**Modern Linguistics** 现代语言学, 2025, 13(9), 561-573
Published Online September 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.1391013

Hans汉斯

# 基于语义匹配的跨模态图文检索与生成研究

**步英杰[1,2*]，伍　乙[1,3]，陈　锐[1,4]，贺国栋[1]，陈　伟[1]**

[1]温州商学院信息工程学院，浙江　温州
[2]澳门大学协同创新研究院，澳门
[3]浙江工商大学萨塞克斯人工智能学院，浙江　杭州
[4]悉尼大学工程学院，澳大利亚　悉尼

收稿日期：2025年6月9日；录用日期：2025年9月5日；发布日期：2025年9月17日

## 摘　要

**在语义学的研究框架中，语言被视为意义的载体，而图像则是感知的表征。随着人工智能的发展，文本与图像之间的跨模态研究成为了新的热点。本文聚焦于文搜图的任务，从语义学角度分析文本如何通过模型转化为相应的图像。具体来说，本文提出了一种基于语义组合特征的CLIP + MLP模型，用于提升模型在细粒度语义对齐中的表现。此外，本文通过构造多层次文本描述，比较原始CLIP模型与CLIP + MLP模型的语义匹配能力，并采用Stable Diffusion 1.5模型进行抽象语义的图像生成测试。结果表明，CLIP + MLP模型在复杂语义结构下表现更优，而Stable Diffusion 1.5模型在抽象风格与隐喻语言的还原中亦展现出一定的语义建构能力。总而言之，本文验证了语义特征建模在跨模态任务中的关键作用，为语义驱动的图文理解提供了有力支持。**

## 关键词

**跨模态，文搜图，语义匹配，相似度计算，CLIP模型，CLIP + MLP模型，Stable Diffusion 1.5模型**

# Research on Cross-Modal Graphic Retrieval and Generation Based on Semantic Matching

**Yingjie Bu[1,2*], Yi Wu[1,3], Rui Chen[1,4], Guodong He[1], Wei Chen[1]**

[1]School of Information Engineering, Wenzhou Business College, Wenzhou Zhejiang
[2]Institute for Collaborative Innovation, University of Macau, Macau
[3]Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou Zhejiang
[4]Faculty of Engineering, University of Sydney, Sydney, Australia

Received: Jun. 9th, 2025; accepted: Sep. 5th, 2025; published: Sep. 17th, 2025

*通讯作者。

文章引用: 步英杰, 伍乙, 陈锐, 贺国栋, 陈伟. 基于语义匹配的跨模态图文检索与生成研究[J]. 现代语言学, 2025, 13(9): 561-573. DOI: 10.12677/ml.2025.1391013

## Abstract

**In the research framework of semantics, language is regarded as a carrier of meaning, while images are representations of perception. With the development of artificial intelligence, cross-modal research between text and image has become a new hot spot. This paper focuses on the task of text searching for images, and analyzes how text can be transformed into corresponding images through models from a semantic perspective. Specifically, this paper proposes a CLIP + MLP model based on semantic combination features for improving the performance of the model in fine-grained semantic alignment. In addition, this paper compares the semantic matching ability of the original CLIP model and the CLIP + MLP model by constructing multi-level text descriptions, and tests the image generation of abstract semantics using the Stable Diffusion 1.5 model. The results show that the CLIP + MLP model performs better under complex semantic structures, while the Stable Diffusion 1.5 model also shows some semantic construction ability in the reduction of abstract style and metaphorical language. All in all, this paper validates the key role of semantic feature modeling in cross-modal tasks and provides strong support for semantics-driven graphic understanding.**

## Keywords

**Cross-Modality, Text Search Map, Semantic Matching, Similarity Calculation, CLIP Model, CLIP + MLP Model, Stable Diffusion 1.5 Model**

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

## 1. 引言

随着自然语言处理(NLP)与计算机视觉(CV)技术的迅速发展，“从文本到图像”的跨模态研究已成为当前人工智能领域的核心任务之一[1]。近年的研究也指出，跨模态检索与生成在图像搜索、视频理解和多模态对话系统等场景中展现了重要应用前景[2]。这一任务不仅体现了模型对语言与图像的综合理解能力，也深刻揭示了人类语义表达与视觉感知之间的内在联系。其中，文搜图技术尤为关键，该技术是指根据自然语言描述检索或生成相应图像[3]。

目前，主流的文搜图方法主要可分为两大类：其一为检索型方法，通过将文本与图像映射到同一语义向量空间中，计算其相似度，从已有图像库中筛选出最匹配的图像；其二为生成型方法，直接基于输入文本生成全新的图像，现如今常用的就是扩散模型实现从语言到视觉的转换。

如今随着技术的不断进步，文搜图技术已被广泛应用于图像搜索引擎、AI 绘画平台等多个实际场景，并在学术与工业界均展现出广阔的发展前景[4]。因此，本文将围绕文搜图任务展开研究，系统探讨其语义基础、模型机制及文本与图像语义相似度的计算方法，并通过实验对比检索型与生成型模型在语义表达层面的表现差异。

## 2. 数据集来源与预处理

### 2.1. 数据集来源

本文所涉及的实验使用的图像数据集是来自 Kaggle 官网的公开数据集 Animals-10，该数据集中包含 10 个动物类别的图像，分别为：马、狗、猫、蝴蝶、鸡、大象、牛、羊、松鼠和蜘蛛，总计超过 26,000

张图像。选取该数据集进行实验的原因，是因为其具备多样性和复杂性。数据集中的图像在角度、背景、光照、清晰度等方面存在显著差异。部分图像拍摄于自然环境，另一些则具有人工背景，有的存在局部遮挡或构图不清。这种数据集可以为跨模态语义建模提供更具挑战性的测试场景。

### 2.2. 数据集预处理

考虑到计算资源的限制，本文所涉及的实验从 Animals-10 数据集中按类别随机抽取每类 500 张动物图像样本，共计 5000 张图像，以在保证类别平衡的前提下控制实验规模。抽取的图片构建了一个语义分布均衡的小型图库，该小型的图库用作跨模态匹配与生成的图像输入。

图像数据在使用前统一进行了尺寸调整(224 × 224 像素)和标准化处理。文本部分的描述数据由人工撰写，旨在模拟真实图文检索场景中的自然语言输入，以支持后续的语义相似度计算与模型匹配实验[5]。

## 3. 相关理论基础

### 3.1. 语义学基础

#### 3.1.1. 语义匹配

语义匹配指的是两个表达在意义层面上的相似性。在语言学中，它通常涉及两个词汇、短语或句子是否在语义上具有等价或相似的解释[6]。其核心目的是判断二者在语义空间中是否表达了相近、相似或相等的概念内容[7]。

#### 3.1.2. 语义相似度

语义相似度是指两个词汇、短语或句子在意义层面上的接近程度[8]。在语义学与自然语言处理领域中，它通常用于衡量语言单位在上下文中的相似性[9]。在本文的跨模态任务中，如果一段文本和一张图像表达了相同或相近的语义。那么，它们在共享的空间中应具有较高的相似度。

#### 3.1.3. 多模态语义理解

文本与图像分别属于语言模态与视觉模态，它们在意义的表现方式上存在差异。表 1 总结了两种模态在语义表达维度上的主要差异。

**Table 1.** Analysis of differences in semantic representation between text and images
**表 1.** 文本与图像在语义表达上的差异分析

| 维度 | 文本 | 图像 |
|---|---|---|
| 表达方式 | 线性、离散、依赖语言规则 | 空间结构、连续变化、以视觉符号呈现 |
| 明确程度 | 借助词语可清晰定义对象与关系 | 需通过模型识别，存在理解模糊性 |
| 歧义处理 | 依赖上下文与语用信息澄清歧义 | 依赖主观判断，歧义难以规约 |

如果本文要将这两种模态建立连接，那么实验中必须要依赖一种共享的语义空间，使得模型能够将语言描述与图像内容映射到统一的高维语义向量表示中。

### 3.2. 统计学基础

#### 3.2.1. 向量空间模型

在本文中，文本和图像分别通过神经网络模型进行编码，并且构成了高维向量。这种做法是建立在向量空间模型的基础之上，将复杂的语义信息变为可计算的数值向量，使得语言与图像在统一的空间中

可以被计算比较[10]。

#### 3.2.2. 归一化处理

为了使来自不同模态的向量在共享语义空间中具备可比性，模型通常需要对其进行归一化处理。具体而言，本文对文本和图像的向量均采用了 L2 归一化，就是使得每个向量的模长调整为 1 [11]。在此处理下，向量仅保留其方向信息，从而确保后续余弦相似度与图像语义匹配得分计算。

#### 3.2.3. 余弦相似度

本文使用余弦相似度来度量文本与图像向量之间的接近程度[12]。其相似度计算示意图如下图 1 所示：

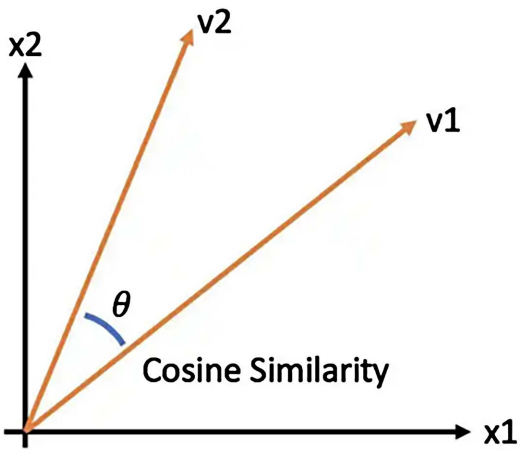


**Figure 1.** Cosine similarity calculation diagram
**图 1.** 余弦相似度计算示意图[①]

其公式如下：

$$Cosine\_similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{1}$$

#### 3.2.4. 图像语义匹配得分(CLIPScore)

本文的 MLP 模块输出的是一个匹配得分，这里本人将其命名为图像语义匹配得分(CLIPScore)。该得分的范围在 0 到 1 之间[13]。顺便一说，该得分不是直接用余弦相似度，而是通过文本和图像向量拼接，再加上它们的差值和乘积信息，经过一个两层神经网络来判别语义是否匹配。其公式如下所示：

$$s = \sigma\left(MLP\left(\left[\vec{t}, \vec{i}, \left|\vec{t} - \vec{i}\right|, \vec{t} \odot \vec{i}\right]\right)\right) \tag{2}$$

### 3.3. 神经网络模型基础

#### 3.3.1. CLIP 模型

CLIP 模型由图像编码器和文本编码器两部分组成[14]。图像编码器负责将图像转换为特征向量，可以是卷积神经网络或 Transformer 模型，如下图 2 所示；文本编码器则负责将文本转换为特征向量，通常是一个 Transformer 模型[15]，如下图 3 所示，这一模型最早由 Radford 等人提出，用于从大规模图文对中学习可迁移的视觉–语言表示[16]。这两个编码器会通过共享一个向量空间来实现跨模态的信息交互与融合[17]。

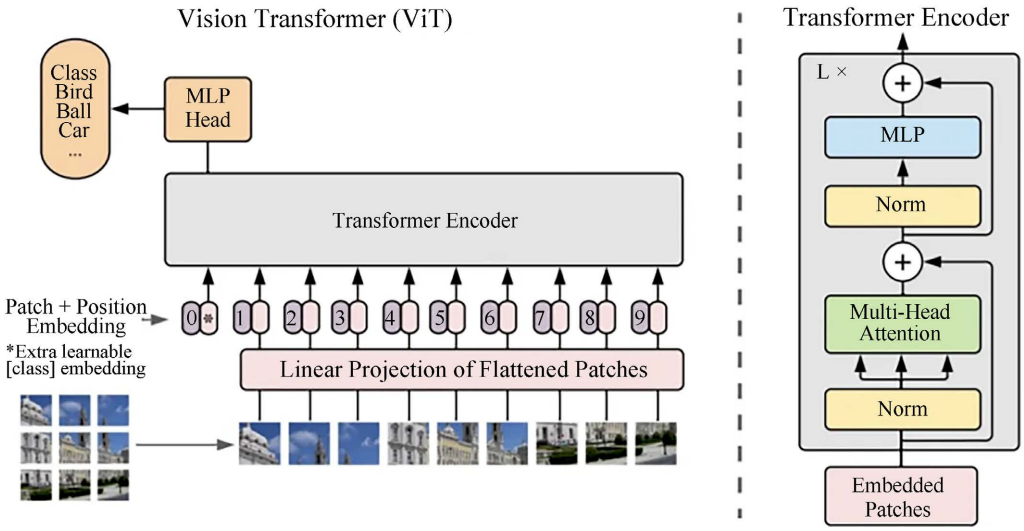


**Figure 2.** Architecture of Image Encoder in graphics editors
**图 2.** 图形编辑器 Image Encoder 架构[②]

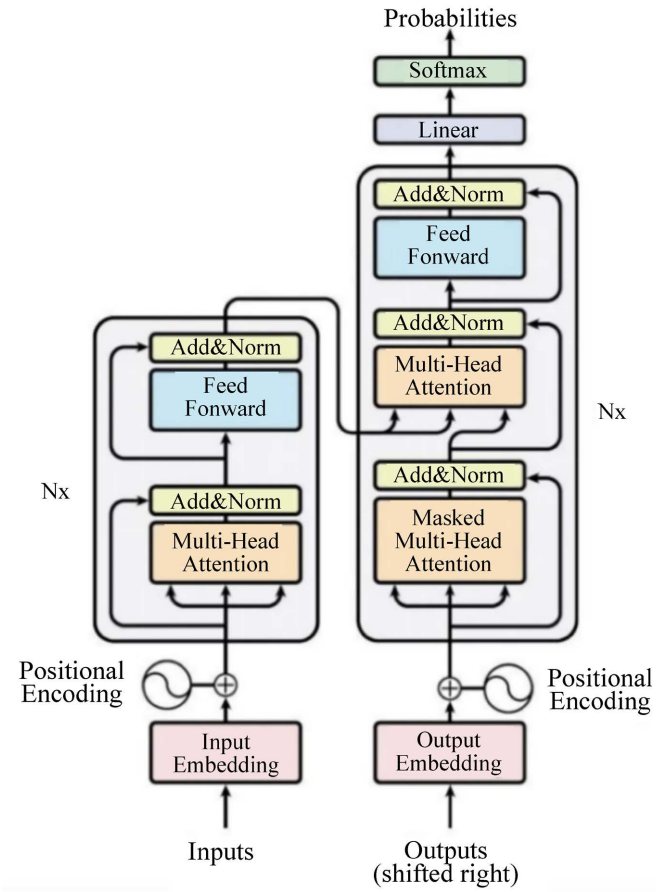


**Figure 3.** Architecture of Text Encoder in text editors
**图 3.** 文本编辑器 Text Encoder 架构[②]

### 3.3.2. Stable Diffusion 模型

Stable Diffusion 模型是一种特殊的扩散模型，称为潜在扩散模型[18]。近年来，基于扩散模型的图像

生成方法被广泛研究，并逐渐取代了 GAN 成为主流生成技术[19]。原始扩散模型往往会消耗更多的内存，因此创建了潜在扩散模型，它可以在称为潜在空间的低维空间中进行扩散过程[20]。模型结构如下图 4 所示：

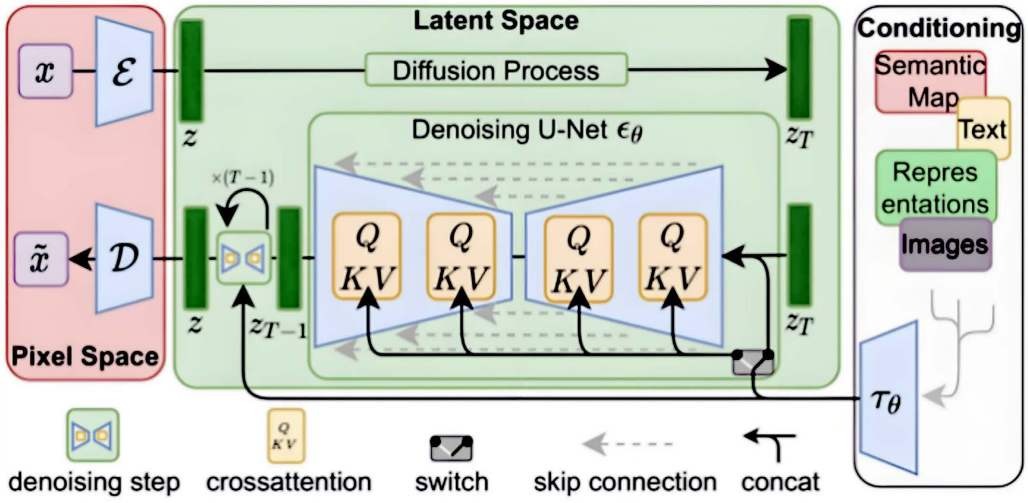


**Figure 4.** Stable Diffusion model architecture
**图 4.** Stable Diffusion 模型架构[③]

### 3.4. 评估指标

本文通过定量进行指标评估，采用的评估指标如下表 2 所示。

**Table 2.** Evaluation indicators
**表 2.** 评价指标

| 指标名称 | 说明 |
|---|---|
| Top-k 准确率 | 正确图像是否出现在前 k 名内 |
| 图像语义匹配得分 | 由 MLP 模块输出的图文匹配得分(0~1) |
| 余弦相似度 | 原始 CLIP 模型输出的语义相似度分值 |

## 4. CLIP + MLP 模型设计与构建

为了提升 CLIP 模型在语义匹配任务中的表现，本文在其基础上引入了一个轻量级的多层感知机(Multi-Layer Perceptron, MLP)判别模块，以增强模型对复杂语义关系的建模能力。具体来说，原始 CLIP 模型通过计算文本向量 $\vec{t}$ 与图像向量 $\vec{i}$ 之间的余弦相似度作为匹配得分，该方法虽然高效，但在处理复杂语义结构或多义表达时可能存在表达能力的瓶颈。为此，本文提出了一种语义组合特征结构：

$$x=\left[\vec{t};\vec{i};\left|\vec{t}-\vec{i}\right|;\vec{t}\odot\vec{i}\right] \tag{3}$$

通过拼接文本与图像向量本身、二者的绝对差以及 Hadamard 积来增强语义交互信息表达能力。该融合特征向量作为输入，进入一个由两层全连接神经网络组成的 MLP 判别器中，进行非线性匹配建模，最终输出一个匹配概率分数 $s\in[0,1]$。其形式化如下所示：

$$s=\sigma\left(W_2\cdot \mathrm{Re}LU\left(W_1\cdot\vec{x}+\vec{b_1}\right)+\vec{b_2}\right) \tag{4}$$

此次设计不仅保留了原始嵌入向量的信息结构，也引入了差异性与交互性特征，有效提升了模型在复杂语义匹配任务中的判别能力。

## 5. 文本到图像的语义检索实验

### 5.1. 基于 CLIP 模型的语义匹配能力分析

本文可以先从数据集随机选取一张图片进行文本描述，方便后期程序从数据集中找到这个图片，随机选取图片(图片编码：OIP-_7Dax4YmCM4SX9_EQn4qLQHaFZ)如下图 5 所示：

**Figure 5.** Randomly selected target image 1 (from the Animals-10 dataset)
**图 5.** 随机选取的目标图片 1 (Animals-10 数据集中)[④]

针对上图，本文构造了三个不同细节层级的文本描述，旨在从浅层语义到深层语义逐步测试模型的语义理解能力，文本描述如下所示：

**句 1:** 一匹马在吃草。

**句 2:** 一匹棕色的马正在低头吃干草，背景是一片田野和蓝天。

**句 3:** 一匹棕色的马低下头，在干枯的草地上吃草。它的鬃毛偏浅，阳光照射下呈现金黄色，前景是一片褐黄色的枯草，背景是模糊的栏杆和蓝天白云。

紧接着，本文可以对每句话一一进行测试，然后通过余弦相似度进行检测，结果如下表 3 所示。

**Table 3.** Different text, same image-cosine similarity
**表 3.** 不同文本同一图片－余弦相似度

| 文本编号 | 余弦相似度 |
|---|---|
| 句 1 | 0.86 |
| 句 2 | 0.90 |
| 句 3 | 0.97 |

由上表可以看出，随着文本描述语义层级的逐步加深，模型输出的余弦相似度得分也随之上升，分别为 0.86、0.90 与 0.97。这一趋势说明，CLIP 模型对文本中所包含的语义信息量与细节丰富度具有较高的响应能力。当文本的语义更加具体时，模型更容易从中提取有效语义特征，从而在图像向量空间中实

现更高程度的语义对齐。

此外，本文还验证了语义粒度对匹配性能的显著影响。比如说，句 1 虽能捕捉图像主体(“马”与“草”)的基本信息，但在空间场景与颜色特征等方面未形成完整的语义结构，导致匹配精度较低。而句 3 中引入了前景、背景、光照、颜色等多维度语义信息，使得模型在语义空间中更接近目标图像，匹配精度也是最高的。

## 5.2. CLIP + MLP 模型性能测试

为了进行新模型性能测试，本文再从数据集中随机选取一张图片进行文本描述，并进行性能测试。随机选取图片(图片编码：OIP-_4M8lLVlk06o0YOtolSlvQHaHL)如下图 6 所示：

**Figure 6.** Randomly selected target image 2 (from the Animals-10 dataset)
**图 6.** 随机选取的目标图片 2 (Animals-10 数据集中)[④]

针对上图，本文构造了五个不同细节层级的文本描述，旨在从深层语义到浅层语义逐步测试模型的语义理解能力，文本描述如下所示：

**句 1:** 一只黄色的小狗。

**句 2:** 一只黄色的幼犬，后面有一片绿色的草地和白色的围栏。

**句 3:** 一只黄色的幼犬被人抱着，背景是一片绿色的草地和白色的围栏。

**句 4:** 一只米黄色的卷毛小狗正被一只手抱起，它的耳朵垂下来，表情安静可爱。背景是一片整齐的绿草坪，后方有一道白色木质围栏横贯画面。

**句 5:** 画面中央是一只被人双手托举的小狗，它的毛色浅黄偏米黄，毛发柔软卷曲，眼神略带忧郁，嘴部饱满。手部戴有戒指，狗狗的耳朵自然下垂，右侧背景是一片修剪整齐的绿色草地，远处延伸着一道略有弯曲的白色围栏，整体画面色彩温和、自然清新。

紧接着，将上述五条文本描述分别输入至 CLIP 模型与 CLIP + MLP 模型进行语义匹配。其中 CLIP 模型输出对应的余弦相似度得分，而 CLIP + MLP 模型则输出图像语义匹配得分。

随后，将两组模型的匹配结果通过热力图形式进行可视化展示，以对比分析两种模型在面对不同语义细粒度描述时的表现差异。热力图结果如下图 7 所示。

通过热力图可以得知，CLIP + MLP 模型在全部 5 条描述中均取得了比原始 CLIP 模型更高的语义相似度得分，尤其是在详细描述的文本中表现更为突出。这说明 CLIP + MLP 模型在处理细粒度语义匹配任务时具备更强的判别能力和鲁棒性。

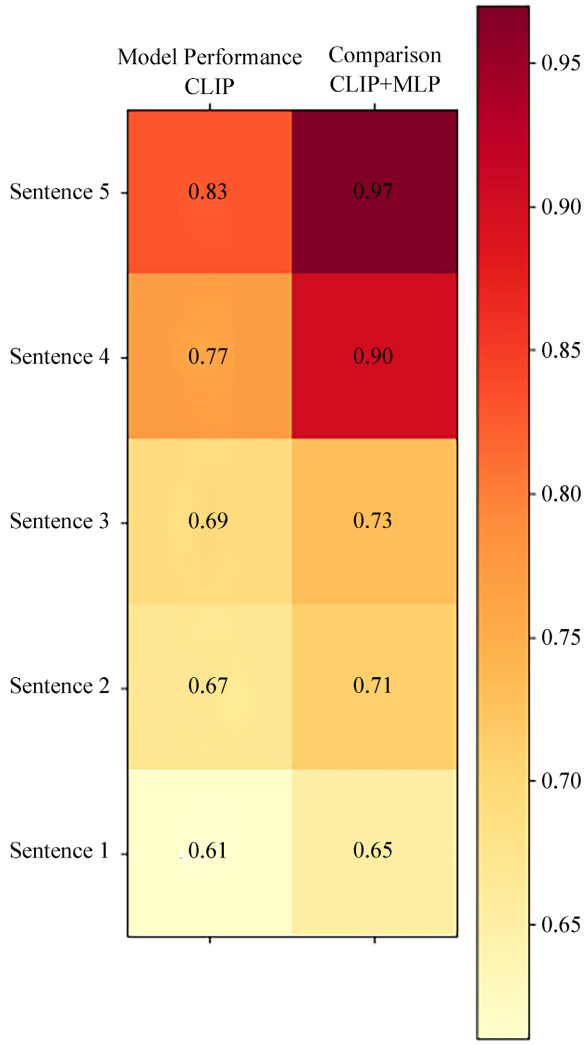


**Figure 7.** Model performance comparison-heatmap
**图 7.** 模型性能比较–热力图

## 6. 文本到图像的语义生成实验

在实验过程中，本文设置相似度阈值为 0.5，用以划分检索型与生成型任务的执行路径。当模型对文本描述与图像之间的相似度评分均低于该阈值时，即可判定数据集中不存在与该文本语义高度匹配的图像，此时任务将转入生成型流程。为深入探讨自然语言在视觉空间中的语义投射机制，生成型实验部分采用了 Stable Diffusion 1.5 模型，尝试通过文本驱动的方式生成与语义内容相对应的图像，从而分析模型对不同语义结构的建构能力。

### 6.1. 基础语义结构的可视化建构能力测试

与检索型方法依赖固定图像库进行匹配不同，生成型模型需要根据文本构建出全新的视觉场景。这一过程不仅涉及语义信息的抽取，还要求模型在内部建立语言与图像之间的意义映射规则。

为突出不同语义结构的表达差异，本文使用了以下三条具有代表性的文本描述：

**句 1:** 森林中的一座游泳池。

**句 2:** 来自非洲的男子在冰上奔跑。

**句 3:** 躺在摇篮里的小猫。

上述的这三句话分别体现出不同类型的语义构造。第一句话涉及空间嵌套语义，这便要求 Stable Diffusion 1.5 模型理解以森林作为场景背景与游泳池的局部包含关系；第二句话则构成一个语义冲突结构，将非洲的男子与冰上奔跑这两个不常见共现语义框架组合在一起，这对该模型的语义推理能力构成挑战；第三句话是典型的状态描写型描述，这便测试了该模型对于实体到位置的关系建模能力。生成图像结果如下图 8 所示：

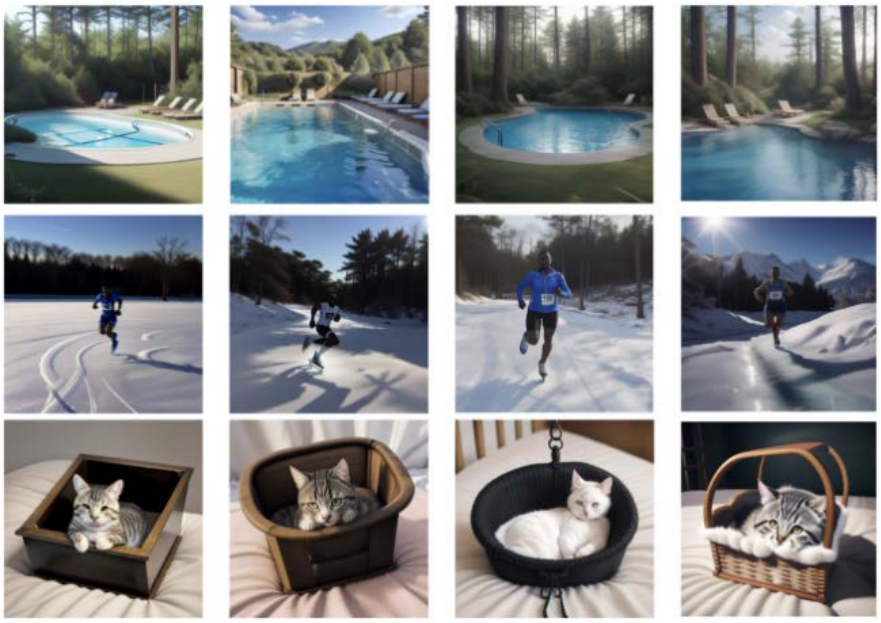

**Figure 8.** Stable Diffusion 1.5 model generates result 1
**图 8.** Stable Diffusion 1.5 模型生成结果 1

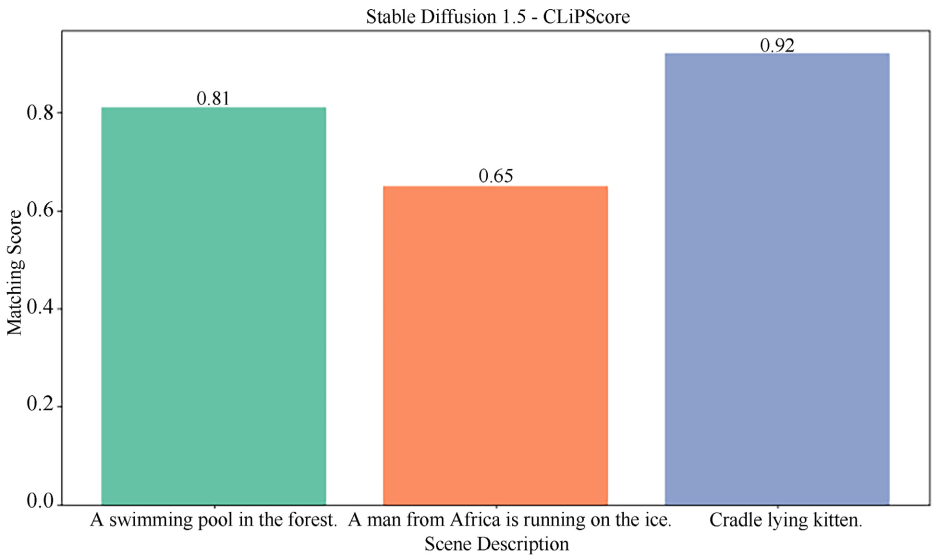


**Figure 9.** CLIPScore result 1
**图 9.** CLIPScore 结果 1

得到图 8 后，便可以对生成型实验中三组文本描述对应的图像语义匹配得分可视化图表，如图 9 所示。

从图 9 的可视化中可以得出，句 3 获得了最高分(0.92)，说明 Stable Diffusion 1.5 模型在静态实体与容器关系的表达上较为准确；句 1 的得分中等(0.81)，说明该模型基本能够理解场景组合语义，但在空间嵌套表达上仍可能存在偏差；句 2 的得分最低(0.65)，这便反映出该模型在处理语义冲突与非常规组合时表现有限。

## 6.2. 抽象风格与语义隐喻生成能力测试

在本部分实验中，为进一步评估模型对高阶语义结构的建构能力，特别引入包含语义抽象与语义隐喻特征的文本描述，具体示例如下：

**句 1:** 一只赛博朋克风格的猫→语义抽象实验

**句 2:** 锅中的鸭子→语义隐喻实验

本文首先设置了“赛博朋克(蒸汽朋克)风格”，这是一种较为抽象的语言。这种从抽象语言中提取语义然后映射到视觉风格，是语义抽象建构的体现。本句话生成图片如下图 10 所示。

**Figure 10.** Stable Diffusion 1.5 model generates result 2
**图 10.** Stable Diffusion 1.5 模型生成结果 2

**Figure 11.** Stable Diffusion 1.5 model generates result 3
**图 11.** Stable Diffusion 1.5 模型生成结果 3

赛博朋克风格作为一种抽象的风格词汇，既不明确指向某个具体颜色、动作或物体，也不构成严格的空间关系，而是通过语境加氛围的一种复合方式来构建意义。但是通过上图可以得知，Stable Diffusion 1.5 模型在生成图像中成功引入了“机械部件、金属结构、发光镜片”等高频赛博朋克视觉符号，表明其在语义抽象层级中具备一定的风格映射能力，能够将抽象修饰语转化为稳定的视觉风格元素。该过程便可视为 Stable Diffusion 1.5 模型在跨模态建模中对抽象语义到视觉风格的理解。

紧接着，可以对下一句话“锅中的鸭子”进行图片生成。本句的描述并不具有高度抽象的风格修饰语，而是包含了一个带有语义隐喻特征的日常生活场景。本句话从字面上看是一个具象的空间搭配，但在语义层面则有“被烹饪”文化隐含意义。生成图片如图 11 所示。

如图 11 所示，Stable Diffusion 1.5 模型对这句话表达的理解相对直接，图像中确实出现了一只鸭子放置于锅中。这说明模型能够捕捉句中实体之间的嵌套关系，但是不能成功实现被烹饪的语义隐喻联想。

对应的图像语义匹配得分的可视化图表如下图 12 所示，这些得分是针对这两个文本描述的。

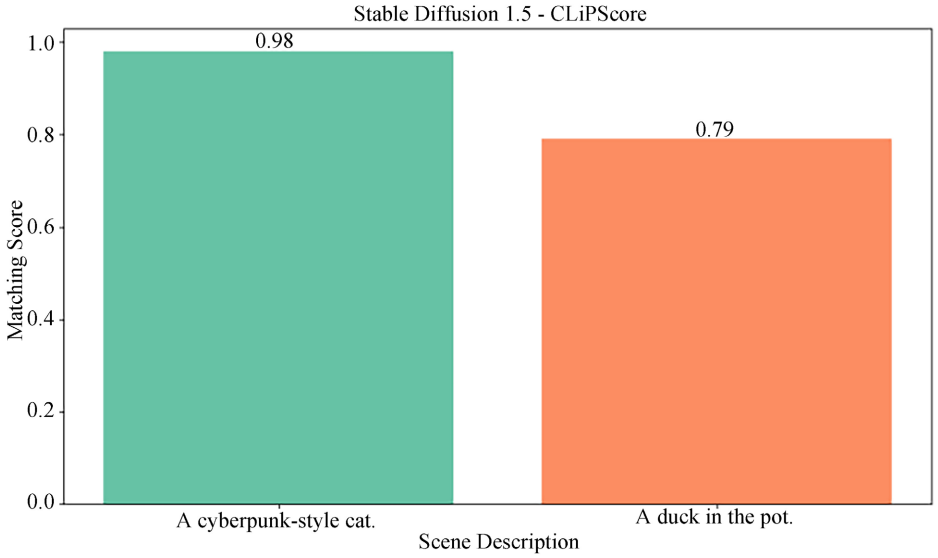


**Figure 12.** CLIPScore result 1
**图 12.** CLIPScore 结果 1

如图 12 结果显示，句 1 的图像语义匹配得分最高，得分为 0.98，这反映 Stable Diffusion 1.5 模型在抽象风格词映射上较为优秀。而句 2 的得分为 0.79，这表明了 Stable Diffusion 1.5 模型对空间嵌套关系虽能部分还原，但在处理文化隐喻存在一定的语义偏移。

## 7. 结语

本文聚焦于“文搜图”这一跨模态语义理解任务，基于 CLIP 模型提出了融合多层感知机(MLP)的改进结构 CLIP + MLP 模型，并通过文本到图像的匹配实验和文本驱动图像生成实验，系统评估了其在不同语义粒度与结构下的建模能力。

在检索型实验中，结果表明原始 CLIP 模型对浅层语义具备基本对齐能力，但在细粒度匹配任务中存在表达能力瓶颈。通过引入 MLP 非线性判别模块，CLIP + MLP 模型能够更有效捕捉语义组合特征，显著提升了模型在复杂语义结构下的匹配精度和鲁棒性。

在生成型实验中，Stable Diffusion 1.5 模型展现了在静态实体描述和抽象风格映射方面的良好性能，能较为准确地将语言描述转换为视觉图像。但在应对语义冲突、空间嵌套及文化隐喻等非常规语义表达

时，仍存在生成结果语义偏移或构型模糊的问题。

综上，本文验证了在跨模态场景下构建更强语义理解能力模型的重要性，同时也揭示了当前主流模型在面对深层语言结构时的局限性。

## 注　释

①图片来源：https://baijiahao.baidu.com/s?id=1821817320449768081&wfr=spider&for=pc (引用日期：2025-06-08)

②图片来源：https://mp.weixin.qq.com/s?__biz=MzIwNDY0MjYzOA==&mid=2247514749&idx=1&sn=bfa35fd34561201469e9fcafd6a45a4f&chksm=968f621c76c6bbfcf94570a18f5f14511e6c8f50d39a20729660d07427a218675d6dd118028e&scene=27 (引用日期：2025-06-08)

③图片来源：https://zhuanlan.zhihu.com/p/20329891447 (引用日期：2025-06-08)

④图片来源：https://www.kaggle.com/datasets/alessiocorrado99/animals10 (引用日期：2025-06-08)

## 基金项目

2023 年大学生创新创业训练计划项目(202313637013)。

## 参考文献

[1] 彭晏飞, 孙鲁. 基于图像分割的语义标注方法[J]. 计算机应用, 2012, 32(6): 1548-1551.

[2] Baltrušaitis, T., Ahuja, C. and Morency, L.P. (2018) Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

[3] 彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望[J]. 计算机研究与发展, 2019, 56(1): 183-208.

[4] Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z. and Shen, H.T. (2025) Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *Proceedings of the IEEE*, **112**, 1716-1754.

[5] 张玉康, 谭磊, 陈靓影. 基于图像和特征联合约束的跨模态行人重识别[J]. 自动化学报, 2021, 47(8): 1943-1950.

[6] 林倞, 杨斌斌. 从感知到创造: 图像视频生成式方法前沿探讨[J]. 光学学报, 2023, 43(15): 155-175.

[7] 张雷, 崔荣一. 基于编辑距离的词序敏感相似度度量方法[J]. 延边大学学报: 自然科学版, 2020, 46(2): 140-144.

[8] 李秋明, 张卫山, 张培颖. 基于句子多种特征的相似度计算模型[J]. 软件导刊, 2016, 15(9): 4-6.

[9] 徐健. 基于多种测度的术语相似度集成计算研究[J]. 情报学报, 2013, 32(6): 618-628.

[10] 张晅. 在跨模态检索技术加持下推动广电数据安全[J]. 影视制作, 2024, 30(7): 78-81.

[11] 赵琼. 基于视频和三维动作捕捉数据的人体动作识别方法的研究[D]: [博士学位论文]. 中国科学技术大学, 2025.

[12] 刘鑫. 内容过滤技术与挖掘算法的设计优化[J]. 2024, 53(5): 42-43.

[13] 付泽润. 基于子兴趣分解的神经协同过滤方法[D]: [硕士学位论文]. 太原: 太原理工大学, 2022.

[14] 刘颖, 郭莹莹, 房杰, 等. 深度学习跨模态图文检索研究综述[J]. 计算机科学与探索, 2022, 16(3): 489-511.

[15] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., *et al.* (2021) Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. *International Conference on Machine Learning*, Online, 18 July 2021, 4904-4916.

[16] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., *et al.* (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, Online, 18 July 2021, 8748-8763.

[17] 郑玉栋. 基于外部注意力机制的多模态模型研究[D]: [硕士学位论文]. 哈尔滨: 黑龙江大学, 2023

[18] 毛琪, 方镇, 陈澜, 等. 基于扩散模型的图像编辑研究现状[J]. 中国传媒大学学报(自然科学版), 2024, 31(4): 38-54.

[19] Li, J., Li, D., Xiong, C. and Hoi, S. (2022) Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. *International Conference on Machine Learning*, Online, 18 July 2021, 12888-12900.

[20] 李帅帅, 何向真, 张跃洲, 等. 融合多情感的语音驱动虚拟说话人生成方法[J]. 计算机应用研究, 2024, 41(8): 2546-2553.