基于自提示多模态大语言模型和语义感知离散扩散模型的图像描述生成算法

陈宇峰[1]，**江爱文**[1,2*]，**黄琪**[1]，**王明文**[1,2]
[1]江西师范大学，计算机信息工程学院，江西，南昌，330022
[2]江西师范大学，数字产业学院，江西，上饶，334000
Email: {chenyufeng, jiangaiwen, huangqi, mwwang }@jxnu.edu.cn

摘要

近年来，非自回归图像描述生成技术凭借其双向传播和并行词语生成的能力受到广泛关注。与此同时，基于离散扩散方法的研究也取得了显著进展。然而，在离散噪声添加与去噪过程中，现有方法仍面临图像文本关联性低、目标物体遗漏、描述准确性不足以及词语重复等关键问题。为应对这些挑战，我们提出一种基于语义感知的离散扩散模型。该模型通过可学习查询机制构建语义感知模块，以捕捉与图像物体级语义特征的潜在关联从而更好地生成图像描述。在此基础模型之上，我们进一步引入自提示优化框架，利用大语言模型生成与图像细节内容更相符的丰富描述。在COCO数据集上的综合实验表明，本方法在图像描述任务中取得一定的提升，其性能优于现有的相关方法。

关键词： 图像描述；离散扩散模型；大语言模型；自提示

Exploring Semantic-aware Discrete Diffusion Model via Self-Prompting Multimodal LLMs for Image Captioning

Yufeng Chen[1]，**Aiwen Jiang**[1,2*]，**Qi Huang**[1]，**Mingwen Wang**[1,2]
[1]School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi , 330022
[2]School of Digital Industry, Jiangxi Normal University, Shangrao, Jiangxi, 334000
Email: {chenyufeng, jiangaiwen, huangqi, mwwang }@jxnu.edu.cn

Abstract

In recent years, non-autoregressive image captioning has gained significant attention due to its capabilities in bidirectional propagation and parallel word generation. Meanwhile, considerable progress has been made in research on discrete diffusion-based approaches. However, during the processes of discrete noise addition and denoising, existing methods still face critical challenges such as weak image-text relevance, object omission, inaccurate descriptions, and word repetition. To address these issues, we propose a semantic-aware discrete diffusion model. This model incorporates a learnable query mechanism to construct a semantic perception module, which captures latent correlations with object-level semantic features in images, thereby improving caption generation. Building upon this foundational model, we further introduce a self-prompting optimization framework that leverages large language models to generate richer descriptions that better align with image details. Comprehensive experiments

*通讯作者

第二十四届中国计算语言学大会论文集，第434页-第447页，济南，中国，2025年8月11日至14日。

on the COCO dataset demonstrate that our method achieves notable improvements in image captioning tasks and outperforms existing relevant approaches

Keywords: Image Captioning , Discrete Diffusion Model , Large Language Model , Self-Prompting

1 引言

图像描述生成作为一项重要的跨模态任务，旨在自动生成与图像内容语义一致且生动全面的文字描述。目前主流方法多采用编码器-解码器结构，以自回归方式进行学习 (Cornia et al., 2020; Li et al., 2022; Yao et al., 2018; Herdade et al., 2019)，逐词解码生成描述。自回归方法存在解码过程中仅允许单向文本信息传递等局限性。为此，近年来研究者开始探索非自回归方法 (Guo et al., 2021; Yan et al., 2021; Yu et al., 2023)，该方法支持双向文本信息传递与并行词语生成。然而，由于缺乏对文本序列信息的依赖，非自回归方法在主流客观评价指标上的表现往往逊色于自回归方法。

随着扩散模型 (Ho et al., 2020; Song et al., 2021)在文生图领域取得显著成功，研究者在反向的图生文领域也展开了探索性工作，从而扩展了非自回归生成方法的应用。Austin等人 (Austin et al., 2021)提出在离散状态空间实现词语级的噪声添加与去噪过程。在离散扩散模型中，前向扩散过程通过转移矩阵逐步添加噪声，将原始文本退化为无意义的"[MASK]"标记；反向过程则逐步从全"[MASK]"状态恢复文本。该研究将转移矩阵作为关键设计要素，这一改进在文本领域取得了一定的成效。He等人 (He et al., 2023b)同样基于离散扩散模型提出了DiffusionBERT。尽管利用了预训练模型的优势，但离散扩散特有的生成策略使其难以捕捉词语依赖关系。此外，由于文本中存在大量无意义的"[MASK]"标记，这增加了捕捉图文关联的难度，可能导致部分生成内容与图像无关。

大型语言模型（LLMs）在当今人工智能技术发展进程中发挥着重要作用。最新研究表明，诸如InstructGPT (Ouyang et al., 2022)和LLaVA (Liu et al., 2023)等大型语言模型都展现出卓越的生成能力。当提供经过精心设计的任务特定指令时，这些LLMs能够有效执行包括图像描述生成在内的下游任务。然而，若直接使用基于模板的提示而未包含详细视觉线索，则可能无法充分捕捉图像语义信息，甚至可能导致幻觉性内容生成。

为缓解上述问题，我们提出了一种基于语义感知的离散扩散图像描述生成模型。该模型通过语义感知模块更好地捕获图像视觉特征中的语义信息，生成图像语义的标记表示，并且将这些语义的标记与文本表示进行融合。这种设计使模型在去噪阶段能够更有效地建模图像与文本的跨模态关联，从而生成语义一致且细节丰富的描述。在一定程度上减轻扩散过程中语义信息及相关实体丢失等问题。

与此同时，基于上述提出的基础模型，我们进一步构建了自提示描述优化框架。通过将图像描述任务抽象为相关问答任务，并为大语言模型提供通用问答模板，从而获取包含图像细节信息的答案。我们将这些答案提示与初始生成的描述作为语义线索，借助大语言模型进一步丰富图像文本描述。该框架通过自提示生成方法，利用大语言模型的推理能力产生包含丰富细节信息的答案提示，继而优化基础模型生成的描述。这种方法不仅缓解了错误内容的出现情况，还能生成更贴合图像细节的丰富描述。

本文的主要贡献可归纳为以下三个方面：

- 提出了一种有效的语义感知离散扩散模型，通过增强离散扩散过程中文本标记间的语义关联性，缓解了去噪阶段图文相关性捕获困难等问题。
- 设计了面向图像描述优化的自我提示框架，通过利用大语言模型的生成能力，有效优化和润色了描述的细节与丰富度。
- 在公开大规模数据集COCO上，将提出的基础模型及优化框架与现有先进方法进行对比实验。结果表明，无论是基础描述还是优化后的描述，其生成质量均获得显著提升。

第二十四届中国计算语言学大会论文集，第434页-第447页，济南，中国，2025年8月11日至14日。

2 相关工作

主流图像描述生成方法主要采用两种策略：自回归与非自回归。早期的图像描述方法普遍基于自回归策略 (Anderson et al., 2018; Karpathy and Fei-Fei, 2015)，采用逐词解码方式生成文本描述。具体而言，在解码阶段，这些方法通过单向融合图像特征与已生成的文本来预测后续词语。随着Transformer在深度学习领域取得了巨大的成功，AoANet (Huang et al., 2019)和M2Transformer (Cornia et al., 2020)等模型也通过使用注意力机制实现图像与文本的跨模态交互。

非自回归模型是一种不依赖已生成标记来预测序列的生成模型。该模型能够实现文本的并行输出也在机器翻译、图像描述生成等多个领域展现出巨大潜力，相关方法如MNIC (Gao et al., 2019)、MIR (Lee et al., 2018)和SATIC (Zhou et al., 2021)等已获得广泛关注。然而，由于缺乏对生成序列内部依赖关系的建模，此类模型往往存在生成准确性不足的问题。

随着预训练模型的发展，非自回归方法在文本生成领域进一步取得显著进展。相关研究已将预训练模型与非自回归方法相结合 (Su et al., 2021)。此外，随着扩散模型在文生图任务中的重大成功，研究者也开始探索将扩散模型作为非自回归文本生成模型的应用。

目前，基于扩散模型的文本生成方法仍处于探索阶段，主要可分为两类。第一类基于连续扩散模型 (Chen et al., 2023; He et al., 2023a)，其遵循传统扩散过程，将文本嵌入连续向量或二进制字节空间后添加高斯噪声进行预测。然而，由于语言标记本身包含语义信息与上下文关联，直接在编码文本特征上添加噪声可能导致语义丢失与词语重复问题。此外，即使词语特征去噪效果良好，也未必能保证与原始词语的一致性 (He et al., 2023a)。

第二类基于离散扩散模型，与连续扩散模型不同，其在词语标记和句子层面进行噪声添加与去除，而非作用于词向量空间。噪声词语被替换为"[MASK]"标记，去噪过程则逐步恢复这些标记。代表性方法包括DiffusionBERT (He et al., 2023b)、Diffusion-NAT (Zhou et al., 2024)和DDCAP (Zhu et al., 2022)、VCC-DiffNet(Cheng et al., 2024)、SD^3(Zhao et al., 2024)等。离散生成策略支持词语并行生成，结合预训练语言模型可快速生成完整句子。但在图像描述任务中，离散噪声添加与去除过程仍存在图文内容不相关、目标遗漏、描述不准确及词语重复等挑战性问题。

大语言模型在图像描述任务中的应用研究已取得一定进展。主流方法主要通过提示工程引导大语言模型生成丰富细致的描述 (Wu et al., 2023)，但这类方法存在生成过程随机性强、可控性不足等局限性。此外，大语言模型可能无法充分理解图像关键细节，从而导致幻觉信息生成。

在视觉问答任务中，Hu等人 (Hu et al., 2023)提出利用生成描述及相关问题提示大语言模型产生准确答案。受此启发，本研究逆向利用大语言模型的视觉问答能力提取图像细节信息，将获取的答案与初始生成的图像描述相结合，实现描述的优化与文本润色，有效地提升了生成描述的准确性与细节完整性。

3 方法

本节提出了一种用于图像描述生成的语义感知离散扩散模型（SeDDM）。如图 1所示，该模型架构包含三个核心组件：语义感知模块、语义融合模块以及基于Transformer的扩散解码器。在此基础上，我们进一步构建了一个大语言模型驱动的自提示优化框架。如图 4所示，该框架通过将图像描述任务形式化为结构化问答范式，有效利用大语言模型的语义推理能力来增强描述生成的质量。

3.1 图像特征提取

给定输入图像I，我们分别从CLIP图像编码器 (Radford et al., 2021)和Faster R-CNN (Ren et al., 2017)中提取区域级视觉特征。从CLIP图像编码器获取的网格特征f_G能够提供与图像的丰富语义对齐，而通过Faster R-CNN提取的区域特征f_R可提供对象级信息，这对描述图像的细节内容至关重要。我们在不同阶段利用这些特征进行视觉引导。

3.2 语义感知模块

我们定义一组可学习的查询标记P_s 作为语义潜在提示。提出的语义感知模块旨在捕获图像中的关键视觉信息，其结构如图 2所示。通过该模块，可学习查询能够更好地捕获视觉特征中

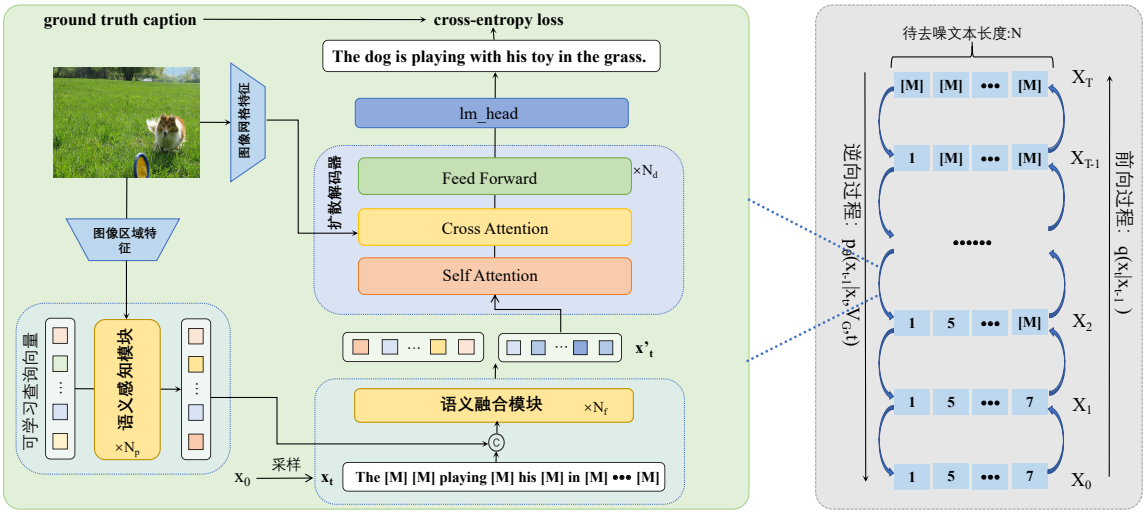


图 1: 语义感知离散扩散模型的结构细节

的语义信息，并实现更好的对齐，生成图像语义的标记表示。

具体而言，首先对可学习查询P_s. 进行自注意力计算建立输入序列内部元素的全连接依赖关系，随后通过区域视觉特征f_R 与可学习查询之间的交叉注意力机制，获得图像对齐的标记表示，使可学习的标记能够从图像特征中捕捉语义信息。我们采用N_s 层的语义感知计算，使可学习查询能够充分捕获视觉细节，最终得到表征$X_s = P_s^{N_s}$。该语义感知模块的计算过程如公式 1所示。

$$\begin{aligned} \widetilde{P}_s^i &= \mathrm{LN}(P_s^i + \mathrm{MHA}(P_s^i, P_s^i, P_s^i)) \\ P_s^{i+1} &= \mathrm{FFN}(\mathrm{LN}(\widetilde{P}_s^i + \mathrm{MHA}(\widetilde{P}_s^i, f_R, f_R))) \end{aligned} \tag{1}$$

其中，LN表示层归一化，MHA代表多头注意力机制，FFN指前馈神经网络层。$i = \{0, 1, ..., N_s - 1\}$，$P_s^0 = P_s$。

3.3 语义融合模块

语义融合模块与后续的离散扩散Transformer协同工作。在离散扩散模型的逆向过程中，模型需要从含有大量掩码标记序列中预测恢复文本。由于文本中含有大量的掩码标记导致缺乏上下文信息，很难捕捉上下文之间的联系以及与图片信息之间的相关性。语义融合模块则可以利用图像语义感知模块先前捕获的视觉语义信息标记，将其与嘈杂的文本进行连接和融合，以获得语义增强的编码表示。该融合过程通过一个融合编码模块实现，以获取语义增强的编码表示。在每次去噪步骤前，我们将语义感知模块的输出X_s 与噪声文本嵌入X_t 进行拼接，得到$X_T{}^{(0)} = [X_s, X_t]$。随后，通过N_f 层的语义融合模块来学习语义增强的表示。该语义融合的过程如公式 2所示：

$$X_T{}^{(i+1)} = \mathrm{FFN}(\mathrm{LN}(X_T^{(i)} + \mathrm{MHA}(X_T^{(i)}, X_T^{(i)}, X_T^{(i)}))) \tag{2}$$

其中，$i = \{0, 1, ..., N_f - 1\}$。

我们将最终输出$X_t{}^{(N_f)}$ 作为离散扩散变换器的输入$X_t{}'$ 用于生成图像描述。

3.4 离散扩散模型

离散扩散将$\mathbf{X}_t = \{w_1, w_2, w_3, ..., w_n\}$ 中的每个词视为具有K 个类别的离散随机变量，其中K 表示词汇表大小。前向加噪过程可如公式 3所示进行表示。

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathrm{Cat}(\mathbf{x}_t; \mathbf{p} = \mathbf{x}_{t-1}\mathbf{Q}_t) \tag{3}$$

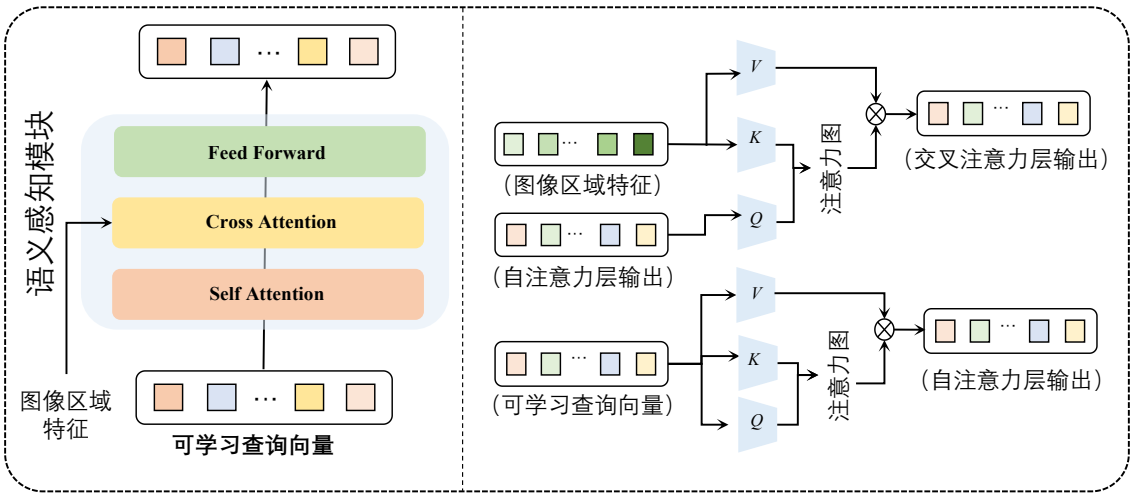


图 2: 语义感知模块结构图

其中，Cat(·) 表示类别概率分布。$\mathbf{Q}_t$ 是一个转移矩阵，该矩阵被独立地应用于序列中的每个标记。其中，$[\mathbf{Q}_t]_{ij} = q(x_t = j|x_{t-1} = i)$定义了从状态$i$到状态$j$的转移概率。

为充分利用预训练语言模型的能力，采用掩码机制对$\mathbf{X}_t$中的每个词语进行噪声处理。在加噪过程中，部分词语保持不变，部分词语则以特定概率被替换为”[MASK]”标记。其中转移矩阵如公式 4所示：

$$[\mathbf{Q}_t]_{i,j} = \begin{cases} 1 & \text{if} \quad i = j = [\mathrm{M}], \\ \beta_t & \text{if} \quad j = [\mathrm{M}], i \neq [\mathrm{M}], \\ 1 - \beta_t & \text{if} \quad i = j \neq [\mathrm{M}], \end{cases} \tag{4}$$

具体而言，在前向过程中，若文本中的某个词在步骤$t-1$时为”MASK”标记，则该词在步骤t 时保持为”MASK”的概率为1。若该词在步骤$t-1$时非”MASK”标记，则在步骤时t时，以β_t的概率变为”MASK”标记，或以$1-\beta_t$的概率保持不变。经过T 步加噪后，整个文本序列将被替换为全掩码序列。

第t步的概率分布$q(\mathbf{x}_t^i|\mathbf{x}_0^i)$可以通过一个简单的闭环直接得到，如公式 5 所示。

$$q(\mathbf{x}_t^i|\mathbf{x}_0^i) = \begin{cases} \overline{\alpha}_t & \text{if} \mathbf{x}_t^i = \mathbf{x}_0^i, \\ 1 - \overline{\alpha}_t & \text{if} \mathbf{x}_t^i = [\mathrm{M}], \end{cases} \tag{5}$$

其中，$\overline{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. $\mathbf{X}_t^i$ 表示第t 步时文本序列中的第i 个标记。

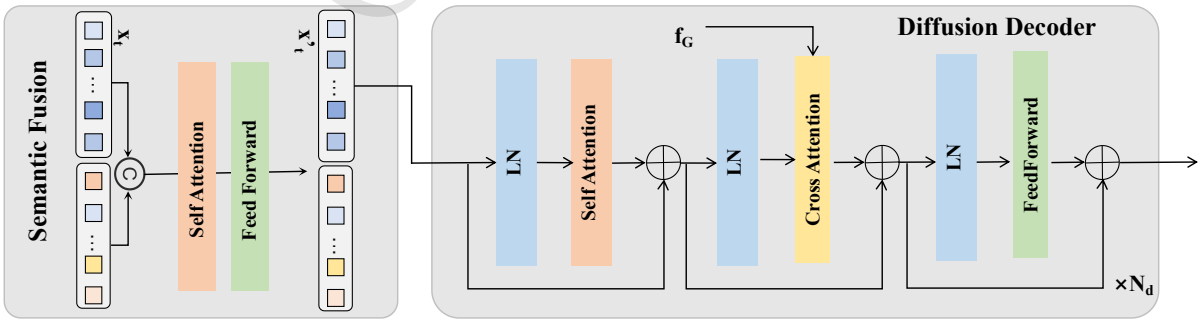


图 3: 语义融合模块与离散扩散解码器结构图

本文采用扩散Transformer网络进行噪声预测。如图 4所示。在交叉注意力层中，网格特征f_G 作为查询向量，为去噪语义增强的$\mathbf{X_t}'$提供引导。时间步t 通过正弦位置编码嵌入。令$\mathbf{z_t} = \mathbf{X_t}'$具体计算公式 6如下。

$$
\begin{aligned}
z_t^i &= z_t^i + \mathrm{MHA}\left(\mathrm{LN}\left(z_t^i\right), \mathrm{LN}\left(z_t^i\right), \mathrm{LN}\left(z_t^i\right)\right) \\
\hat{z}_t^i &= z_t^i + \mathrm{MHA}\left(\mathrm{LN}\left(z_t^i\right), V_G, V_G\right) \\
z_t^{i+1} &= \hat{z}_t^i + \mathrm{FFN}\left(\mathrm{LN}\left(\hat{z}_t^i\right)\right)
\end{aligned} \tag{6}
$$

在训练阶段，将$\mathbf{X}_t$ 中的所有"[MASK]"标记恢复为原始文本形式，并采用交叉熵损失函数训练基础模型。在在推理阶段，模型首先基于$\mathbf{X}_t$ 预测中间结果$\widetilde{\mathbf{X}}_0$，随后施加适量噪声生成$\mathbf{X}_{t-1}$．通过迭代执行上述过程，最终获得去噪结果$\mathbf{X}_0$。该逆向扩散过程可形式化表示为公式 7:

$$
P_\theta(\boldsymbol{x}_0) = \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_R, f_G, t) \tag{7}
$$

3.5 自提示图像描述优化

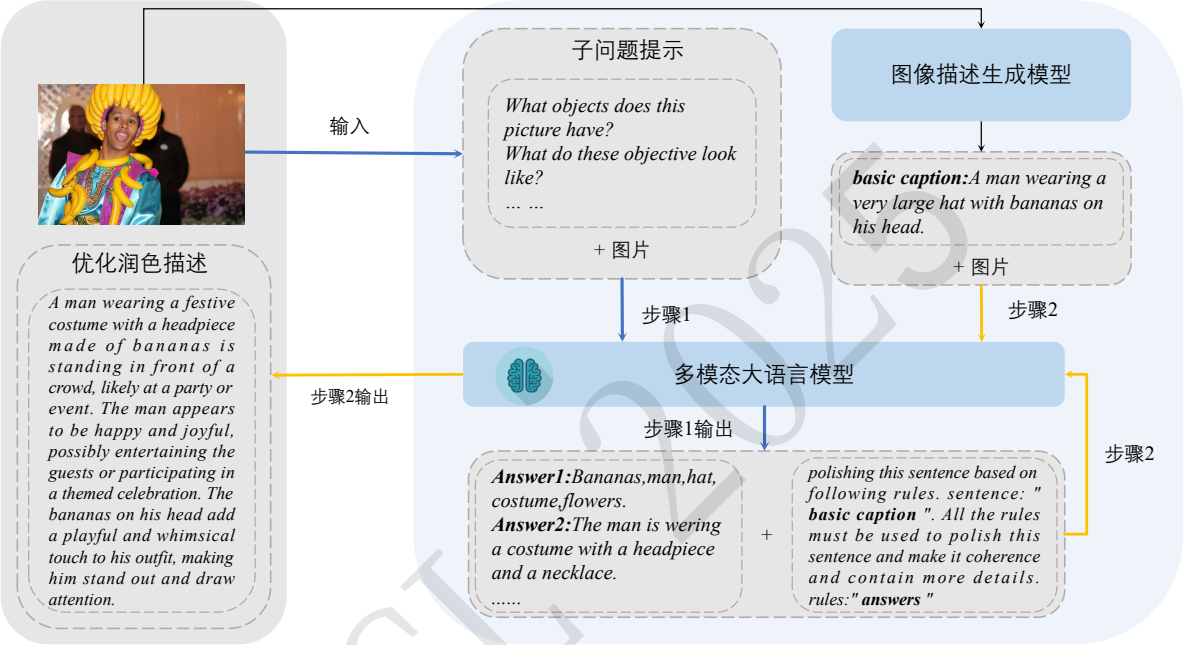


图 4: 基于多模态大语言模型的自提示图像描述优化框架

为从人类认知视角实现图像描述生成任务，本文提出了一种自提示优化框架，通过多模态大语言模型对语义感知离散扩散模型（SeDDM）生成的初始描述进行深度分析与优化，从而产生更准确、更丰富的图像文本描述。该方法的整体架构如图 4所示。

表 1: 子问题提示示例

What objects does this picture have?
What do these objects look like?
What is the background of this picture?
What feelings do the background and objects show?

在图像描述生成任务中，主体对象的准确识别与表征构成核心要素。本框架通过多模态大语言模型的推理能力，从视觉实体、场景信息以及情感语义等多维度进行图像细粒度分析。我们将描述优化任务解构为若干可操作的子问题，通过子问题求解获取的细粒度视觉知识作为外部知识注入来增强描述质量。具体的子问题设计如表 1所示。

如图 4所示，本框架采用一种基于两阶段优化的多模态描述增强策略：第一阶段，将输入图像与一组精心设计的子提示问题输入LLaVA-1.5(Liu et al., 2024b)多模态大语言模型

（MLLM）。其中，图像作为视觉信息输入到多模态大语言模型中，而子提示问题则引导模型关注图像中的特定语义内容。通过这种处理方式，模型生成与视觉内容高度相关的结语义信息，使之提取图像中的关键元素，包括对象、背景及其所蕴含的情感信息等。第二阶段，将第一阶段获得的语义提示答案与SeDDM生成的基础描述Dbase进行融合，并配合预设的指令模板再次输入MLLM，以增强其文本生成能力。通过这一方式，模型能够生成细节更加丰富且更贴合图像内容的描述，从而提升最终图像描述的质量。其具体指令模板定义如下：

polishing this sentence based on following rules.
sentence: " **basic caption** ".
All the rules must be used to polish this sentence and make it coherence and contain more details.
rules:" **answers** "

其中，**basic caption**为SeDDM生成的基础描述，**answers**为MLLM基于子提示问题生成的回答。

4 实验

4.1 数据集与评测指标

COCO数据集作为当前图像描述生成领域最广泛使用的基准数据集，包含82,783张训练图像、40,504张验证图像和40,775张测试图像，每张图像均配有5条人工标注描述文本。本研究采用Karpathy划分方案 (Karpathy and Fei-Fei, 2015)，使用5,000张图像作为验证集，5,000张作为测试集，其余用于模型训练。

针对基础语义感知离散扩散模型（SeDDM）的性能评估，我们采用五个主类通用指标：BLEU@N (Papineni et al., 2002)、METEOR (Banerjee and Lavie, 2005)、ROUGE (Lin, 2004)、CIDEr (Vedantam et al., 2015)以及SPICE (Anderson et al., 2016)。

对于自提示优化框架的评估，由于优化文本缺乏标准参考，为保证评价的公平性与客观性，受Ning(Ning et al., 2023)等人启发，我们引入大语言模型Llama3 (Huang et al., 2024)作为语言评价器，从两个维度进行评分：1) 丰富度（评分区间0-1）；2) 细节度（以细节数量计分）。此外，我们还使用无参指标CLIPScore(Hessel et al., 2021)来计算图像与文本之间的相似度，其主要通过将图片与本文信息经过CLIP编码，并计算其跨模态余弦相似度的值作为图文对齐度评价的结果。计算过程可以表示为公式 8：

$$\text{CLIPScore}(\mathbf{c}, \mathbf{v}) = \max(cos(\mathbf{c}, \mathbf{v}), 0) \tag{8}$$

其中，c表示CLIP文本嵌入向量，v表示图像嵌入向量。此外，我们也进行了人工评估以全面衡量生成质量。

4.2 实验细节

本研究采用ViT-B/16架构的预训练CLIP模型 (Dosovitskiy et al., 2021)和Faster-RCNN作为特征提取器。主要参数配置如下：可学习查询向量个数设为20，语义感知模块深度($N_s = 6$)，语义融合模块深度($N_f = 3$)．离散扩散模型基于DDCap模型架构 (Zhu et al., 2022)实现，优化器选用AdamW (Ilya Loshchilov, 2019)，损失函数为交叉熵，学习率设置为2e-5。实验在单块NVIDIA 3090 GPU上完成，批量大小为64，训练周期为30轮次。自提示框架中的多模态大语言模型采用LLAVA-1.5-7B (Liu et al., 2024b)，基础描述生成器为语义感知离散扩散模型SeDDM。

4.3 实验结果与分析

本文提出的语义感知离散扩散模型（SeDDM）在COCO Karpathy测试集上取得了具有竞争力的性能表现，具体结果如表 2所示。

作为一种非自回归方法，本模型相较于自回归方法展现出稳健的性能优势。特别地，我们的模型一定程度上改善了传统非自回归方法的缺陷，在部分指标上达到了与ViTCap模型相当的结果。

相较于现有非自回归方法，本模型在相关评估指标上表现出较好的性能提升。在METEOR、CIDEr和SPICE指标上，较DiffCap、Bit Diffusion、DDCap等最新扩散模型均

有显著改进，表明本方法在语义准确性、描述丰富性和场景理解深度方面具有明显优势，能够生成更贴合图像内容、更具语言多样性的高质量描述。同时，SeDDM在这三项指标上分别超越SCD-net模型0.6、1.5和0.6个点，也表明本模型能够生成更丰富、更准确的图像描述。值得注意的是，SCD-net采用基于检索的策略，将检索到的图像描述作为先验知识来指导新描述的生成。这种方法相当于直接提供了一个图像描述用以指导生成，虽然SCD-net在B@4和ROUGE指标上得分较高，但本模型在其他指标上均取得更优结果，这也凸显了我们的方法能够更好地捕捉文本关联性与图像相关性，从而生成更优质的图像描述。

表 2: 本模型与当前先进的自回归及非自回归方法在COCO Karpathy测试集上的对比结果

模型	B@4↑	METEOR↑	ROUGE↑	CIDEr↑	SPICE↑
自回归方法					
RFNet (Jiang et al., 2018)	35.8	27.4	56.5	112.5	20.5
Up-Down (Anderson et al., 2018)	36.2	27.0	56.4	113.5	20.3
GCN-LSTM (Yao et al., 2018)	36.8	27.9	57.0	116.3	20.9
Transformer (Sharma et al., 2018)	34.0	27.6	56.2	113.3	21.0
AoANet (Huang et al., 2019)	37.2	28.4	57.5	119.8	21.3
VitCap (Fang et al., 2022)	35.7	28.8	57.6	121.8	22.1
非自回归方法					
MIR (Lee et al., 2018)	32.5	27.2	-	109.5	20.6
MNIC (Gao et al., 2019)	30.9	27.5	55.6	108.1	21.0
SATIC (Zhou et al., 2021)	32.9	27.0	-	110.0	20.6
Bit Diffusion (Chen et al., 2023)	34.7	-	**58.0**	115.0	-
DDCap (Zhu et al., 2022)	34.4	28.1	57.1	117.9	21.6
DiffCap (He et al., 2023a)	31.6	26.5	57.0	104.3	19.6
SCD-Net (Luo et al., 2023)	**37.3**	28.1	**58.0**	118.0	21.6
Prefixdiffusion (Liu et al., 2024a)	31.8	26.6	56.1	109.3	20.4
SeDDM (Ours)	35.2	**28.7**	57.7	**119.5**	**22.2**

针对大语言模型自提示优化在图像描述任务中的有效性，本研究评估了三种方案：（1）基于SeDDM生成描述的基线方法；（2）直接使用MLLM进行描述优化；（3）采用自提示优化（self-prompt）框架的方法。直接使用MLLM进行描述优化我们利用以下提示来进行优化图像描述（polishing this sentence.sentence: ” ”. make it coherence and contain more details.）实验结果如表 3所示，基线方法在所有指标上得分较低，这主要源于基线模型仅能生成单句描述。相比之下，直接使用大语言模型优化可使各项指标获得提升。在CLIPScore上提升了1.2，细节度提高了3.1，丰富度增加了0.24。而自提示优化框架则在所有评价维度上取得更优结果：在CLIPScore、细节度和丰富度上分别增加1.6、4.6和0.37。以上综合实验表明，该框架在描述丰富性和细节度方面表现突出，能生成更全面、细致的图像描述。同时，在图像文本相关性指标CLIPScore上也有一定的提升。这些结果综合说明了自提示方法通过利用MLLM的推理能力给图像描述任务提供了丰富和细粒度的外部知识，从而提高了MLLM的生成能力，使得优化润色效果得到一定的提升。

4.4 可视化对比

为清晰展示模型效果，我们在图 5中提供了八个样本的视觉对比结果。实验表明，本模型能够更精确地描述图像内容并生成更丰富的文本描述。具体而言：在第一幅图像中准确生成“in the air”的悬空状态空间描述。在第五幅图像中精确捕捉“a pink shirt ”粉色衬衫的细节特征。

表 3: 使用自提示框架的比较结果

	CLIPScore↑	细节度↑	丰富度↑
SeDDM	30.8	3.6	0.31
MLLM w/o. self-prompt	32.0	6.7	0.55
MLLM w/. self-prompt	**32.4**	**8.2**	**0.68**

第二十四届中国计算语言学大会论文集，第434页-第447页，济南，中国，2025年8月11日至14日。

在第七幅图像中正确识别“watching a projection screen”观看投影屏幕的行为活动。这些结果证明，SeDDM能够有效提取图像中的深层语义信息，从而生成质量更高的文本描述。

DDCap : A man doing a trick on a motor bike.
ours : A person on a motorcycle doing a trick in the air.
GT : A person up in the air with a motor bike.

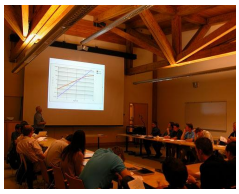
DDCap : A dog is jumping up to catch a frisbee.
ours : A group of people watching a dog play frisbee.
GT :A dog and its trainer doing Frisbee tricks for a crowd.

DDCap : A train is going over a bridge over a river.
ours : A train traveling across a bridge over a river.
GT : A train traveling on top of a bridge spanning a river.

DDCap : A man walking across a street under a traffic light.
ours : A man walking across a rain soaked street holding an umbrella.
GT : A person walking across a street in the rain.

DDCap : A woman holding a box of pizza in a box.
ours : A woman in a pink shirt holding a box of pizza.
GT : A woman is holding a box of pizza.

DDCap : A small dog sitting on a bed near a window.
ours : A dog sitting on a cushion looking out a window.
GT : A dog that is looking out of a window.



DDCap : A group of people sitting around a table.
ours : A group of people sitting around a table watching a projection screen.
GT : A group of people at long tables are looking at a big screen.

DDCap : A teddy bear is wrapped up in a bed.
ours : A teddy bear laying in a bed with a blanket.
GT : A teddy bear tucked into bed under a blanket.

图 5: SeDDM与相近工作DDCap的生成结果可视化对比

图 6展示了自提示优化框架的生成结果可视化分析。图中结果表明，该框架在细粒度图像细节捕捉方面表现优异，并能有效缓解多模态大语言模型(MLLM)产生的描述偏差。具体而言，在第二个示例中，MLLM生成的“the top of the hydrant”这一表述与图像中的物体及其空间语境存在偏差；相比之下，本框架生成的描述聚焦于场景光照效果（如“illuminated by a light”），避免了缺乏视觉依据的空间关系描述。同时，该框架能够有效抑制大语言模型常见的错误性描述。例如在第五个示例中，MLLM生成的“a group of people are enjoying a kite-flying session”缺乏视觉证据支持，属于典型的错误内容；而本框架则准确描述了空中飞舞的风筝及其环境氛围特征。因此，本文所提出的自提示框架能够改善了MLLM的部分缺陷，在文本润色方面展现出良好的优化效果。

4.5 消融实验

为了验证语义感知离散扩散模型中各模块的有效性，本研究在COCO数据集上进行了系统的消融实验。结果如表 4所示。实验采用三组配置进行对比分析：基础Base指的是仅使用标准扩散Transformer架构直接进行离散扩散去噪的图像描述生成；Base加上SF表示在基础架构上增加了语义融合模块，通过线性层将图像区域特征映射到文本空间的维度进行特征融合后进行离散扩散去噪生成的图像文本描述；Base加上SF和SP则整合了语义融合模块和语义感知模块，构成完整的SeDDM模型。实验结果表明，采用语义融合模块的离散扩散模型相比基础模型在性能指标上获得了一定提升，这主要归因于该模块能够利用图像区域特征进行特征融合，但由于缺乏对跨模态深层语义关系的建模，其性能提升存在局限。相比之下，同时集成语义感知和语义融合模块的完整模型在所有评估指标上均展现出显著优势。这一结果充分证明了语义感知模块在图像描述生成任务中的关键作用，该模块不仅优化了视觉与文本特征的对齐精度，还在去噪过程中有效建立了图像与文本之间的语义关联，从而显著提升了生成描述的语义准确性和表达质量。

图像	MLLM(LLAVA)	MLLM with self-prompt
	A large white dog is standing next to two people on a bench, one of whom is wearing a blue shirt. The dog appears to be sitting on the bench, and the people are sitting nearby.	A large white dog is sitting on a bench next to two people, who are accompanied by a backpack and handbag. In the background, there is a bicycle and trees, creating a serene and leisurely atmosphere.
	A yellow fire hydrant is situated next to a building at night, with a blue light streak coming from it. The hydrant is located on the sidewalk, and the blue light appears to be emanating from the top of the hydrant.	A yellow and silver fire hydrant stands next to a building at night, with a tree nearby. The scene is illuminated by a light, creating a mysterious atmosphere.
	A herd of sheep grazing on a lush green field, with some standing in the foreground and others scattered throughout the scene.	A herd of sheep grazing on a lush green field, surrounded by dirt, water, and trees. The sheep are peacefully enjoying the serene natural environment, with mountains in the background.
	A young boy is standing behind a stuffed animal, which is sitting on a counter. The boy appears to be smiling, and the stuffed animal is wearing a name tag.	A young boy is smiling in the background, while a stuffed animal cat with a note that says "Kitty Bear Award" sits on a counter. The scene is playful and happy, with the boy enjoying the company of his beloved stuffed toy.
	On a sunny day at the beach, a group of people are enjoying a kite-flying session. There are at least five kites soaring in the sky, each with different colors and designs. The kites are spread out across the sky, with some flying higher and others closer to the ground. The scene is lively and full of excitement as the kites dance in the wind, creating a vibrant and memorable experience for everyone present.	The sky is a beautiful shade of blue, and several kites are flying gracefully in the air over a sandy beach. The kites come in various shapes and sizes, adding a lively and colorful touch to the scene. The atmosphere is calm and serene, making it an ideal day for kite flying and enjoying the outdoors.

图 6: 本图展示了自提示优化框架在图像描述任务中的生成结果。其中，红色标记表示存在错误的描述内容，绿色标记则代表经过优化后生成的多样化且丰富的描述文本

表 4: 在COCO数据集上进行消融实验结果，”Base”表示扩散Transformer，”SF”则代表语义融合模块，”SP”则代表语义感知模块。

Base	SF	SP	B@4↑	METEOR↑	ROUGE↑	CIDEr↑	SPICE↑
✓			34.5	28.4	57.6	118.1	21.8
✓	✓		34.9	28.5	57.5	118.9	21.9
✓	✓	✓	**35.2**	**28.7**	**57.7**	**119.5**	**22.2**

表 5: 人工评估得票率%

指标	MLLM	self-prompt MLLM
流畅性↑	41.2%	**58.8%**
多样性↑	35%	**65.0%**
连贯性↑	39.6%	**60.4%**
图文相似性↑	43.4%	**56.6%**

对于自提示框架的有效性验证如表 3所示，相较于直接优化方法，采用自提示优化框架CLIPScore提升了0.4，细节度提高了1.5，丰富度增加了0.13。这表明使用了自提示优化方法能够在原有多模态大语言模型的基础上，通过对图片当中相关信息的捕获，进一步提升其在图像描述任务上的生成能力。

4.6 人工评估

由于多模态大语言模型生成的结果缺少参考标准，实验同时也进行人工评估。具体评估方案如下：我们从测试集中随机抽取100张图像样本，邀请5名评估人员分别对多模态大语言模型（MLLM）直接优化生成与采用自提示方法优化后的描述结果进行对比评估。评估设置四个维度：语言流畅性、描述多样性、逻辑连贯性以及图文相似性，采用投票机制确定各指标下的最优描述。

根据表 5 所示实验结果的深入分析，自提示优化框架在获得流畅性投票率58.8%的同时,在描述多样性方面表现突出，获得了65.0%的投票率。同样的在连贯性和图文相似度也分别获得了60.4%、56.6%。这证明了该方法能有效激发模型的生成表达能力。同时根据综合评估结果也能证明，相较于传统的MLLM直接润色生成方式，我们提出的自提示优化框架能系统性地提升图像描述质量，在语言表达、内容丰富度等多个层面均取得显著改进。该结果与自动评估结论一致，进一步表明了自提示优化框架能有效提升图像描述质量，其生成效果优于大语言模型的直接润色生成。

5 总结与展望

本文提出了一种图像描述生成方法与优化框架，通过引入语义感知融合等模块构建了语义感知的离散扩散模型（SeDDM），该模型在去噪过程中有效提升了图文相关性的学习能力，缓解了传统离散扩散模型在非自回归描述任务中存在的的相关问题，从而生成更精细的图像描述。基于多模态大语言模型设计的自提示策略进一步优化了生成结果，显著提升了语言丰富度、视觉细节描述和图文相关性。在COCO数据集上的综合实验验证了该方法的有效性和优越性。证明了扩散模型在图像描述任务中的应用潜力，同时也展示了通过自提示策略实现多模态大语言模型零样本学习的有效途径。

与此同时，像COCO这类数据集的标注均采用单句描述图像内容，在当前多模态大语言模型迅速发展的背景下，此类简略的标注难以满足文本丰富性和表达多样性的需求，从而在一定程度上制约了模型生成高质量描述的能力。而大语言模型在许多多模态任务中已经展现出优异的性能，但也仍存在进一步提升的空间。未来的研究将致力于开发更为先进的优化策略，可以采用过滤机制、思维链和大模型代理等技术手段使多模态大语言模型能够自己思考和反馈从而进一步提升模型在图像描述生成任务中的表现，使其能够生成更加丰富、准确和生动的文本描述。

参考文献

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, pages 17981–17993.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. 2023. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *The Eleventh International Conference on Learning Representations*.

Qimin Cheng, Yuqi Xu, and Ziyang Huang. 2024. Vcc-diffnet: Visual conditional control diffusion network for remote sensing image captioning. *Remote Sensing*, 16(16).

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10584.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2022. Injecting semantic concepts into end-to-end image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17988–17998.

Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. Masked non-autoregressive image captioning. *arXiv: 1906.00717*.

Longteng Guo, Jing Liu, Xinxin Zhu, Xingjian He, Jie Jiang, and Hanqing Lu. 2021. Non-autoregressive image captioning with counterfactuals-critical multi-agent learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Yufeng He, Zefan Cai, Xu Gan, and Baobao Chang. 2023a. Diffcap: Exploring continuous diffusion on image captioning. In *arXiv: 2305.12144*.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023b. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4521–4534.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: transforming objects into words. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *IEEE/CVF International Conference on Computer Vision*, pages 2951–2963.

第二十四届中国计算语言学大会论文集，第434页-第447页，济南，中国，2025年8月11日至14日。

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *IEEE/CVF International Conference on Computer Vision*, pages 4633–4642.

Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. How good are low-bit quantized llama3 models? an empirical study. *arXiv:2404.14047v1*.

Frank Hutter Ilya Loshchilov. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018. Recurrent fusion network for image captioning. In *15th European Conference on Computer Vision*, page 510–526.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. 2022. Comprehending and ordering semantics for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Guisheng Liu, Yi Li, Zhengcong Fei, Haiyan Fu, Xiangyang Luo, and Yanqing Guo. 2024a. Prefix-diffusion: A lightweight diffusion model for diverse image captioning. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12954–12965.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. 2023. Semantic-conditional diffusion networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23359–23368.

Munan Ning, Yujia Xie, Dongdong Chen, Zeyin Song, Lu Yuan, Yonghong Tian, Qixiang Ye, and Liuliang Yuan. 2023. Album storytelling with iterative story-aware captioning and large language models. *ArXiv*, abs/2305.12943.

Long Ouyang, Jeffrey Wu, Xu Jiang, and et al. 2022. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Yixuan Su, Deng Cai, Yan Wang, David Vandyke, Simon Baker, Piji Li, and Nigel Collier. 2021. Non-autoregressive text generation with pre-trained language models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, April.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. In *arXiv: 2303.04671*.

Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. 2021. Semi-autoregressive image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 2708–2716.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *The European Conference on Computer Vision*.

Hong Yu, Yuanqiu Liu, Baokun Qi, Zhaolong Hu, and Han Liu. 2023. End-to-end non-autoregressive image captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.

Yu Zhao, Hao Fei, Xiangtai Li, Libo Qin, Jiayi Ji, Hongyuan Zhu, Meishan Zhang, Min Zhang, and Jianguo Wei. 2024. Synergistic dual spatial-aware generation of image-to-text and text-to-image. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 122354–122383. Curran Associates, Inc.

Yuanen Zhou, Yong Zhang, Zhenzhen Hu, and Meng Wang. 2021. Semi-autoregressive transformer for image captioning. In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 3132–3136.

Kun Zhou, Yifan Li, Xin Zhao, and Ji-Rong Wen. 2024. Diffusion-NAT: Self-prompting discrete diffusion for non-autoregressive text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1438–1451.

Zixin Zhu, Yixuan Wei, Jianfeng Wang, Zhe Gan, Zheng Zhang, Le Wang, Gang Hua, Lijuan Wang, Zicheng Liu, and Han Hu. 2022. Exploring discrete diffusion models for image captioning. In *arXiv: 2211.11694*.