大模型标注基础上中国文化类视频的多模态话语意义构建研究

秦子敏[1]

（1 北方工业大学，北京 100144）

摘　要：本研究基于 Kress & van Leeuwen 的视觉语法理论和 Halliday 的系统功能语法理论，以 Chat GPT 4o 为标注工具，深入探讨了中国文化类视频《筷子：这个餐具绝没那么简单！》中的视觉模态与语言模态如何进行意义构建。通过对该视频中多模态话语的细致解读，研究挖掘了其中所蕴含的文化意义，以及 Chat GPT 4o 在标注准确性、多模态语言处理能力及指令理解精度方面的局限性。希望本文能为内容创作者在讲述中国故事时提供参考价值，同时为大模型开发与训练提供有益的启示。

关键词：多模态话语分析；国际传播；中国文化；大模型

Doi：doi.org/10.70693/rwsk.v1i5.930

一、引言

在 2021 年 5 月 31 日举行的中国共产党第十九届中央政治局第三十次集体学习中，习近平总书记着重指出，讲好中国故事，传播好中国声音，展示真实、立体、全面的中国，是加强我国国际传播能力建设的重要任务[1]。随后，各类讲述中国文化的视频及其多模态话语研究引起了人们的重视。

然而，早期的多模态话语分析多是提出某个分析框架并选取少量数据进行示例，结合 Elan 等软件进行手动标注。近年来，多模态话语研究正逐步进入全新的大数据阶段。李梦洁等指出，未来研究可结合如 ChatGPT, Gemini 等人工智能工具，以提升自动处理多模态语料的效率，提供实时数据分析反馈，并优化数据的可视化效果[2]。

二、理论基础

（一）视觉语法理论

视觉语法是社会符号学领域的重要理论。该理论借鉴了 Halliday 关于语言三大功能的假说，并结合电影研究的相关理论，提出了与这三大功能相对应的视觉图像分析的三个层面：再现意义、互动意义和构图意义[3]。图像的再现意义分为叙事性再现和概念性再现。图像的互动意义是指图像符号如何构建并维持观看者与图像中人物的关系。包括接触、社会距离以及视角和情态[4]。图像的构图意义涵盖了三个方面：信息值、显著性和取景。

（二）系统功能语法

根据 Halliday 的系统功能语法，语言具有三大元功能，即表示概念意义的概念功能，表示说话人和听话人关系以及说话人对所说内容的态度的人际功能，以及表示语篇意义的语篇功能。其中，概念功能主要通过及物系统来体现。人际功能的实现主要是通过语气、情态、时态和人称代词来完成的。语篇功能主要通过主要通过主位结构、信息结构和衔接的方式得以体现[5]。

三、研究设计

（一）研究对象

《中国范儿》(Feel of China)是由中国网精心制作的一档双语微视频系列，节目采用英语解说和中英文双语字幕。该视频是典型的多模态话语，以语言模态和视觉模态为主，听觉模态为辅，深入浅出地介绍了节气、服饰、传统工艺、节日等文中国特色文化符号，展现了中华文化的深厚底蕴和广泛影响。在具体分

作者简介：秦子敏(2000—)，女，硕士，研究方向为多模态话语分析与国际传播。

析时选取了其中的《筷子》篇为例，对其中的图像、文字符号进行分析，探究其意义构建。

（二）语料标注与统计

大语言模型（Large Language Model, LLM）是基于深度学习的自然语言处理模型，能够理解和生成自然语言文本。通过分析大量文本数据进行训练，这些大模型可对文本、语音、图像等进行分析和理解，在文献阅读与总结、情感分析、编写代码、机器翻译、访谈分析、语言和话语分析、以及多模态语言处理等方面表现出极大的应用前景[6]。常见的大语言模型有 Chat GPT、文心一言、智谱清言、kimi 等。

本研究采用视觉语法和系统功能语法对《筷子》篇进行分析，结合前沿的大模型 Chat GPT 4o 版本进行标注(数据取自 2024 年 9 月 20 日至 28 日)。相较于之前的版本，Chat GPT 4o 能够更准确地理解上下文，从而提供更相关的回答。同时在多模态处理、多语言处理、个性化定制等方面均有显著提升，能够很好地实现本研究的目的。

具体而言，在图像抓取阶段，Chat GPT 4o 可以从视频当中随机截取关键帧，避免人为挑选有意义帧产生的主观性；在标注阶段，基于视觉语法和系统功能语法的理论框架，可以提升标注效率;在数据统计阶段，可以生成图表，帮助标注者处理一些耗时低效的重复性工作。由于其无法像人类一样灵活应变，在截取图像时可能会抓取很多重复的画面以及转场画面，并且生成的内容可能产生误导性，所以需要人工校对，以保证标注结果的准确性。笔者用 Chat GPT 4o 从视频中抓取了 50 个关键帧，人工剔除掉重复帧后剩余 39 帧及其对应的字幕用于分析。

（三）研究问题

本研究通过分析大模型标注基础上中国文化类视频《筷子》中的多模态话语意义构建，旨在回答以下几个问题:

该视频中的视觉模态和语言模态如何进行意义构建?

大模型在标注视频的过程中存在哪些局限性与不足?

四、《中国范儿》的多模态话语分析

（一）再现意义分析

表 1 再现意义分布

分类	过程类型	标注次数	占比
叙事性再现	行为过程	23	59.0%
	心理过程	1	2.6%
概念性再现	分类过程	4	10.3%
	象征过程	11	28.2%

根据 Kress & van Leeuwen，再现意义分为叙事性再现和概念性再现两种，有无矢量是区分两者的标志。叙事性再现存在矢量，再现行为或事件，强调事件的展开和变化的过程，此类图像通常叙述一个故事或解释一个过程，体现的是“做什么”。概念性再现没有矢量，再现的则是某事物具有普遍意义的、稳定的本质，体现的是“是什么”。叙事性再现可以分为行为过程、反应过程、言语过程和心理过程，而概念性再现可以分为分类过程、分析过程和象征过程[3]。视频的主题是筷子，因此大部分场景都是人使用筷子的画面，而人和筷子之间本身就是一种矢量关系，所以视频中大部分都是叙事性再现中的行为过程，比如图 1 的团圆饭上，小孩用筷子给老人夹取食物，大人、小孩和筷子之间分别存在矢量，因此属于行为过程。而且老人（反应者）满是欣慰地看向正在夹菜的小孩（现象），他们的目光也形成了矢量，构成反应过程。这一幕生动地展示了家庭的代际关系和文化传承，与此同时，老人小孩坐在中间，大人位于两边，体现了长幼有序的中华传统美德。

图 1

（二）互动意义分析

1. 接触

表 2 接触类型分布

分类	标注次数	占比
索取	2	5.1%
提供	37	94.9%

Kress & van Leeuwen 指出，接触包含两个要素，索取和提供[3]。如果图像中的参与者直视观看者，与其进行了眼神交流，则表达了索取的意义；反之，如果图像中的参与者并未直视观看者，没有眼神交流，则为提供。这两者的区别主要在于图像中的参与者与观看者是否具有眼神交流。由于视频的目的是展示筷子所蕴含的文化内涵，即提供信息，所以大部分图像表达的都是“提供”。

图 2

图 2 中有四位参与者，参与者之间的目光形成了矢量，具有眼神交流，但参与者与观看者之间并没有眼神接触，因此，这个画面表达的是“提供”的意义。结合视频内容可知，此时正值春节时期，中间那位邻居独自在家过年，于是主人招呼他一同过来用餐，周围的主人都在用筷子给客人夹菜，彼此之间存在眼神交流，客人眼含泪花，满是兴奋与激动。筷子不仅提供了餐食，也提供了邻里之间的温情，这也是邻里和睦和团圆的中国文化内涵。

2. 社会距离

表 3 社会距离分布

分类	标注次数	占比
特写镜头	29	74.4%
中景镜头	9	23.1%
长景镜头	1	2.5%

社会距离与镜头的框架有关，可以分为特写镜头，中景镜头和长景镜头。特写镜头表示亲密或个人关系。中景镜头表示社会关系，长景镜头表示公共关系[3]。《筷子》篇中特写镜头居多，主要为了展示筷子的细节和使用、还有美食与烹饪等与筷子紧密联系的元素。此外，图 3 展示了中国人与外国友人共享美食的场景，特写镜头表明两人之间的空间关系较为亲近，这可能暗示了他们之间的亲密关系或友好的社交互动。两人的表情都很轻松，给人一种轻松愉悦的氛围，拉近了参与者与观看者之间的关系，召唤了观众的情感，有利于增强传播效果。画面与配文“文明多姿，世界才能多彩”协同作用，通过展示不同文化背景下的人们和谐共处，强化了文化多样性对于世界的重要性。

图 3

3. 视角

视角可以分为水平视角和垂直视角。其中水平视角由正面视角和倾斜视角构成。正面视角给观看者带来一种卷入之感，倾斜视角则给观看者一种超然之感。垂直视角包括仰视、俯视和平视。不同视角反应了参与者与观看者之间的权势关系，俯视表明权势在观看者一方，仰视表明权势在参与者一方，平视则说明观看者和参与者之间是平等的关系[3]。

图 4

图 5

这两张图均为倾斜视角，观看者可以从侧面感受到筷子的精致，即筷子可以用精密的雕刻刀进行题诗，作画，刻字，并非只是作为餐具那么简单。这种视角使得观看者能够清晰地看到筷子上的细微纹理，体现了手工制作的工艺美。这些细节展示了工匠对品质的追求和对中国传统文化的尊重。

4. 情态

在 Kress & van Leeuwen 的视觉语法中，情态指的是世界的真实性和可信性。而图像细节的最小和最大表示(如颜色饱和度和丰富度)影响信息的可靠性，可细分为高情态、中情态和低情态。在色彩饱和度方面，较高的情态意味着更大的可信度和真实性，而较低的情态则更加抽象，超出了世界的真实性[3]。

视频中的筷子均取材于温暖的黄棕色竹子，而这也是筷子最初的色泽。《中国范儿》视频的目的是让世界了解更真实的中国，选用这样的颜色作为筷子的呈现，不仅最贴合人们的传统认知，还再现了筷子的自然本质，这种高情态给人以真实自然之感。而图 6 主要描绘了西周时期的“匕匙”这一具体文物，这两把“匕匙”带有铜绿的氧化痕迹，展示了金属的质感和岁月的痕迹，这种高情态十分具有历史真实性。勺子的造型和纹饰展示了当时的工艺水平和社会审美观念，与配文“可切可捞”协同作用，表明其作为饮食器具的功能和历史价值，反映了古代中国文明的发达程度和人类智慧。

图 6

（三）构图意义分析

1. 信息值

信息值靠图像中元素的构图位置来实现。主要分为上下、左右、中心和边缘[3]。根据人的阅读习惯，左边的信息是已知的，而右边的则是新信息。上方的信息往往是理想的，而下方的信息往往是现实的。位于中心的信息是核心的，而位于边缘的信息是从属的。

图 7

图 7 则是中心-边缘信息结构，这种分布符合人的一般认知顺序，有利于观看者对重要信息的提取，并给视频留下深刻的印象。参与者包括两位老人、两个成年人和两个孩子，他们构成了一个典型的三代同堂的家庭结构。老人坐在中间的位置即“主座”是一种尊重和礼貌的表现，强调了他们在家庭中的重要地位。而这也是中国敬老爱老的宴席礼仪。每位家庭成员都满是兴奋与激动，体现了阖家团圆的中国文化内涵。

2. 显著性

显著性指的是用图像中的元素来吸引观看者，可以通过大小、聚焦程度、色调对比、色彩对比、前景化及某些文化因素实现[3]。在这个视频中，图像主要通过大小，前景化和色彩对比来吸引观看者。比如图中的红衣服、红福字和大红灯笼都突出了人们对来年美好生活的向往和春节的喜庆。而图 5 则用前景化及模糊背景的方式突出了筷子上图画的精妙绝伦以及作画的精细，引导观众将注意力聚焦于正在被雕刻的筷子，增强了视觉效果。此外，视频中还善用特写镜头使得美食前景化，刺激观看者的味蕾，使其对中式美食印象深刻。

3. 取景

取景是指图像中是否存在空间分割线条，这些线条表示图像中各成分之间在空间上被分离或被连接的关系。可以通过矢量、颜色和视觉形状等连接或分离不 同的形状[3]。图 8 是一个庭院，石栏杆和石柱作为分割线，将院子框起来分为院内和院外。它们使得这幅图有了聚焦，吸引观看者将目光聚焦在院内手拿烟花追逐嬉戏的孩子们，突出了春节热闹的气氛。这种空间布局具有一定的对称性，体现了中国人对和谐与平衡的追求，符合“中庸之道”的设计理念。

图 8

（四）概念意义分析

概念意义主要通过及物系统来体现。Halliday 将其细分为六个具体的过程，即物质过程、心理过程、言语过程、行为过程、存在过程和关系过程[5]。笔者用 Chat GPT 4o 将关键帧对应的字幕进行转录和标注，示例如下：

1. Chopsticks can take on themselves a variety of art forms such as inscriptions, paintings, pyrography, inlays

and carvings.

2. Eating with chopsticks is a real trick for first-time learners.

3. The world becomes colorful thanks to diversified civilizations, and the cultural tensions created by differences possess a unique form of charm.

4. They are the crystallization of Chinese wisdom as well as a symbol of the oriental civilization.

5. First of all, it's because many Chinese food are served hot, making it necessary for the diner to use a tool to avoid direct contact with hot food.

以上为 Chat GPT 4o 随机抓取的五个句子，其中小句 1 和 5 是物质过程，5 中动词是 are served,目标是 many Chinese food,这句描述了使用筷子的缘由和必要性。物质过程常与使用筷子就餐的画面一同出现，在视频中出现的频率仅次于关系过程。小句 2、3、4 都是关系过程，且为包孕型过程归属式。关系过程主要用于强调参与者之间的关系，在此类宣传中国文化视频的中很常见。视频制作者使用关系过程不仅将筷子与其背后的饮食文化、团圆文化建立联系，还能进一步与世界文明产生联结。比如小句 4 表明筷子是中国智慧及东方文明的象征，强调其背后的文化内涵。小句 3 的配图是外国友人和国人一同使用筷子用餐的画面，画面结与字幕协同作用，共同强调了文化差异所带来的文化张力别有韵味。

（五）人际意义分析

Halliday 认为语言的人际功能是指除了具有表达讲话者的亲身经历和内心活动的功能外，还具有表达讲话者的身份、地位、态度、动机和其对事物的推断、判断和评价的功能。人际功能的实现主要是通过语气、情态、时态和人称代词来完成的。其中，语气有四种基本类型，分别是陈述语气、疑问语气、祈使语气和感叹语气[5]。因为陈述语气的主要目的是提供信息，而《中国范儿》系列视频的初衷是选取独具中国特色的文化符号进行展示和介绍，让世界了解更真实的中国，感受中国文化底蕴。所以《筷子》篇中的人际意义主要靠陈述语气来实现。比如，小句 1 用陈述语气展示了筷子的多姿多彩，可以融合题词、刻诗、绘画、烙画、镶嵌、雕镂等艺术形式。而小句 6 中疑问语气的使用则能够吸引观众的注意力，鼓励观众参与到这种文化体验当中，使其对使用筷子的缘由印象深刻。

6.Why do Chinese people eat with chopsticks?

（六）语篇意义分析

语篇意义主要通过主要通过主位结构、信息结构和衔接的方式得以体现，小句 1、2、3、4 的主位均是非标记性主位，而小句 5 中的“First of all”则是标记性主位。这些主位无论是“Chopsticks”“Eating with chopsticks”还有“They”都将筷子分解成了具体的几个方面进行阐述，通过述位展示了筷子背后的文化属性。同时，用“They”代替“Chopsticks”作为主位，这种切换促进了语篇的连贯性。而标记性主位的使用则强调了使用筷子的缘由，反映了中国饮食文化中热菜的传统，强调了烹饪和食用方式的独特性，更是中国文化的象征，体现了民族的智慧和生活方式。

五、大模型在标注过程中的局限性与不足

Chat GPT 4o 在标注过程中除了带来了一些积极的影响，比如提升效率、避免主观性等，也存在某些局限性与不足。

第一，标注结果无法做到完全准确。Chat GPT 4o 主要是基于自身的数据库，结合预训练的模型对语料进行标注。但该模型无法像人一样进行复杂的思考，缺乏灵活应对的能力，导致在标注时存在差异。比如 Chat GPT 4o 在随机截取关键帧时会将许多重复帧或者无意义的转场画面收录进去，从而产生误导，影响结果的准确性。

第二，对多模态语言的处理尚未完善。Chat GPT 4o 尚不具备直接标注语音、视频等多模态语言的能力，输入指令后得到的大都是搭载某个软件提取图像、音频，再输入指令进行标注的回复。与此同时，Chat GPT 4o 未能大规模处理多模态语料，大批次输入语料后的标注结果误导性很大，不如分批标注准确。

第三，对指令的理解精度有限。在实际操作过程中，有一些内容在 Chat GPT 4o 预训练的数据库以外，但其在理解指令时并不具备处理这些区域以外的能力。比如，在生成可视化图表的过程中，笔者想要进行一下修改，让其按照手绘图表的样示进行调整，而生成的结果却大相径庭。

六、 结语

本研究以中国文化类视频《筷子：这个餐具绝没那么简单! 》为研究对象，采用 Chat GPT 4o 进行标注，

并基于 Kress & van Leeuwen 的视觉语法和 Halliday 的系统功能语法理论，对视频中的视觉模态与语言模态进行深入分析。研究发现，视频通过叙事性再现和概念性再现，构建了筷子的使用过程和文化内涵；在互动意义上，揭示了视频与观众之间的互动关系和文化价值观的传递；在构图意义上，探讨了视频如何通过信息值、显著性和取景等视觉修辞手法，实现意义构建。在语言模态上，《筷子》善用关系过程来展示筷子与其他中国文化之间的关系，以及陈述语气来介绍中国文化，而主位述位进行切换也保证了语篇的连贯性，此外，本研究还评估了 Chat GPT 4o 在标注过程中的局限，例如标注准确性、多模态语言处理能力及指令理解精度等方面。研究结果表明，大模型标注技术为多模态话语分析提供了新的工具和思路，但仍需进一步探索和改进，以提升其在学术研究中的应用价值。

参考文献：

[1] 袁小陆，乃瑞华.“文化中国”国际传播多模态话语意义建构研究[J].外语教学, 2022, 43(05): 23-29.
[2] 李梦洁，冯德正，邓谊.国际多模态话语分析的进展与前沿[J].现代外语, 2024, 47(03): 419-430.
[3] KRESS, G.& T. van Leeuwen. .*Reading Images: The Grammar of Visual Design* (second edition)[M]. London: Routledge. 2006: 179-224.
[4] 李战子.多模态话语的社会符号学分析[J]. 外语研究, 2003(5)：1-8+80.
[5] Halliday, M.A.K..L. *An Introduction to Functional Grammar* (second edition)[M]. Beijing: Foreign Language Teaching and Research Press.2000: 21-76.
[6] 许家金，赵冲，孙铭辰.大语言模型的外语教学与研究应用[M]. 北京:外语教学与研究出版社.2024: 1-127.
[7] QIN Zimin. Multimodal Discourse Analysis of Bilingual Micro-video “Feel of China” from the Perspective of International Communication[J]. 言語と文化の研究，2024:88-103.

Study on the Meaning Construction of Multimodal Discourse in Chinese Cultural Videos Based on LLMs

Qin Zimin1

1School of Humanities and Law, North China University of Technology, Beijing 100144, China

Abstract: This study is based on Kress & van Leeuwen’s visual grammar and Halliday’s systemic functional grammar.Using Chat GPT 4o as an annotation tool, it explores how the visual modality and verbal modality of the Chinese cultural video construct meaning. Through the detailed interpretation of the multimodal discourse in the video, this study shows the cultural meaning embedded in it. It further analyzes the shortcomings of Chat GPT 4o, pointing out its limitations in terms of annotation accuracy, multimodal language processing capability, and comprehension precision of instruction. It is hoped to provide reference value for content creators in telling Chinese stories, as well as useful insights for large language model’s development and training.

Keywords: multimodal discourse analysis; international communication; Chinese culture; large language models