

基于数据合成的多模态讽刺隐喻理解大模型的构建

戴凌睿^{1,2}, 李浩^{1,2}, 吴云芳^{2,3*}

¹北京大学软件与微电子学院/ 北京市, 102600

²北京大学多媒体信息处理国家重点实验室/ 北京市, 100871

³北京大学计算机学院/ 北京市, 100871

2201210183@stu.pku.edu.cn, {fst1,wuyf}@pku.edu.cn

摘要

讽刺和隐喻是文学与语言表达中常见的修辞手法, 以往相关研究多聚焦于分类任务上, 且更多的基于英文数据进行探索。随着大模型与多模态大模型的不断涌现, 模型对各种自然语言处理任务与多模态任务的处理能力得到了显著的提高。本文利用GPT-4o进行自动数据合成, 来训练多模态大模型, 实现了图文多模态讽刺隐喻综合理解任务。本文训练出能理解图片或图文讽刺隐喻内容, 并进行详细解释或配文的参数量较小的多模态大模型, 并且保证了模型具备良好的鲁棒性和通用性能。本文精心设计了数据构造方法, 包括数据源的选择, 指令数据的合成, 回复数据的合成, 来获得了一批高质量的多模态讽刺隐喻指令微调数据。我们选用了当前表现较好的多模态大模型作为骨干模型, 使用合成数据并结合公开多模态图文数据集进行训练。在模型评测方面, 本文分别从讽刺隐喻理解能力和通用能力进行评测, 验证了模型的可用性。本文的数据以及模型权重将在后续放置在<https://github.com/652897698/Multimodal-LLMs-for-Sarcasm-and-Metaphor-Undrerstanding>

关键词: 讽刺识别; 隐喻识别; 多模态大模型; 数据合成

Fine-tuning Multimodal LLMs for Sarcasm and Metaphor Undrerstanding Based on Data Synthesis

Lingrui Dai^{1,2}, Hao Li^{1,2}, Yunfang Wu^{2,3*}

¹School of Software and Microelectronics, Peking University, Beijing, China

²National Key Laboratory for Multimedia Information Processing, Peking University

³School of Computer Science, Peking University, Beijing, China

2201210183@stu.pku.edu.cn, {fst1,wuyf}@pku.edu.cn

Abstract

Sarcasm and metaphor are widely used rhetorical devices in literature and language. Previous research has primarily focused on classification tasks and has largely concentrated on English datasets. With the emergence of large language models and multimodal large language models, there has been a significant improvement in their ability to handle a variety of natural language processing and multimodal tasks. In this paper, we utilize GPT-4o to automatically synthesize data for training a multimodal large language model, aiming to enable multimodal understanding of sarcasm and metaphor across text and images. We aim to train a multimodal LLM capable of providing detailed explanations and generating descriptive captions, with an emphasis on achieving strong robustness and generalization capabilities. To this end, we carefully design a data construction pipeline which includes source selection, synthetic instruction generation, and response generation, to produce a high-quality multimodal

instruction-tuning dataset focused on sarcasm and metaphor. Using multimodal models with strong performance as our base model, we train them with both synthetic and publicly available datasets. For evaluation, we assess the model's comprehension of sarcasm and metaphor alongside its general capabilities, demonstrating its effectiveness. The data and model weights of this paper will be placed at <https://github.com/652897698/Multimodal-LLMs-for-Sarcasm-and-Metaphor-Understanding>.

Keywords: Sarcasm Recognition , Metaphor Recognition , Multimodal LLM , Data Synthesis

1 引言

讽刺和隐喻是文学与语言表达中常见的修辞手法，讽刺通过说反话或夸张进行批评 (Dews and Winner, 1995)，隐喻则用一种事物暗喻另一种事物 (Lagerwerf and Meijers, 2008)，二者都具有含蓄性。过去关于隐喻和讽刺的研究大多集中在分类任务上。隐喻相关的任务主要解决隐喻词的判断，或一句话是否包含隐喻信息。讽刺相关任务则主要判断这句话或一个图文对，一个视频中是否包含讽刺含义。在隐喻的研究中，大多还是以英文数据集为基础，中文相关的数据则相对偏少。同时，对于讽刺和隐喻现象进行解释的工作也较少。

随着大模型的不断发展，包括纯文本大模型 (Brown et al., 2020; Touvron et al., 2023)，以及多模态大模型 (Liu et al., 2023; Dai et al., 2023)，使得模型在自然语言处理、计算机视觉、多模态等领域任务中展现出卓越的泛化能力与语义理解深度。在此背景下，人工智能研究范式也发生深刻转变，任务重心从模式识别为主的分类任务，逐步向更具挑战性的生成任务迁移。因此本文尝试探索在讽刺和隐喻理解这一领域，利用大模型来完成的效果，并且从传统的讽刺隐喻的分类任务上转向生成任务，即对讽刺和隐喻现象的解释，或者是创作讽刺和隐喻的内容。

本文利用GPT-4o (Hurst et al., 2024)生成合成数据，对参数量较小的多模态大模型进行微调，训练出在讽刺与隐喻理解领域上更优秀的多模态大模型，同时尽可能地保留了模型的通用能力。本文设计了多种数据合成的方法，来保证合成数据的多样性与复杂性。本文选择公开的多模态讽刺分类数据集MMSD2.0 (Qin et al., 2023)以及自行爬取的小红书数据作为本文的数据源来进行数据合成。其中的合成数据包括指令数据的合成与回复数据的合成。在指令数据的合成中，我们将其划分为三类不同的合成方法，分别为：对种子指令集合进行改写与衍生；对数据生成讽刺隐喻相关的随机指令；对数据生成任意的通用随机指令。在回复数据的合成中，本文对带有分类标签的MMSD2.0数据进行额外处理，来保证回复数据的准确性。

本文使用上述的合成数据，并混合公开的多模态通用指令微调数据，通过SFT训练讽刺隐喻多模态大模型。本文通过讽刺隐喻理解评测和通用能力评测来综合评判模型的能力。在讽刺隐喻理解评测中，我们进行了在MMSD2.0数据上的判别测试以及使用大模型出题并评判的讽刺隐喻的综合理解能力测试。在通用能力评测中，本文使用公开多模态评测集MMStar (Chen et al., 2024a)以及MMBench (Liu et al., 2024b)的分数作为对模型的通用能力评估。综合来说，本文的贡献如下：

- 在讽刺隐喻理解领域上设计并实现了多种利用多模态大模型自动合成数据的方法，并构建了讽刺隐喻综合理解指令微调数据集。
- 利用构建的讽刺隐喻综合理解指令微调数据集对参数量较小的多模态大模型进行训练，得到讽刺隐喻领域多模态大模型。并对模型进行了讽刺隐喻能力与通用能力上的综合评测，验证了模型的可用性。

*通讯作者Corresponding author

©2025 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

2 相关工作

2.1 讽刺识别与隐喻识别

在过去关于讽刺与隐喻的研究中，模型主要基于Transformer (Vaswani et al., 2017)架构，并在其基础上进行模型的修改，特征的添加，损失函数的设计等。如果是多模态上，则会额外设计不同模态之间的交互。

在讽刺识别的研究中，主要包括模型的设计和讽刺数据集的构建。Cai(2019)等人选取文本特征、图像特征以及图像属性作为三种模态，进而提出了一种多模态层次融合模型，并同时构建了多模态图文讽刺分类数据集MMSD。他们在社交平台twitter上通过带有讽刺tag，如#sarcasm来爬取数据，来构建图文讽刺分类数据集。Qin (2023)等人在Cai的基础上构建了MMSD2.0多模态图文讽刺识别数据集。他们发现了MMSD存在的一些虚假线索，并重新标注，并进行一致性和质量检测，得到最终数据。Wu (Wu et al., 2024b)等人设计了一种统一视觉模态和语言模态的多模态软提示模型，来解决少样本设置下的讽刺检测。Zhang (2024)等人采用文本编码器和图像编码器分别对文本和图像进行编码，并以CLIP (Radford et al., 2021)作为教师模型，通过对比学习的方式实现文本与图像特征的语义对齐。

在隐喻识别中，传统的隐喻识别采用基于特征的方法。其中手工设计的特征包括词语的抽象性和具体性 (Turney et al., 2011)。后续随着深度学习的发展，在隐喻识别任务上通常采用预训练模型与语言学性质相结合的方法，Minjin Choi (2021)等人同样以RoBERTa作为骨干网络，并基于MIP(Metaphor Identification Procedure) (Group, 2007)和SPV(Selectional Preference Violation) (Wilks, 1975; Wilks, 1978)的理论设计模型，提高模型在隐喻检测上的性能。在数据构建上，Mohammad等人发布的数据集MOH (Mohammad et al., 2016)从WordNet中提取的1639个句子（1230个字面义句子和409个隐喻性句子）组成，体现了440个动词的用法。动词用法的隐喻性标注是通过众包方式获得的。VUA数据由阿姆斯特丹自由大学（VU Amsterdam）隐喻语料库在2018年和2020年的隐喻检测共享任务中发布，分别称为VUA-18 (Leong et al., 2018)和VUA-20 (Leong et al., 2020)。性。

2.2 多模态大模型

大语言模型在自然语言处理任务上展现出令人惊叹的能力，基于大语言模型为骨干网络，涌现出越来越多强大的多模态大模型。GPT-4 (Achiam et al., 2023)作为早期大语言模型的代表，虽以文本处理为主，但其强大的语言理解和生成能力为后续多模态扩展奠定了基础。后续版本如GPT-4V进一步整合视觉能力，支持图像理解和文本生成，例如生成视频字幕或分析医疗影像，但是该模型为闭源模型。

早期的开源多模态大模型以LLaVA (Liu et al., 2023)和InstructBLIP (Dai et al., 2023)为代表。LLaVa对于视觉信息的处理基于视觉编码器和映射网络，大语言模型基座则使用Llama-2 (Touvron et al., 2023)进行初始化，这类结构成为了近来多模态大模型的主流。InstructBLIP对视觉信息的处理则基于Q-Former板块，这种架构继承自BLIP-2模型 (Li et al., 2023a)。在后续的发展中，类似的多模态大模型大量涌现。InternVL (Chen et al., 2024b)系列模型通过扩大视觉处理器规模，为模型性能提升奠定了硬件基础。在训练策略方面，采用渐进式对齐训练策略，先在大规模嘈杂数据上开展对比学习，让模型初步掌握视觉与语言的关联；随后在高质量数据上进行生成学习，促使模型能够精准适配多种任务需求。在InternVL1.5以及InternVL2版本中，模型展现出强大的图像处理能力。DeepSeek系列中，DeepSeek-VL (Lu et al., 2024)采用了一种混合视觉编码器，能够在固定的标记预算内高效处理高分辨率图像，同时保持相对较低的计算开销。DeepSeek-VL2 (Wu et al., 2024a)采用MoE架构，在语言模型部分引入专家并行，动态选择不同的专家网络处理不同任务，显著提升训练效率并降低计算成本。DeepSeek-VL2在一系列常用的基准评测集上进行了评估，与其他多模态大模型相比表现优异。

2.3 使用大模型合成数据

对于缺少数据集的任务上来说，使用人工进行标注是十分昂贵的，在大模型出现后，为实现数据标注和数据合成这一复杂过程的自动化提供了前所未有的机遇。Llava使用图片的描述或图片的bounding box信息输入GPT-4，让GPT-4自动合成SFT数据。Llama3.1 (Grattafiori et al., 2024)中会使用自己生成的数据来训练自己。具体来说，会让当前迭代轮次的Llama模型根

据收集的prompt生成k个回答，然后选择RM模型打分最高的模型的回答，作为下一轮SFT数据的一部分。Tan (2024)等人对使用大模型进行数据标注进行了详细的调查。他们指出指令的多样性 (Li et al., 2023b; Song et al., 2024) 对大模型的学习十分重要，最近的研究探索了各种方法来使原始数据集中的指令多样化并进行扩充。Wang (2022)等人使用一些人工编写的种子指令，并通过“生成后过滤”的流程对其进行迭代扩充，来提高指令的多样性。或者专门训练一个指令生成模型 (Meng et al., 2023; Wang et al., 2024) 来生成指令，以增加指令的多样性。在完成数据标注与数据合成后，最常用的应用就是用来微调一个较小的模型来提高该模型在特定领域的能力，即用大模型的能力蒸馏到小模型上。许多中小型大语言模型都在由较大的大语言模型生成的合成数据集上进行训练 (Xu et al., 2023)。即使不采用大模型蒸馏到小模型的策略，用模型生产的数据再训练自己进行自蒸馏有时也是有效的做法 (Yang et al., 2024)。

3 指令微调数据合成

本文主要利用GPT-4o自动合成图文多模态指令微调数据，来训练参数量更小的多模态大模型，使其更好地适应讽刺和隐喻相关的任务，并保持模型的通用能力。整体的实验流程如图 1所示。本文设计了三种数据合成的方式来合成讽刺和隐喻理解相关的数据。我们将合成的所有数据以及公开的多模态通用指令微调数据集进行混合，进行模型的训练。本节主要介绍指

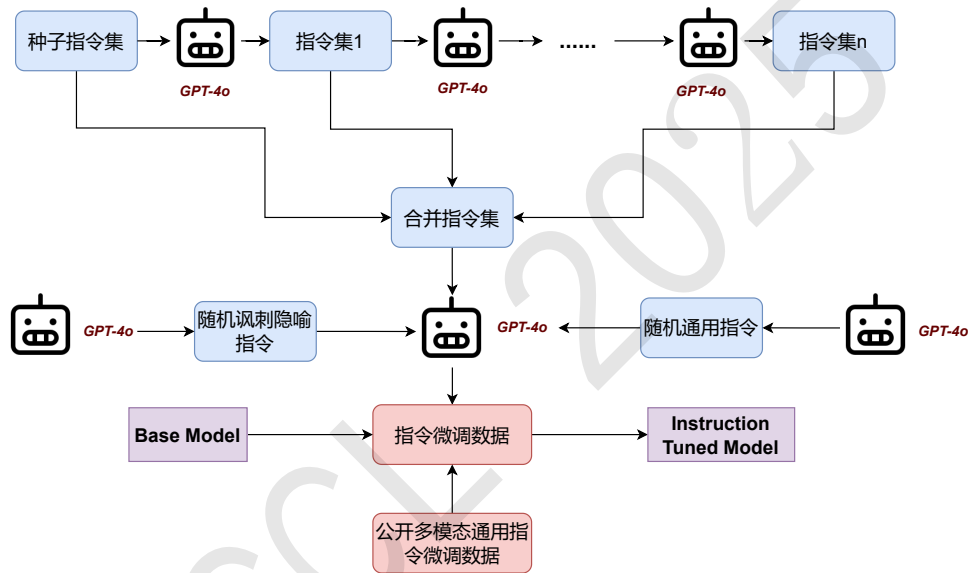


Figure 1: 整体实验流程图

令微调数据生成的方法，包括数据源的选择，指令数据的合成以及回复数据的合成。

3.1 数据源

为了进行数据合成，需要收集图文数据以及图片数据。我们希望收集的图片或图文包含隐喻信息或讽刺信息的概率尽可能大，因此本章使用两个数据源，一个是公开的多模态讽刺分类数据集MMSD2.0，一个是通过一些固定tag进行爬取的小红书数据。

MMSD2.0是对多模态讽刺数分类数据集MMSD的修正版。MMSD通过在推特爬取带有#sarcasm标签的图文数据，形成多模态讽刺数据集。MMSD2.0去除了话题标签词以及表情符号词，来防止模型只需学习这种虚假的相关性就能做出正确预测。然后他们通过众包对不合理样本重新标注，并进行了一致性检测和质量检测，得到最终的数据。该数据的统计信息如表1。该数据拥有配对的图文对以及相应的讽刺判别的标签，在实际合成时使用图文对或单独的图片进行数据合成。

小红书是一款具有广泛影响力的社交平台，用户可以在平台上发布图片，视频，推文等内容，相比其他社交平台，小红书的图文数据更多且质量更高，因此考虑在小红书上爬取可能与

MMSD2.0	Train	Validation	Test
Sentences	19816	2410	2409
Positive	9572	1042	1037
Negative	10240	1368	1372

Table 1: MMSD2.0数据统计

讽刺隐喻相关的推文。具体来说，我们会使用一些tag来进行推文爬取，包括“讽刺图片”，“隐喻图片”，“创意广告”，“引人深思的图片”，“创意内涵作品”，“寓意深刻的图片”。在实际使用时，只使用爬取的图片来进行数据合成。其中一条推文的图片示例如图2。



Figure 2: 小红书爬取的图片示例

3.2 指令数据合成

为了保证最终训练的模型对用户不同的指令输入具有鲁棒性，以及保持模型的通用能力，我们需要保证指令的复杂性以及多样性。因此，本实验在合成指令时划分为三大类别。

第一大类为讽刺隐喻固定指令的改写以及衍生，用于保证模型对讽刺隐喻的基本能力。首先，会由人工编写几个固定的指令模板作为种子指令集。然后使用GPT-4o对指令集进行改写以及衍生，生成新的指令。再将新生成的指令作为下一轮的指令模板。不断迭代该过程，最后将所有的指令进行合并，获得最终的指令集。具体的指令模板内容可以参考附录A。合成的流程如图1的上半部分所示。本文初始指令分为划分为两种。第一类为解释类指令，即给出图文或图片，要求回复给出对应的讽刺或隐喻的解释。第二类为创作类指令，即给出图片，创作出可能的文本，使图文对形成讽刺或隐喻，例如为某张图片创作一条具有讽刺或隐喻性质的标语。在使用GPT-4o进行改写以及衍生时，我们使用如图3的prompt，其中中文和英文分别设计。

第二大类为与讽刺隐喻相关的随机指令，来更大程度上丰富与讽刺隐喻相关指令的多样性与复杂性。具体来说，会直接要求GPT-4o根据一张图片或者一对图文，自行从零生成与讽刺隐喻相关的可能的指令，而不是像第一大类一样依据种子指令进行改写和衍生。并且在生成指令时会同时输入图片，因此生成的该指令只能作用于该图片或该图文，而不像第一大类生成的指令可以作用于任何图片或图文。在使用图文数据进行合成时，prompt如图8所示，在使用图片数据进行合成时，prompt如图9所示。

第三大类为随机通用指令，来保证模型对需要合成的多模态数据的数据分布上的通用能力。具体来说，会直接要求GPT-4o生成一些图片相关的通用指令。整个生成过程与第二大类类似，在此不再赘述。使用的prompt如图10所示。

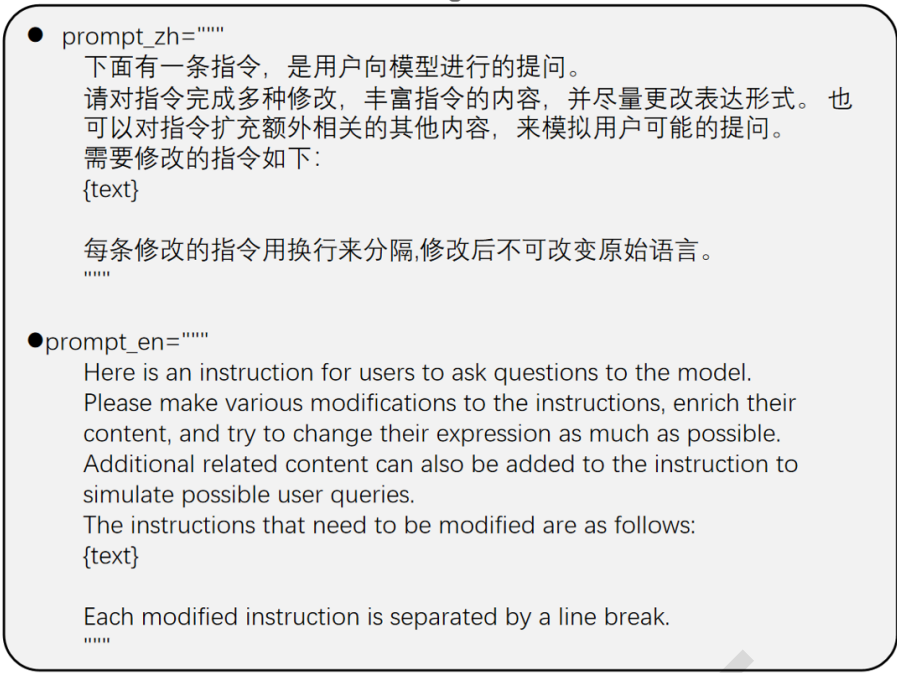


Figure 3: 输入GPT-4o来帮助改写指令的prompt

3.3 回复数据合成

在生成回复数据时，依据指令数据的三大类依次生成回复数据。

在第一大类中，对于讽刺隐喻种子指令的改写以及衍生，我们已经生成了指令集。在生成回复时，对于每一个图片或图文数据，以相同概率在指令集中随机抽取一条指令，将其作为GPT-4o的prompt，让GPT-4o给出响应，并将GPT-4o的回复作为回复数据。由于MMSD2.0数据给出了多模态讽刺的分类标签，因此在让GPT-4o标注MMSD2.0的讽刺解释数据时，需要将该分类标签告诉GPT-4o。具体来说，本实验会在指令的最后依据数据的标签以及语言插入下面的内容：讽刺数据插入(It is known that this picture-text pair contains sarcasm.)或(已知这个图文对包含讽刺表达)，非讽刺数据插入(It is known that this picture-text pair does not contain sarcasm.)或(已知这个图文对不包含讽刺表达)。而在训练自己的模型的时候，再将这部分插入的内容删掉。在标注MMSD2.0的讽刺文本创作数据时，以50%的概率将原始的推文作为示例输入给GPT-4o，此时在prompt中插入 (Example text that could constitute sarcasm:’+’+text+’ But remember that the generated text is as different as possible from the example)，在训练自己的模型的时候，再将这部分插入的内容删掉。

第二大类和第三大类的回复生成比较类似，在使用GPT-4o进行合成时，指令和回复的生成由一步完成，即GPT-4o生成的内容包含了需要训练模型的指令和回复。第二大类生成的指令与第一大类的区别在于在生成的指令中会针对性的插入与本张图片相关的内容，来满足用户输入的多样性。并且在标注过程中，通过prompt基本保持了中文和英文的数据比例，来使模型支持双语能力。示例如图4所示。最终获得如表2的标注训练数据，共计9400条。

指令与回复的类型	MMSD2.0	小红书
第一大类（讽刺隐喻改写类）	3800	3000
第二大类（随机讽刺隐喻类）	500	800
第三大类（随机通用类）	500	800

Table 2: 合成数据统计

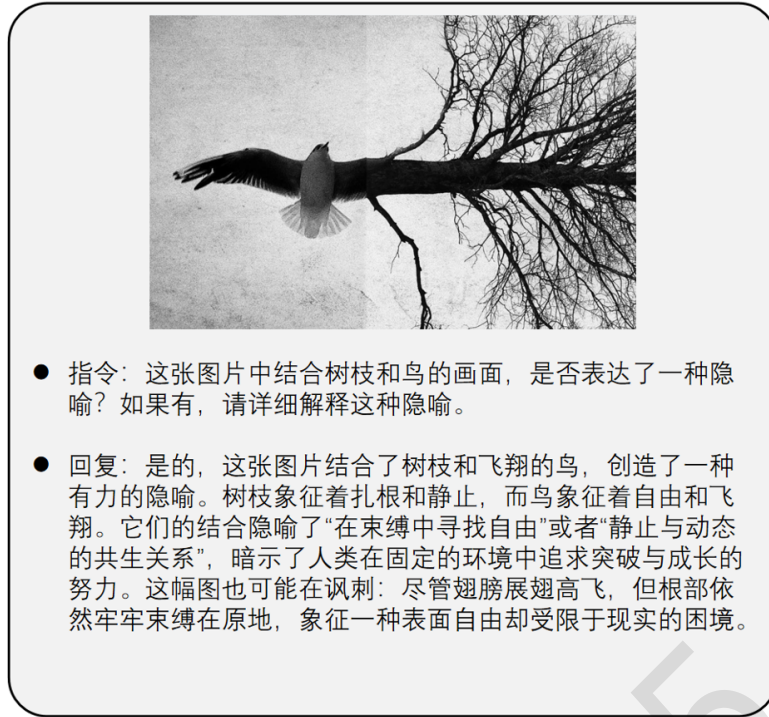


Figure 4: 任务相关的随机指令与回复示例

4 实验设置

4.1 模型训练

我们选择InternVL2与DeepSeek-VL作为骨干模型。具体在实验中，选择InternVL2-8B与deepseek-vl-7b-chat作为实际训练的模型。

除了使用上一节的9400条合成数据外，还额外添加了公开的多模态图文数据集，来进一步维持模型的通用能力表现。具体来说，我们参考了Llava1.5中SFT的数据设置 (Liu et al., 2024a)，采用该设置的子集作为补充通用数据。在上述通用数据中进行随机采样2000条作为通用数据的补充样本。因此总训练数据为3.3小节提到的9400条数据与2000条通用多模态指令微调数据，合并为11400条数据作为最终的SFT数据。

本实验使用LORA (Hu et al., 2022)对InternVL2-8B与deepseek-vl-7b-chat进行SFT，lora rank设置为8。学习率为 $1e-4$ ，采用余弦衰减，共训练2个epoch。

4.2 模型评测

模型评测分为讽刺隐喻理解能力的评测以及通用能力的评测。

在讽刺隐喻理解能力的评测上，可以分为判别任务和生成任务进行评测。

由于MMSD2.0数据带有讽刺分类的标签，可以直接在该测试集上评测讽刺判别准确率，因此在该数据的测试集上进行测试。

对于生成任务，包括隐喻讽刺的解释以及隐喻讽刺文本的创作，我们使用大模型来进行自动评测，采用两两比较的方法来对模型进行测试。即收集两个模型的答案，让大模型来判断谁的答案更好。具体来说在MMSD2.0的测试集上和小红书未使用的数据上分别抽取100条数据来作为讽刺隐喻综合能力测试的评测集。由于从讽刺隐喻指令库中收集指令来对模型进行评测可能影响公平性，因为经过训练的模型可能见过这些指令，尽管使用的数据是未见过的。因此使用GPT-4o对每条测试数据随机设计一个与讽刺隐喻可能相关的指令，再将设计的指令与测试数据输入需要对比的两个模型，获得两个模型的回答，最后再使用大模型对这两条数据进行对比，判别哪个答案更好，最终使用判别结果的胜率作为讽刺隐喻综合能力的评价标准。我们使用GPT-4o作为裁判，使用如图5中的prompt进行评测。

```

● prompt = ""
    你现在作为一个公正的裁判，请对输入的图片以及对应的
    指令，判断两个输出的答案哪个更好，并给出理由
    指令: {insturction}
    答案1: {out1}
    答案2: {out2}

    请严格按照下面的格式输出：
    答案1/答案2
    理由: xxx
    ""

```

Figure 5: 使用GPT-4o作为裁判的prompt

在通用能力的评估上，我们使用公开多模态评测集MMStar以及MMBench作为模型的评测集来评估模型的通用能力

5 实验结果与分析

5.1 讽刺隐喻理解评测

首先在MMSD2.0测试集上对模型进行测试。由于测试对象为大模型，无法直接拿到分类标签，因此通过prompt来提示大模型生成答案，并通过大模型的结构化输出来得到分类的结果。基线模型参考提出MMSD2.0中的部分实验结果。其中InternVL2-8B-ms和deepseek-vl-7b-chat-ms为使用4.1小节的全量数据进行训练得到的模型。InternVL2-8B与deepseek-vl-7b-chat为原始版本的模型。HFM来自Cai等人 (2019)的论文，Multiview CLIP来自Qin等人 (2023)的工作。实验结果如表3所示。其中小模型的结果是使用了完整的训练集的结果，大模型则分别对比了零样本实验和2000条MMSD2.0训练数据的结果。

	Model	Acc	F1	Precision	Recall
小模型-All TrainData	HFM	0.7057	0.6688	0.6484	0.6905
	Multiview CLIP	0.8564	0.841	0.8033	0.8824
大模型-零样本	InternVL2-8B	0.5512	0.6438	0.4889	0.9421
	deepseek-vl-7b-chat	0.4806	0.6159	0.4518	0.9672
大模型-2000 TrainData	InternVL2-8B-ms	0.7882	0.7694	0.7242	0.8206
	deepseek-vl-7b-chat-ms	0.7447	0.7535	0.6447	0.9064

Table 3: MMSD2.0测试结果

原始的MMSD2.0数据中训练集包含19816条数据，在实际的训练数据中，使用了4800条训练数据，而来自MMSD2.0的数据共2000条（在构造数据时进行了类别的划分，挑选了2000条只与讽刺判别和解释相关的指令），且这2000条中也同样包含没有分类指令的任务在其中，并且指令形式各不相同。因此在该测试集上并不会追求高指标，仅用来观测讽刺判别的效果。

通过观察实验结果，未在MMSD2.0上训练的大模型表现出较低的结果，模型召回率高，准确率低，说明这两类大模型更容易把图文识别为包含讽刺信息。一方面由于讽刺判别存在主观因素，另一方面则来源于模型本身的偏见性，若未经数据进行训练则表现出较差的结果。在经过相关数据训练后，得益于该模型本身较强的多模态理解能力，模型的偏见得到纠正后模型在该数据上的表现得到了显著提升。对比另外两个使用了完整训练数据的基线模型，结果会高于HFM，低于Multiview CLIP。但由于使用的训练数据量较少，并且进行训练的指令并不是专门为该任务所设计的，因此并没有得出使用全部数据完整训练后模型的上限。这里大模型仅仅使用10%的训练数据就得到了不错的结果。

在讽刺隐喻的生成与理解评测上，实验遵循4.2小节提出的使用大模型进行评测的方法进行评测。为了方便描述，下文将经过4.1节提到的全部混合数据训练的模型称为新模型，实验结果如表4所示。

模型	新模型胜率(%)	新模型获胜次数	原始模型获胜次数	无法评价次数
InternVL2-8B	55.5	111	89	0
deepSeek-vl-7b-chat	58.5	117	81	2

Table 4: 讽刺隐喻理解综合评测结果

经过所有数据训练后的模型获得了相对更高的胜率。该结果说明经过专门设计的讽刺隐喻数据微调的模型可以在讽刺与隐喻相关指令上获得更好的结果，但是两者的差别不是非常显著，也证明原始版本的模型在面对讽刺和隐喻相关的自由指令时也可以较好的进行回答。一些具体的例子被放在附录C中，包括GPT-4o设计的题目，微调前的模型与微调后的模型的输出，以及GPT-4o给出的判决结果以及判决理由。

5.2 通用能力评测

对于模型通用能力的评测，选择使用公开的多模态评测集来对模型进行评测。为了探究各种数据配置对通用能力表现的影响。实验测试以下几种模型并将其编号：1. 原始版本的模型；2. 使用4.1节的全量混合数据训练的模型；3. 不使用2000条补充的通用数据以及3.3节的第三大类的数据（领域内的通用数据）训练的模型；4. 仅使用2000条第一大类的讽刺衍生指令训练的模型。除了原始模型以外，剩下的模型随着编号的提升，指令多样性逐步降低。实验使用评测集MMBench和MMStar来对模型进行评测，其中MMBench分为英文评测和中文评测，实验结果见表5，表格中模型从上到下的顺序依次为上述测试模型的编号的顺序。

	编号	训练数据量	MMbench_EN	MMbench_CN	MMStar
InternVL2-8B	1	-	0.8178	0.8264	0.6113
	2	11400	0.8032	<u>0.8221</u>	<u>0.6100</u>
	3	8100	<u>0.8084</u>	0.8213	0.5980
	4	2000	0.7981	0.8204	0.6033
deepseek-vl-7b-chat	1	-	0.7423	0.7234	0.4120
	2	11400	0.7448	<u>0.7208</u>	0.4193
	3	8100	0.7363	0.7122	0.4073
	4	2000	0.7345	0.7105	0.3880

Table 5: 通用能力评测结果

观察表格可知，在Internvl2模型上，原始版本的Internvl2-8B在各个通用评测集上表现最好。经过微调后的模型不论何种数据配置，在通用评测集上的表现均有下降，但下降幅度均较小。综合来看，加入了领域内的通用数据和领域外通用数据进行训练的1号模型平均来说在各个通用评测集上的下降幅度最小，尤其在MMStar评测集上较好的保持了模型原始的能力。但不论怎样，各种数据配置的模型在通用数据上的表现均达到可以接受的程度，都没有在通用评测集上表现出大幅下降的趋势。

在DeepSeek-VL模型上，经过全部数据训练过的模型在两个评测集上超过了原始模型的表现。而在去除领域内的通用数据和领域外通用数据（3号模型），或只使用指令改写方法的讽刺模型（4号模型），在通用测试集的表现均有一定程度的下降。

综合来看，随着模型编号的提升（2号-4号），配置的指令微调数据的多样性与通用性逐渐降低，在通用测试集上的表现也逐渐降低。在使用全量数据（2号）训练的模型尽可能维持住了通用能力，在评测集上的表现略低或略高于原始模型。尽管在训练数据多样性最差4号模型中，我们也使用了指令的改写与衍生来尽可能拓展了模型指令以及回复的多样性，而不是全部使用单一的指令来让模型适应一个任务，来尝试维持模型的通用性能。由于1号模型在通用评测集上的表现在可接受的范围且差距不明显，本实验不再探究更多样的领域外的通用数据对比对模型在通用能力上表现的影响。

5.3 MMSD2.0补充实验

由于在MMSD2.0上使用了2000条与讽刺判别和解释相关的指令，本小节简单探究了不同

训练数据量以及使用中英文训练数据对模型结果的影响，其中en代表使用英文的prompt和英文的response；zh代表使用中文的prompt和中文的response。由于评测时间原因，本小节仅挑选MMSD2.0测试集中的200条进行测试，且仅测试InternVL2-8B模型，实验结果如表6所示。

数据设置	训练数据量	Acc	F1	Precision	Recall
500en	500	0.66	0.7142	0.5782	0.934
500zh	500	0.73	0.6966	0.7126	0.6813
500en+500zh	1000	0.695	0.7264	0.6136	0.8901
1000en	1000	0.72	0.7522	0.6296	0.934
1000zh	1000	0.745	0.7582	0.6667	0.8791
1000en+1000zh	2000	0.805	0.8	0.75	0.8571
all	11400	0.805	0.8	0.75	0.8571

Table 6: MMSD2.0上不同数据量以及中英文搭配的实验结果

观察上述实验结果，在2000条以内的数据量时，随着训练数据量的提升，模型在测试集上的表现会得到提升。在同样的数据量的情况下，中英文混合数据对模型的效果会低于只单独使用一种语言进行训练，这也在预期之中，因为混合两种语言时在测试时只使用一种语言的prompt，而想从另外一种语言上学习到类似的内容比从同一种语言上直接学习要困难。在使用一种语言的数据配置的基础上，再额外添加另一种语言的数据的情况下，模型在测试时的效果会得到提升。在使用500en+500zh的数据配置的情况下，与仅使用500en或500zh的结果的变化并不明显。这可能是由于训练数据量过小导致模型对中英文数据之间的理解能力还不够，尽管模型在中英文上的表现都足够强，但在领域数据内还需要更多的训练。而在1000en+1000zh的数据配置的情况下，对比1000en和1000zh的结果则有明显的提升。对比只使用领域内的2000条数据训练与使用了全量训练数据训练的模型，即使用了领域外的讽刺隐喻数据以及通用数据的模型，在MMSD2.0上的测试结果几乎相差不大，在表格中表现为相同的结果（观察两者输出并不相同）。其他的补充数据无法让模型迁移到MMSD2.0数据的讽刺识别上。

由于表6中最后两行的实验结果相同，为此我们重新设计了新的prompt来对最后两行的数据设置训练的模型进行测试，此次使用中文的prompt来测试，实验结果如表7所示。

数据设置	训练数据量	Acc	F1	Precision	Recall
1000en+1000zh	2000	0.755	0.7742	0.6667	0.9231
all	11400	0.8	0.76	0.84	0.6923

Table 7: MMSD2.0上使用中文prompt的部分实验结果

可以看出在更换中文prompt后，两种数据设置训练模型的结果产生了差异，可知模型对prompt敏感，这也保持了大模型的一贯作风，尤其在本任务中我们并没有选择单个固定的指令来完成单一的任务，而是选择了丰富的指令形式与不同的任务形式来对模型进行训练。

6 结语

在本文中，我们设计了多种多模态数据合成的方法，来构建讽刺隐喻多模态数据。本文利用不同的数据源，使用GPT-4o合成了讽刺隐喻领域的多模态图文指令微调数据。在进行数据构建时，本文将其分为指令数据的合成与回复数据的合成，并在合成时使用了三种不同的方法，来保证数据的复杂性与多样性。基于该数据训练的多模态模型，本文对其讽刺隐喻能力以及通用能力进行了综合评估，验证了该合成数据的有效性以及模型的可用性。使用该方法的模式可以借鉴使用到其他各个细分的多模态领域，来训练其他领域的垂域多模态大模型。本文的局限性在于对于提到的三种类型的合成数据以及通用多模态数据的配比并没有展开详细的研究，只选择了部分固定的数据配比方式进行实验。在未来的工作中，我们会进一步对上述不同方法合成数据的比例，以及通用数据的配比进行探索。

参考文献

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.
- Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.
- Shelly Dews and Ellen Winner. 1995. Muting the meaning a social function of irony. *Metaphor and Symbol*, 10(1):3–19.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. 2024. Gpt-4o system card.
- Luuk Lagerwerf and Anoe Meijers. 2008. Openness in metaphorical and straightforward advertisements: Appreciation effects. *Journal of Advertising*, 37(2):19–30.
- Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 via metaphor detection shared task. In *Proceedings of the workshop on figurative language processing*, pages 56–66.
- Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 via and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023b. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR.
- Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 23–33.
- Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. *arXiv preprint arXiv:2307.07135*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Yue Wang, Haoke Zhang, Juntao Li, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. 2024. Sass: Self-alignment with semi-supervised instruction data generation.
- Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.
- Yorick Wilks. 1978. Making preferences more active. *Artificial intelligence*, 11(3):197–223.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024a. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.
- Zichen Wu, Hsiu-Yuan Huang, Fanyi Qu, and Yunfang Wu. 2024b. Mixture-of-prompt-experts for multi-modal semantic understanding. *arXiv preprint arXiv:2403.11311*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*.

Ming Zhang, Ke Chang, and Yunfang Wu. 2024. Multi-modal semantic understanding with contrastive cross-modal feature alignment. *arXiv preprint arXiv:2403.06355*.

附录A

我们使用如下人工编写的指令作为初始种子指令集，对于图文数据，人工编写的指令集如图6：对于图片数据，人工编写的指令集如图7，注意此处相比图文数据额外增加了创作型指

- 这里有一个图文对，文本为{text}，请帮我判断一下这个图文对是否包含讽刺，并给出解释。请先输出：有讽刺/无讽刺。再给出解释。
- 这里有一个图文对，文本为{text}，请帮我判断一下这个图文对是否包含隐喻，并给出解释。请先输出：有隐喻/无隐喻。再给出解释。
- 这个图片和文本：{text}，构成讽刺表达吗？如果有，请给出详细的解释。
- 这个图片和文本：{text}，构成隐喻表达吗？如果有，请给出详细的解释。
- This is a text-image pair. The text is {text}. Please determine whether this text-image pair contains metaphors and provide an explanation. Please firstly output: Contains metaphors / Does not contain metaphors. Then provide an explanation.
- This is a picture-text pair, with the text being {text}. Please determine whether this picture-text pair contains sarcasm and provide an explanation. Please firstly output: Contains sarcasm / Does not contain sarcasm. Then provide an explanation.
- Does this picture and the text: {text} form a metaphor? If so, please explain it to me.
- Do this picture and the text: {text} form sarcasm? If so, please explain it to me.

Figure 6: 图文数据的初始指令集

令，即给出一张图片生成出可能相配对的讽刺或隐喻的文本内容。

附录B

此附录介绍在使用GPT-4o合成3.2节所提到的第二、三类指令数据与回复数据的prompt。对于第二类指令与回复数据合成时，我们仍将其划分为图文数据与图片数据，对于图文数据，prompt如图8所示。对于图片数据，prompt如图9所示。

对于第三类指令与回复数据合成时，我们仅对图片数据进行合成，prompt如图10所示。

附录C

此附录展示一些利用GPT-4o进行讽刺隐喻理解评测的一些样例，包括GPT-4o设计的题目，微调前后模型的输出，以及GPT-4o给出的判别结果以及解释。

示例1:

输入图片如图11

题目: 这幅图通过“公正”和“公平”的对比传递了什么隐喻意义？请详细解释。

InternVL2-8B-ms:

这幅图通过“公正”和“公平”的对比传递了隐喻意义：左边展示的是“公正”，即每个人都站在相同的高度，但实际结果却是有人无法看到比赛，这隐喻了公正可能忽视个体差异，导致结果的不平等；右边展示的是“公平”，即通过提供不同的平台（如梯子），让每个人都能站在相同的高度，这隐喻了公平考虑到了个体差异，通过提供机会让每个人都能获得平等的结果。

InternVL-8B:

这幅图通过“公正”和“公平”的对比，传递了关于社会正义和公平的隐喻意义。

- 请为这张图片写一段文字，来使得构成的图文对包含隐喻。如果难以完成，则拒绝回答。
- 请为这张图片写一段文字，来使得构成的图文对包含讽刺。如果难以完成，则拒绝回答。
- Please write a paragraph of text for this picture to make the combination of the picture and text contain a metaphor. If it is difficult to complete, decline to answer.
- Please write a paragraph of text for this picture to make the combination of the picture and text contain sarcasm. If it is difficult to complete, decline to answer.
- 请帮我判断一下这个图片是否包含讽刺，并给出解释。请先输出：有讽刺/无讽刺。再给出解释。
- 请帮我判断一下这个图片是否包含隐喻，并给出解释。请先输出：有隐喻/无隐喻。再给出解释。
- 这个图片有讽刺表达吗？如果有，请给出详细的解释。
- 这个图片有喻表达吗？如果有，请给出详细的解释。
- 请判断这张图片是否含有讽刺或者隐喻的意味，并给出解释。
- 这张图片构成隐喻或者讽刺吗？如果有，分别帮我解释一下。
- Please tell me whether this picture contains sarcasm and provide an explanation. Please firstly output: Contains sarcasm / Does not contain sarcasm. Then provide an explanation.
- Please tell me whether this picture contains metaphors and provide an explanation. Please firstly output: Contains metaphors / Does not contain metaphors. Then provide an explanation.
- Do this picture form sarcasm? If so, please explain it to me.
- Do this picture form a metaphor? If so, please explain it to me.

Figure 7: 图片数据的初始指令集

首先，从图中可以看到，左边写着“这是公正”，右边写着“这是公平”。在左边，三个人站在不同的高度，但他们的双臂都伸向同一个方向，似乎在努力达到同一个目标。而在右边，三个人站在相同的高度，但他们的双臂伸向不同的方向，似乎在各自追求自己的目标。

这个对比隐喻了公正和公平的区别：

1. ****公正****：公正指的是每个人都应该得到他们应得的，无论他们的起点如何。图中左边的人虽然站在不同的高度，但他们的目标是一样的，这象征着公正。公正强调的是结果的平等，即每个人都应该得到他们应得的回报，无论他们的背景如何。
2. ****公平****：公平指的是每个人都应该有相同的机会去追求他们的目标。图中右边的人站在相同的高度，但他们的目标不同，这象征着公平。公平强调的是起点的平等，即每个人都应该有相同的机会去追求他们想要的目标，无论他们的背景如何。

通过这个对比，这幅图传达了一个重要的信息：公正和公平是两个不同的概念，但它们都是社会正义的重要组成部分。

GPT-4o评价：

答案1

理由：答案1准确解释了图片中“公正”和“公平”的对比，并且结合图片中的高度差异，以及额外工具的分配，清晰地说明了每种情况的隐喻意义。它正确地指出“公正”忽视个体差异导致结果不平等，而“公平”通过调整资源使每个人在结果上得到平等体验。相比之下，答案2虽然提及了社会正义和公平的概念，但对图片的描述存在偏差（如“手臂伸向不同方向”等与实际图示不符），且对“公正”和“公平”的讨论较为抽象，与图片的具体内容有一定脱节，因此较不精确。

示例2：

输入图片如图12

题目：根据这张图片，设计一个包括讽刺意味的广告语，用于提醒人们警惕“自我毁灭式的盲目行为”。

deepseek-v1-7b-chat-ms:

“当理智被吞噬，你以为你是谁？——别让内心的怪兽吞噬你的未来。”

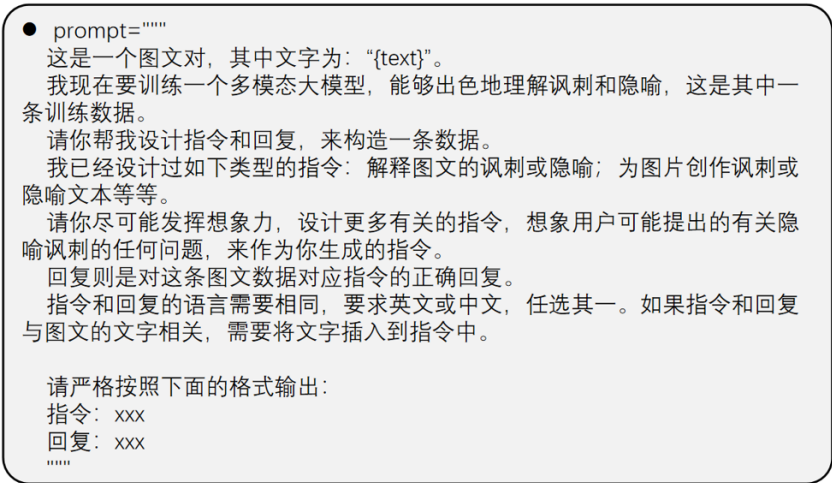


Figure 8: 使用GPT-4o构造图文讽刺隐喻随机指令及回复的prompt

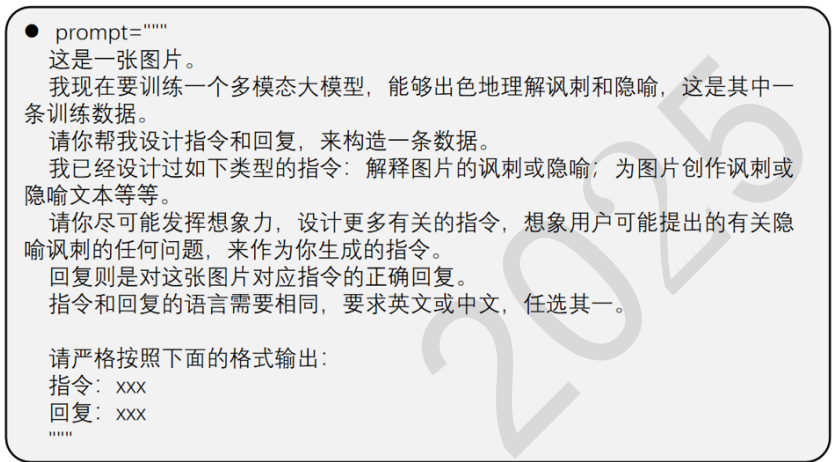


Figure 9: 使用GPT-4o构造图片讽刺隐喻随机指令及回复的prompt

deepseek-vl-7b-chat:

这幅图像呈现了一个引人注目的视觉效果，以一个黑白对比鲜明的人头侧影为主体，背景是纯蓝色的。人头侧面朝右，眼睛闭着，仿佛在沉思或冥想。然而，头部的右侧被一个黑色的锯齿形图案遮盖，暗示着一种危险或威胁。

图像的质感和纹理增加了其深度，头部的白色部分有明显的磨损和污迹，暗示了时间的流逝或忽视。蓝色背景上点缀着细小的白色点，营造出一种空灵或梦幻的氛围。

结合这些元素，可以设计一个讽刺意味的广告语，提醒人们警惕“自我毁灭式的盲目行为”。例如：

“当你的生活被锯齿形的危险吞噬，你的头颅却盲目地忽视磨损的迹象。”

这个广告语利用了图像中锯齿形图案的象征意义，暗示了自我毁灭的行为，同时强调了忽视警告标志（在这里是头部的磨损）的盲目性。

GPT-4o评价:

答案1 理由：答案1的广告语直接针对“自我毁灭式的盲目行为”主题，利用“怪兽”和“吞噬”的隐喻生动且紧扣图像中锯齿形图案所表达的危险感，更加简洁、有力，能直观传达警示意义。相比之下，答案2对图像的分析虽详尽，但广告语本身冗长且复杂，讽刺意味和情感冲击力较弱。

●

prompt = ""

我现在要训练一个强大的通用多模态大模型。请你根据这张图片帮我设计指令和回复，来构造一条数据。
任务包括但不限于：VQA（视觉问答），OCR（光学字符识别），Visual Reasoning（视觉推理）
设计的指令和回复可以尽量复杂且包含推理，具体根据图片内容决定。
指令和回复的语言需要相同，要求英文或中文，任选其一。
只设计一条指令和回复。

请严格按照下面的格式输出：

指令：xxx

回复：xxx

""

Figure 10: 使用GPT-4o构造图片讽刺隐喻随机指令及回复的prompt

A cartoon illustration of a sports field with a goal. A player in a blue jersey is standing on a wooden crate to take a shot. Other players in red and blue jerseys are also standing on crates. The text "这是公正" (This is justice) is on the left and "这是公平" (This is fairness) is on the right.

Figure 11: 示例1图片

第二十四届中国计算语言学大会论文集，第210页-第226页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

225

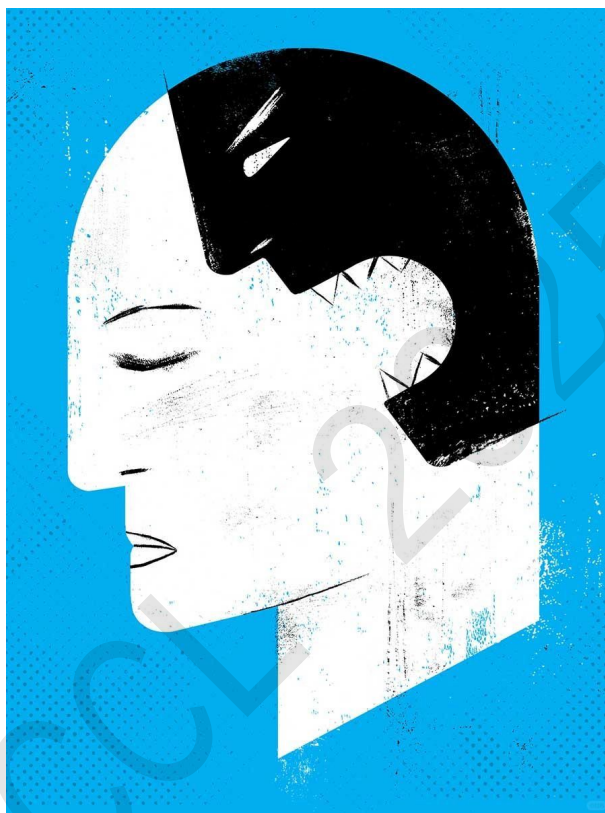


Figure 12: 示例2图片