全地存储于云端数据库中。同

时未登录的用户可直接进入网页，切换不同微调模型与短视频隐喻聊天助手进行交互。注册登录的用户还可以自定义更换用户头像、访问与聊天助手过往的历史对话记录，进一步提升用户的交互连续性与使用便利性。

前端界面展示如下图4.3所示。左上角提供微调模型选择入口，支持实时切换不同的模型配置。右上角则包含用户操作菜单，可以查看历史记录。点击用户头像可以进行更换头像和登出系统的功能。网页底部设有输入框区域，支持用户上传本地短视频文件或输入较长文本内容以与模型进行复杂对话。系统处理完用户请求后，模型生成的回答将展示在对话窗口中，同时在回答区域右下角显示本轮交互的响应时间。

图 4.3 前端界面展示

系统在用户打开网页后默认加载的是 Qwen2-VL-2B 的未微调基座模型，用户可以自己选择不同微调模型进行体验。值得强调的是，虽然这些模型经过了特定任务的微调，但得益于参数高效微调方法，其核心知识能力并未被破坏，原有的大模型通用性得以保留。因此，在增强短视频隐喻理解和生成能力的同时，模型仍具备良好的泛化能力，能够流畅应对开放领域的各种问题，体现出大模型在多任务场景下的强大适应性与实用价值。

在此，展示了一个用户与短视频隐喻聊天助手之间的具体对话过程，以直观体

现模型在理解视频内容与生成隐喻性语言方面的能力。如图 4.4 所示。

图 4.4 与短视频隐喻助手聊天功能展示

切换微调模型后，模型不仅能够按照设定的格式生成规范化的回答，还展现出对视频内容更深入的理解与分析能力。这表明，通过在短视频隐喻数据集上的指令微调，大模型的感知能力与语言生成能力得到了明显增强。

4.6 本章小结

本章介绍了不同微调方法微调大模型的详细实验、结果分析及模型部署。第一节明确了整体实验架构，主要包括选择基座模型、七种主流微调方法的超参微调、实验结果分析和模型部署。第二节通过文本生成的自动化指标与人工评估相结合的方式，对三种最新的多模态大语言模型进行了综合评估，最终选择 Qwen2-VL-2B 作为基座模型。第三节开展微调实验，通过控制变量法对 LoRA、LoRA+、DoRA、rsLoRA、PiSSA、冻结微调和 Galore 微调进行逐一超参调整，以期达到最好的效果。第四节分析了微调实验结果，微调后模型的性能与不足。第五节中将模型部署在云端服务器上，设计出一个可与短视频隐喻聊天助手交互的网页系统。

第五章　总结与展望

鉴于当前短视频隐喻领域的研究尚处于起步阶段，本文聚焦于构建一个面向中文短视频语境的隐喻生成数据集，并基于大模型进行微调训练，从而实现一个具备良好生成性能，能够支持隐喻理解与对话生成的短视频隐喻聊天助手。

本文的主要贡献在于首先构建了面向短视频隐喻任务的中文数据集。数据集涵盖多类语义场景，具有丰富的隐喻表达。同时通过商用大模型处理数据格式以及人工校审加强隐喻标签的细节辅助增强数据，为后续短视频隐喻研究提供了高质量的训练语料，填补了当前领域在该方向的空白。随后，本文提出并验证了多种参数高效微调方法在大语言模型上对短视频隐喻生成任务的应用效果。实验表明，不同微调策略在隐喻生成能力方面存在显著差异，为大模型在视频隐喻生成任务中的应用提供了新思路。最后，设计并实现了一个可落地的短视频隐喻生成系统。该系统通过网页使模型与用户进行交互，展示了大模型在高阶语义生成任务中的实际应用潜力。

然而，尽管本文已在多个层面展开了探索，但仍存在以下不足与可改进之处：

（1）在数据构建方面，当前的数据集规模与内容仍有扩展空间。由于隐喻的理解与生成高度依赖丰富的常识性知识、文化语境与社会背景，单纯依赖原始语料难以支撑复杂语义建模。未来工作中，可引入结构化知识库或知识图谱，构建知识增强型隐喻语料资源，为模型提供更强的语义支撑，提升其对隐喻的准确建模与生成能力。

（2）在模型能力方面，当前所使用的大语言模型在短视频隐喻的理解与生成任务中仍存在一定的性能瓶颈。近年来，扩散模型[48]逐渐兴起，并在图像生成等领域展现出卓越的表达建模能力。今年出现了新的扩散大语言模型[49]将扩散机制引入到语言模型中，通过逐步去噪的生成过程，使模型在语义构建上更为精细，尤其在处理复杂、模糊与抽象的语言现象时具备天然的适应性与表达力。基于此，未来可进一步探索扩散大语言模型在短视频隐喻任务中的适用性与性能潜力。

总的来说，本文为短视频隐喻生成任务的研究提供了初步的理论与实践探索，未来期待研究者在短视频隐喻领域继续开疆扩土，推动该领域的进一步发展。

参考文献

[1] KREUZ R J, ROBERTS R M. The empirical study of figurative language in literature[J]. Poetics, 1993, 22(1-2): 151-169.

[2] LAKOFF G. The contemporary theory of metaphor[Z]. 1993.

[3] HUSSAIN Z, ZHANG M, ZHANG X, et al. Automatic understanding of image and video advertisements[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1705-1715.

[4] MCQUARRIE E F, MICK D G. Visual rhetoric in advertising: Text-interpretive, experimental, and reader-response analyses[J]. Journal of consumer research, 1999, 26(1): 37-54.

[5] MORENO R, MAYER R E. A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia instructional messages.[J]. Journal of Educational psychology, 2000, 92(1): 117.

[6] MESSARIS P. Visual persuasion: The role of images in advertising[M]. sage publications, 1996.

[7] 张驰, 郭媛, 黎明, 等. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用, 2021, 57(11): 57-69.

[8] BIRKE J, SARKAR A. A clustering approach for nearly unsupervised recognition of nonliteral language[C]//11th Conference of the European chapter of the association for computational linguistics. 2006: 329-336.

[9] STEEN G J, DORST A G, HERRMANN J B, et al. Metaphor in usage[M]. Walter de Gruyter GmbH & Co. KG, 2010.

[10] MOHLER M, BRUNSON M, RINK B, et al. Introducing the lcc metaphor datasets[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016: 4221-4227.

[11] 赵秀凤. 概念隐喻研究的新发展——多模态隐喻研究[J]. 外语研究, 2011, 1(1).

[12] PETRIDIS S, CHILTON L B. Human errors in interpreting visual metaphor[C]//Proceedings of the 2019 Conference on Creativity and Cognition. 2019: 187-197.

[13] INDURKHYA B, OJHA A. An empirical study on the role of perceptual similarity in visual metaphors and creativity[J]. Metaphor and Symbol, 2013, 28(4): 233-253.

[14] ACHLIOPTAS P, OVSJANIKOV M, HAYDAROV K, et al. Artemis: Affective language for visual art[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11569-11579.

[15] CHAKRABARTY T, SAAKYAN A, WINN O, et al. I spy a metaphor: Large language models and diffusion models co-create visual metaphors[A]. 2023.

[16] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents: Vol. 1[A]. 2022: 3.

[17] ZHANG D, ZHANG M, ZHANG H, et al. Multimet: A multimodal dataset for metaphor understanding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3214-3225.

[18] HWANG E, SHWARTZ V. Memecap: A dataset for captioning and interpreting memes[A]. 2023.

[19] ANURUDU S M, OBI I M. Decoding the metaphor of internet meme: A study of satirical tweets on black friday sales in nigeria[J]. AFRREV LALIGENS: An International Journal of Language, Literature and Gender Studies, 2017, 6(1): 91-100.

[20] AKULA A R, DRISCOLL B, NARAYANA P, et al. Metaclue: Towards comprehensive visual metaphors research[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 23201-23211.

[21] YOSEF R, BITTON Y, SHAHAF D. Irfl: Image recognition of figurative language[A]. 2023.

[22] KALARANI A R, BHATTACHARYYA P, SHEKHAR S. Unveiling the invisible: Captioning videos with metaphors[A]. 2024.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[24] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[25] LI Y, LIN C, GUERIN F. Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling[C]//Proceedings of the 29th international conference on computational linguistics. 2022: 6468-6479.

[26] 张鹤, 王鑫, 韩立帆, 等. 大语言模型融合知识图谱的问答系统研究[J]. 计算机科学与探索, 2023: 1.

[27] YOUN J, TAGKOPOULOS I. Kglm: Integrating knowledge graph structure in language models for link prediction[A]. 2022.

[28] LI Y, WANG S, LIN C, et al. Framebert: Conceptual metaphor detection with frame embedding learning[A]. 2023.

[29] BAKER C F, FILLMORE C J, LOWE J B. The berkeley framenet project[C]//COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics. 1998.

[30] SPEER R, CHIN J, HAVASI C. Conceptnet 5.5: An open multilingual graph of general knowledge [C]//Proceedings of the AAAI conference on artificial intelligence: Vol. 31. 2017.

[31] STOWE K, CHAKRABARTY T, PENG N, et al. Metaphor generation with conceptual mappings [A]. 2021.

[32] LEWIS M, LIU Y, GOYAL N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[A]. 2019.

[33] SINGH I, BARKATI A, GOSWAMY T, et al. Adapting a language model for controlled affective text generation[A]. 2020.

[34] 王婷, 王娜, 崔运鹏, 等. 基于人工智能大模型技术的果蔬农技知识智能问答系统[J]. 智慧农业, 2023, 5(4): 105.

[35] 潘雪峰, 王超, 卢智增. ChatGPT 在健康谣言鉴别中的实证探讨与应用展望[J]. 情报探索, 2024, 1(1): 1.

[36] XU X, YAO B, DONG Y, et al. Mental-llm: Leveraging large language models for mental health prediction via online text data[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2024, 8(1): 1-32.

[37] CHEN Z, WANG W, CAO Y, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling[A]. 2024.

[38] ZHANG B, LI K, CHENG Z, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding[A]. 2025.

[39] WANG P, BAI S, TAN S, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution[A]. 2024.

[40] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation[A]. 2021.

[41] AGHAJANYAN A, ZETTLEMOYER L, GUPTA S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning[A]. 2020.

[42] QIN Y, WANG X, SU Y, et al. Exploring universal intrinsic task subspace via prompt tuning[A]. 2021.

[43] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models.[J]. ICLR, 2022, 1(2): 3.

[44] ZHAO J, ZHANG Z, CHEN B, et al. Galore: Memory-efficient llm training by gradient low-rank projection[A]. 2024.

[45] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[46] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.

[47] ZHANG T, KISHORE V, WU F, et al. Bertscore: Evaluating text generation with bert[A]. 2019.

[48] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[49] NIE S, ZHU F, YOU Z, et al. Large language diffusion models[A]. 2025.

致　谢

写到这里，代表着我的大学生活即将告一段落。转眼间，四年的时光就这样在一件件堆叠的小事中溜走了。

首先我要感谢我未来读研的导师和师兄——天津大学刘安安老师和修贤师兄对我论文提供的灵感和帮助，不仅为我提供了做实验的服务器，还在我遇到问题时给予我技术上的指导和支持。也要感谢上海大学吴汉舟老师在每周毕设组会上给我毕设提供的宝贵建议。

感谢我的室友李澜鑫、何文慧、李梦茜。来到通信学院后，和你们的每一次欢声笑语我都记在脑海里。因为有你们的存在，那些一起熬夜做项目、一起复习的日子也变得充满回甘。

感谢上大学以来遇到的新朋友和从小一直玩到现在的老朋友们，你们对我的爱让我能够一次一次勇往无前。

感谢吴凡同学在大学的最后能和我相遇，让我的大学生活又增添了一抹亮丽的色彩。所谓相逢恨晚，希望未来我们能继续并肩而行。

最后感谢父母这么多年的养育和关怀，是你们对我的支持、鼓励以及在我做每个决定时给我的参考意见，让我变成了更好的自己。感谢弟弟常常关心挂记我这个姐姐。

大学四年是我人生美好的回忆与宝贵的财富，未来我将带着我的所学所思所感，继续上路出发。

谢雨梦

上海大学

2025 年 5 月 11 日


上海大学
SHANGHAI UNIVERSITY