GIIFT: 图引导的归纳无图像多模态机器翻译

Jiafeng Xiong
Department of Computer Science
University of Manchester
jiafeng.xiong@manchester.ac.uk

Yuting Zhao
Department of Advanced Information Technology
Kyushu University
zhao.yuting.095@m.kyushu-u.ac.jp

Abstract

多模态机器翻译(MMT)展示了视觉信息对机器翻译的重要帮助。然而，现有的 MMT 方法在利用模态差异时面临挑战，因为它们通过强制严格的视-语对齐同时局限于其训练的多模态域内进行推理。在这项工作中，我们构建了新的多模态场景图，以保留和整合特定模态的信息，并介绍了 GIIFT，这是一个两阶段的 G 图引导 I 归纳 I 图像 F 无 MM T 框架，该框架使用跨模态图注意力网络适配器在统一的融合空间中学习多模态知识，并将其归纳泛化到更广泛的无图像翻译领域。在英语到法语和英语到德语任务的 Multi30K 数据集上的实验结果表明，我们的 GIIFT 超越了现有方法，并实现了最先进的水平，即使在推理过程中没有使用图像。在 WMT 基准上的结果显示，与无图像翻译基线相比有显著的提升，证明了 GIIFT 在归纳无图像推理方面的实力。

1 介绍

多模态机器翻译（MMT）旨在通过融入多模态数据，特别是视觉输入，以改进传统的仅文本神经机器翻译（NMT）。现有的方法大多集中于强制图像与文本之间的对齐来改进 MMT，并已证明受益于对齐的视觉信息在消除文本歧义方面的效果。然而，在训练和推理阶段同时具有对齐形式的多模态数据集，使得 MMT 模型在正常的纯文本机器翻译环境中无法进一步泛化。尽管图像的丰富信息可以带来超越文本层面的翻译益处，当对齐的图像信息成为翻译过程推理的必需时，MMT 模型的应用将受到严重限制。因此，无法摆脱推理阶段对齐图像的需求是限制 MMT 模型灵活应用的关键瓶颈。

为了解决上述瓶颈，已经有几种方法试图探索在多模态翻译（MMT）模型中实现无图像推断的解决方案。例如，在训练期间从文本或文本场景图中合成视觉幻觉用于无图像推断（Li et al., 2022; Fei et al., 2023）；从图像到文本标题任务迁移学习到文本到文本翻译任务（Gupta et al., 2023）。然而，这些努力仍未能始终如一地达到彻底弥合多模态与无图像推断之间差距的性能。在推进 MMT 中的无图像推断方面仍面临关键挑战。

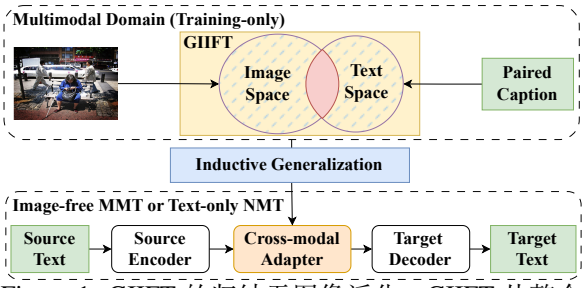

Figure 1: GIIFT 的归纳无图像泛化。GIIFT 从整个多模态领域进行归纳学习，然后通过跨模态泛化来实现 MMT 或仅文本 NMT 的无图像推理。相比之下，以往的模型只能从图像和文本的有限重合部分进行学习。

首先，先前的模型学习不充分，是因为它们强制对齐模态间的差距，通常是指图像和文本之间的内在信息不平衡（Schrodi et al., 2025）。考虑图 1 中的情况，这种约束限制了从多模态域中的红色重叠部分推广出的无图像推理能力。然而，最近的研究（Ramasinghe et al., 2024; Schrodi et al., 2025）显示，接受这种差距并利用全部信息（图 1 中的整个集合）显著提升了多模态学习。其次，目前无图像的多模态翻译（MMT）方法，包括图像幻觉或检索，未能实现跨模态泛化。与 MMT（Fei et al., 2023）相比，无图像翻译性能显著下降，而通过对齐模态间的差距导致的附加信息损失进一步加剧了问题。第三，几乎所有的 MMT 模型都是直推式的（Sutskever et al., 2014b），这使得 MMT 模型无法从多模态域归纳地推广至仅文本的更广泛应用域。面对这些挑战，我们在本研究中探讨以下研究问题：RQ1：我们如何能够充分表示不同的模态以接受模态间的差距？RQ2：我们能否通过跨模态泛化有效地进行无图像推理而不降低性能？RQ3：MMT 是否具备将多模态域推广至更广泛的仅文本域的归纳能力？

我们提出了 GIIFT，这是一种两阶段的 G 图

引导 I 归纳 I 图像- F 免费 MM T 框架。如图 1 所示，GIIFT 在第一阶段从整个多模态领域中学习，并旨在在第二阶段实现图像无关 MMT 或纯文本 NMT 的归纳泛化。提出了图结构的新颖多模态场景图 (MSG) 和语言场景图 (LSG)，用于表示多模态领域，其中每种模态通过统一空间中的图表示来通知和丰富另一种模态。具体来说，我们从图像中提取视觉关系，从文本中提取语言关系，然后通过全局超级节点保留和整合它们以构建 MSGs。LSG 是语言版本，通过全局超级节点保留语言关系。这些关系和节点的特征通过 M-CLIP (Carlsson et al., 2022) 得到互相丰富和统一初始化。

为了实现跨模态推广到无图像或更广泛的纯文本领域，GIIFT 被设计为通过跨模态 GAT 适配器在第一阶段通过 MSG 从多模态领域归纳学习多模态知识，并在第二阶段通过 LSG 将其推广到无图像推理，基于可替换的骨干网络，mBART (Liu et al., 2020) 。

总体而言，主要贡献包括：
(i) 我们构建了新的 MSGs 和 LSGs，以在统一空间中充分表示不同模态，从而弥合模态差距。
(ii) 我们提出了 GIIFT 框架，通过一个新的跨模态 GAT 适配器，从 MSGs 或 LSGs 的两阶段连续学习中实现无图像推理的归纳泛化。
(iii) 在 En → { Fr, De } Multi30K 上的实验结果表明，即使在推理时没有图像，GIIFT 的表现也优于大多数现有的 MMT 方法。在 En → { Fr, De } WMT (Bojar et al., 2014) 上，GIIFT 相较于最佳无图像基准 CLIPTrans，平均提升 +1.92(8.00 %) BLEU 和 +2.82(4.80 %) METEOR，表明从多模态 Multi30K 到其他仅文本 NMT 领域的有效归纳。
(iv) 进一步分析表明，所提出的方法可以通过 MSGs 有效弥合模态差距，并通过跨模态泛化实现稳健的无图像推理。GIIFT 在多模态推理下与自身整体持平，表现出在充分弥合多模态推理与无图像推理之间差距的一致性能。

在这项工作中，我们构建了统一的 MSGs 来弥合差距，并通过 LSGs 来泛化图像无关推理的知识。我们进一步设计了一个两阶段的 GIIFT 框架，包括一个用于归纳学习和泛化的跨模态归纳 GAT 适配器。我们的方法结构如下：1）介绍我们的归纳图像无关推理的问题定义（小节 2.1）。2）MSG 和 LSG 场景图生成的细节（小节 2.2）。3）GIIFT 框架的描述（小节 2.3）。

设 $\mathcal{D}_m$ 是一个包含三元组 (i, c_s, c_t) 的多模态多语言数据集，其中 i 是图像，(c_s, c_t) 是其源语言和目标语言的说明，同时设 $\mathcal{D}_l$ 为一对一的文本平行语料对集 (t_s, t_t) 。传统的 MMT 方法在训练过程中将 i 与 c_s 对齐，然后基于这种对齐关系进行无图像推断，但视觉知识仍然与 c_s 和 $\mathcal{D}_m$ 相关联。我们引入的无图像推断方法则从 i 和 (c_s, c_t) 中学习多模态知识，并在 $\mathcal{D}_m$ 中跨模态地泛化到 $\mathcal{D}_m$ 或翻译对 $(t_s, t_t) \in \mathcal{D}_l$ 。

为了在数据预处理过程中保持模态特定信息并提取复杂关系（例如，人和物体的空间关系和事件状态），我们使用多模态大语言模型（LLM）作为图像解析器，并使用现成的文本解析器分别构建 MSG 和 LSG。图 **??** 显示 MSG 包括一个视觉超级节点、一个图像场景图（ISG）和一个文本场景图（TSG），而 LSG 则用文本超级节点替换视觉超级节点并省略 ISG。

（1）图像场景图。通过使用 LLaVA-34B (Liu et al., 2023) 从图像 i 中获得 ISG。为了获得连贯且结构良好的输出，我们将采样温度设置为 0 以进行确定性生成，对于需要探索性输出的更具挑战性的情况将其提高到 0.4。我们的提示（详细信息见附录 A ）包括：(i) 任务描述，指定如何形成关系并生成结构化场景图内容；(ii) 负面示例，说明输出格式中的常见错误以及如何纠正它们；(iii) 格式示例，提供抽象但结构良好的场景图模板而不使用具体对象，从而防止提示污染。因此，我们生成包含独特和关系型视觉信息的 ISG，例如事件状态。
（2）文本场景图。通过使用 FACTUAL (Li et al., 2023) 从文本 c_s 或 t_s 解析 TSG。FACTUAL 比大型语言模型对于大规模语料库而言更轻量和高效。TSG 将实体及其关系编码为 ISG 的文本类比。因此，我们获得具有结构化语言关系和独特语义信息的 TSG。
（3）多模态场景图。对于每对 (i, c_s) ，我们通过引入超级节点，将 ISG 和 TSG 合并为 MSG，该节点编码来自 M-CLIP 图像编码器的整体图像嵌入。超级节点连接到所有普通的 ISG 和 TSG 节点，以联合模态特定信息并提供多样化的细粒度信息，作为接受模态差距并建立多模态关系的重要桥梁。我们通过 M-CLIP 文本编码器嵌入所有普通节点和关系特征，从而实现多模态关系和归纳基础的统一表示。
(4) 语言场景图。为了跨模态的泛化，我们为文本对 (t_s, t_t) 构建了语言场景图（LSG），通过 M-CLIP 文本编码器仅保留具有超级节点的文本场景图（TSG），该超级节点表示整个文本嵌入，并与 MSG 共享统一的隐藏空间。同样，我们使用 M-CLIP 文本编码器嵌入所有普通节点和边。超级节点连接到所有普通的 TSG 节点，使得在无图像推理时能够利用多层次的文本信息。

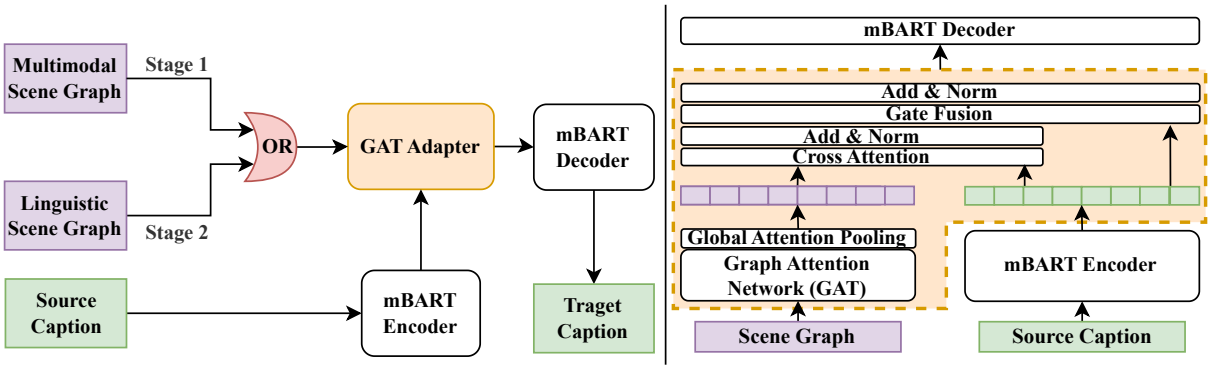

Figure 2: 左图：两阶段 GIIFT 框架概述。阶段 1：通过 MSGs 的多模态学习。阶段 2：通过 LSGs 的跨模态泛化。右图：跨模态 GAT 适配器的架构概述，它用于归纳学习并融合多模态知识到骨干架构 mBART 中。

1.1 GIIFT 框架

GIIFT 利用场景图作为模态桥，通过一个基本的跨模态 GAT 适配器来引导主干 mBART 在两阶段训练管道中，分别归纳地学习多模态知识和进行图像无关的翻译，使用 MSG 和 LSG。

1.1.1 两阶段训练框架

阶段 1：通过 MSG 实现多模态学习。如图 2 (左) 所示，阶段 1 在配对的图像和标题的 MSG 上训练一个共享的 GAT 适配器，导入引导 mBART 翻译的多模态知识。在 MSG 中，$\mathcal{N}_i^{\text{SG}}$ 是来自 ISG 或 TSG 的普通节点 i 的邻居，反映了来自图像或文本的结构关系。GAT 中第 $l, (0 \le l \le L)$ 层节点 i 的嵌入 $\boldsymbol{Z}_i^{(l)}$ 的递归计算如下：

$$\boldsymbol{Z}_i^{(l)} = \sigma(\sum_{j \in \mathcal{N}_i^{\text{SG}}} \alpha_{ij}^{(l)} \phi(\boldsymbol{W}[\boldsymbol{Z}_i^{(l-1)} \| \boldsymbol{Z}_j^{(l-1)} \| \boldsymbol{E}_{ij}]) + \alpha_{i,\text{SN}}^{(l)} \phi(\boldsymbol{W}[\boldsymbol{Z}_i^{(l-1)} \| \boldsymbol{Z}_{\text{SN}}^{(l-1)} \| \boldsymbol{E}_{i,\text{SN}}])), \quad (1)$$

在这里，$\boldsymbol{Z}_i^{(l-1)}$ 是在前一层节点 i 的嵌入（初始节点嵌入 $\boldsymbol{Z}_i^{(0)}$ 通过 M-CLIP 文本编码器），$\boldsymbol{Z}_{\text{SN}}^{(l-1)}$ 是来自超节点在第 $l-1$ 层的全局多模态嵌入，它被传递到所有普通节点，$\boldsymbol{E}_{ij}$ 或 $\boldsymbol{E}_{i,\text{SN}}$ 表示通过 M-CLIP 文本编码器在场景图中关系的边缘嵌入，$\alpha_{ij}^{(l)}$ 和 $\alpha_{i,\text{SN}}^{(l)}$ 是注意力权重，$\sigma(\cdot)$ 是一个激活函数，$\phi(\cdot)$ 是 LeakyReLU。从公式 (1)，我们可以观察到共享权重 $\boldsymbol{W}$ 通过联合处理来自 ISG 或 TSG 节点的局部文本嵌入（$\mathcal{N}_i^{\text{SG}}$）和来自超节点的全局多模态上下文，实现多粒度多模态关系的学习。这个共享权重 $\boldsymbol{W}$ 对于捕捉这些多模态关系至关重要，并将在阶段 2 的跨模态泛化过程中被利用。

初始 MSG 超节点提供全局图像嵌入 $\boldsymbol{Z}_{\text{SN}}^{\text{MSG}(0)}$，并从所有普通节点聚合信息，如下所示：

$$\boldsymbol{Z}_{\text{SN}}^{\text{MSG}(l)} = \sigma(\sum_{i \in \mathcal{N}_{\text{SN}}^{\text{MSG}}} \alpha_{\text{SN},i}^{(l)} \phi(\boldsymbol{W}[\boldsymbol{Z}_{\text{SN}}^{(l-1)} \| \boldsymbol{Z}_i^{(l-1)} \| \boldsymbol{E}_{\text{SN},i}])) \quad (2)$$

其中 $\mathcal{N}_{\text{SN}}^{\text{MSG}}$ 包含 MSG 中的所有普通节点。

通过 GAT 适配器中的全局注意池 (Li et al., 2015) 层，我们接着获得多模态图表示 $\boldsymbol{Z}_g^{\text{MSG}} \in \mathcal{D}_m$：

$$\boldsymbol{Z}_g^{\text{MSG}} = \text{AttnPool}(\{\boldsymbol{Z}_i^{\text{MSG}} : i \in \mathcal{V}_{\text{MSG}}\}), \quad (3)$$

其中 $\mathcal{V}_{\text{MSG}}$ 表示 MSG 中的节点集合，包括普通节点和超级节点。

然后，$\boldsymbol{Z}_g^{\text{MSG}}$ 被输入 mBART 解码器进行翻译生成。mBART 编码器保持冻结状态以维持稳健的输出，这样解码器可以学习一个平衡的表示来基于多模态表示 $\boldsymbol{Z}_g^{\text{MSG}}$ 和 mBART 嵌入 $\boldsymbol{H}$ 之间的门机制（见公式 (??)）生成翻译。

阶段 2: 通过 LSG 的跨模态泛化。如图 2 (左) 所示，阶段 2 输入与文本构建的 LSG 相同的 GAT 适配器，允许多模态知识跨模态地泛化到更广泛的无图像领域。在 LSG 中，每个普通节点 i 表示一个具有初始嵌入 $\boldsymbol{Z}_i^{(0)}$ 的文本实体，并且是通过 M-CLIP 文本编码器生成的全局文本嵌入的超级节点 $\boldsymbol{Z}_{\text{SN}}^{\text{LSG}(0)}$。普通节点在层 l 的节点 i 的嵌入 $\boldsymbol{Z}_i^{(l)}$ 为：

$$\boldsymbol{Z}_i^{(l)} = \sigma(\sum_{j \in \mathcal{N}_i^{\text{LSG}}} \alpha_{ij}^{(l)} \phi(\boldsymbol{W}[\boldsymbol{Z}_i^{(l-1)} \| \boldsymbol{Z}_j^{(l-1)} \| \boldsymbol{E}_{ij}]) + \alpha_{i,\text{SN}}^{(l)} \phi(\boldsymbol{W}[\boldsymbol{Z}_i^{(l-1)} \| \boldsymbol{Z}_{\text{SN}}^{\text{LSG}(l-1)} \| \boldsymbol{E}_{i,\text{SN}}])), \quad (4)$$

LSG 超级节点更新为：

$$\boldsymbol{Z}_{\text{SN}}^{\text{LSG}(l)} = \sigma(\sum_{i \in \mathcal{N}_{\text{SN}}^{\text{LSG}}} \alpha_{\text{SN},i}^{(l)} \phi(\boldsymbol{W}[\boldsymbol{Z}_{\text{SN}}^{(l-1)} \| \boldsymbol{Z}_i^{(l-1)} \| \boldsymbol{E}_{\text{SN},i}])), \quad (5)$$

其中 $\mathcal{N}_{\text{SN}}^{\text{LSG}}$ 表示所有 LSG 普通节点。LSG 有助于将共享的多模态知识权重 $\boldsymbol{W}$ 从阶段 1 中的 $\mathcal{D}_m$ 泛化到阶段 2 中无图像域的 $\mathcal{D}_l$。

与方程 (3) 类似，我们得到了 LSG $\boldsymbol{Z}_g^{\mathrm{LSG}}$ 的图表示。mBART 解码器通过来自 $\mathcal{D}_m$ 的广义知识 $\boldsymbol{Z}_g^{\mathrm{LSG}}$ 得到增强，并适应于不含图像的领域 $\mathcal{D}_l$ ，同时保留未冻结的 mBART 编码器隐藏状态。

我们采用具有残差连接的多层 GAT 来学习多模态知识，并将其跨模态推广到更广泛的无图像领域。图（右）展示了 GAT 适配器的架构，该适配器融合了 mBART 编码器的输出并增强了 mBART 解码器。我们将通过全局注意力池记为 MSG 或 LSG 的图表示记为 $\boldsymbol{Z}_g$ 。图表示 $\boldsymbol{Z}_g$ 与 mBART 编码器隐藏状态 $\boldsymbol{H}$ 之间的融合输出 $\boldsymbol{O}$ 通过交叉注意力执行如下：

一个门机制 $\boldsymbol{g}$ 平衡嵌入流：

其中 $\odot$ 表示元素级乘法，$[\cdot\|\cdot]$ 表示拼接，$\boldsymbol{H}'$ 为 mBART 解码器的输入。这个两阶段过程展示了跨模态 GAT 适配器在表示和推广结构化多模态关系中的中心归纳作用。

2 实验

数据集。我们在两个基准上进行实验：Multi30K (Elliott et al., 2016) 和 WMT2014 (Bojar et al., 2014) 。Multi30K 是一个广泛使用的 MMT 基准，作为 Flickr30k 的多语言扩展。WMT 是一个仅使用文本的多语言 NMT 数据集。为了评估，我们在 Multi30K 的三个标准测试集上进行 EN → { DE, FR } 的翻译任务：Test2016、Test2017 和 MSCOCO。我们还训练和测试 EN → { DE, FR } 在 WMT 上使用来自 Multi30K 的多模态知识，以评估归纳的无图像推断。我们下采样 WMT 训练集以匹配 Multi30K 的大小，同时保持验证和测试集不变。

实现细节。我们在 A100 GPU 上训练 GIIFT，使用 AdamW 优化器（多项式衰减）。批处理大小为 64，学习率为 $2e^{-5}$ （阶段 1）和 $1e^{-5}$ （阶段 2）。GAT Adapter 有 9 层，与 M-CLIP 具有相同的 1024 维度。文本解码使用大小为 5 的束搜索。实现基于 PyTorch 和 Huggingface Transformers 库。我们分别通过 SacreBLEU (Post, 2018) 和 evaluate 库 (Banerjee and Lavie, 2005) 报告 BLEU (Papineni et al., 2001) 和 METEOR。结果是三次运行的平均值，BLEU 的提前停止耐心设为 5。所有表格将最好结果加粗，次佳结果加下划线。基线数据来自论文或代码库。

2.1 Multi30K 上的结果

表 1 和 2 包含了在 EN → { DE, FR } Multi30K 上 BLUE 和 METEOR 的翻译性能比较。我们观察到以下几点：

(1) 通过保留来自图像和文本的全部信息来吸收模态差距的有效性。无图像的 GIIFT 在最强基线 Soul-Mix（即使在有图像测试的情况下）上实现了新的最先进水平，平均提升 +0.61(1.37 %) BLEU 和 +1.01(1.55 %) METEOR，并超越了最佳基于场景图的无图像基线 UMMT，提升了 +13.525(42.27 %) BLEU 和 +18.865(36.07 %) METEOR。这些基线将字幕与图像对齐，导致 MMT 信息的显著丢失。

(2) 对于无图像推理的跨模态泛化的鲁棒性。GIIFT (ours) 模型通过跨模态泛化多模态知识用于无图像推理，GIIFT (ours$^{\#}$) 采用多模态推理，在 6 项基准中有 5 项获得前两名并整体持平。GIIFT (ours) 甚至超过了 GIIFT (ours$^{\#}$)，平均提升 +0.21(0.63 %) BLEU 和 +0.1(0.17 %) METEOR。相比之下，UMMT 在没有图像的情况下显著下降，平均评分低于 UMMT$^{\#}$ 的 -5.4(-12.76 %) BLUE 和 -5.45(-8.53 %) METEOR（带多模态推理）。

2.2 在 WMT 上的结果

在 Tab. 1 中，CLIPTrans (Gupta et al., 2023) 优于其他无图像推理基线，因此我们采用它作为我们的主要基线，并使用其官方库进行实验。类似 GIIFT（我们的），这个基于 mBART 的双阶段模型在 Stage 1 中在 Multi30K 上训练，在 Stage 2 中在 WMT 上训练。

表 3 显示，GIIFT（我们的）在总体上取得了最高的 BLEU 和 METEOR 分数，同时在没有使用 Multi30K 中的图像进行训练的情况下显著超越了 GIIFT（没有阶段 1）这两个指标。这验证了 GIIFT 的归纳能力，可以通过跨模态泛化实现稳健的无图像推理。相比之下，CLIPTrans 必须在阶段 1 中为图像生成字幕，然后在阶段 2 中通过迁移学习进行对齐的字幕翻译。因此，它的完整模型在无图像推理时受到对齐的阻碍，阻止了其学到的 Multi30K 视觉知识泛化到 WMT。此外，CLIPTrans（没有 阶段 1）优于两阶段的 CLIPTrans，而对于 EN → FR 翻译，其性能甚至下降到与其骨干模型 mBART 相当。

为了进一步验证 GIIFT 中不同组件的有效性，我们也展示了在 Tab. 4 中消融版本在 EN → { DE, FR } Multi30K 上的表现。完整模型 GIIFT （我们的）在所有基准测试中的指标都达到最高。移除门控机制 (GIIFT (w/o. gate)) 导致 BLEU 和 METEOR 分数显著下降，这突出门控机制在融合多模态知识和 mBART 信息中的关键平衡作用。此外，省略多模态学习阶段 (GIIFT (w/o. Stage 1)) 会导致性能下降，这强调学习可推广的多模态知识的重要性。GIIFT (unfrozen) 的性能明显低于 GIIFT（我们的），并更接近于基础模型 mBART。这强调

Model	EN → DE			EN → FR			Mean Δ
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
mBART (NMT backbone) (Liu et al., 2020)	41.12	36.63	32.89	63.37	57.01	47.28	-2.08
MMT Model with Multimodal Inference							
DCCN (Lin et al., 2020)	39.70	31.00	26.70	61.20	54.30	45.40	-5.41
GMNMT (Yin et al., 2020a)	39.80	32.20	28.70	60.90	53.90	-	-5.14
CAP-ALL (Li et al., 2021)	39.60	33.00	27.60	60.10	52.80	44.30	-5.56
Gated Fusion* (Wu et al., 2021)	42.00	33.60	29.00	61.70	54.80	44.90	-4.13
Gumbel-Attention (Liu et al., 2022)	39.20	31.40	26.90	-	-	-	-6.73
UMMT# (Fei et al., 2023)	37.40	-	-	56.90	-	-	-7.68
RG-MMT-EDC (Tayir et al., 2024a)	42.00	33.40	30.00	62.90	55.80	45.10	-3.60
Soul-Mix (Cheng et al., 2024)	44.24	37.14	34.26	64.75	57.47	49.25	-0.61
GIIFT (ours#)	43.32	37.47	34.66	65.17	59.11	49.76	-0.21
MMT Model with Image-free Inference							
ImagiT (Long et al., 2021)	38.50	32.10	28.70	59.70	52.40	45.30	-5.68
VALHALLA (Li et al., 2022)	41.90	34.00	30.30	62.20	55.10	45.70	-3.58
VALHALLA* (Li et al., 2022)	42.70	35.10	30.70	63.10	56.00	46.50	-2.78
UMMT (Fei et al., 2023)	32.00	-	-	50.60	-	-	-13.53
CLIPTrans (Gupta et al., 2023)	43.87	37.22	34.49	64.55	57.59	48.83	-0.7
GIIFT (ours)	44.04	38.41	34.94	65.61	58.05	49.72	

Table 1: Multi30K 上的 BLEU。Δ 是与 “GIIFT (ours)” 的差距。* 表示集成模型，# 表示用图像和文本进行训练和测试的模型。“GIIFT (ours#)” 在一个阶段中用未冻结的 mBART 进行训练。

Model	EN → DE			EN → FR			Mean Δ
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
mBART (NMT backbone) (Liu et al., 2020)	69.59	65.07	60.15	82.40	77.63	71.58	-1.35
MMT Model with Multimodal Inference							
DCCN (Lin et al., 2020)	56.80	49.90	45.70	76.40	70.30	65.00	-11.73
GMNMT (Yin et al., 2020a)	57.60	51.90	47.60	74.90	68.60	62.60	-11.88
CAP-ALL (Li et al., 2021)	57.50	52.20	46.40	74.30	68.60	62.60	-12.15
Gated Fusion* (Wu et al., 2021)	67.80	61.90	56.10	81.00	76.30	70.50	-3.48
Gumbel-Attention (Liu et al., 2022)	57.80	51.20	46.00	-	-	-	-14.72
UMMT# (Fei et al., 2023)	57.20	-	-	70.70	-	-	-13.42
RG-MMT-EDC (Tayir et al., 2024a)	60.20	53.70	49.60	77.20	72.00	64.90	-9.48
Soul-Mix (Cheng et al., 2024)	69.93	63.59	59.94	83.24	78.23	73.48	-1.01
GIIFT (ours#)	70.65	65.59	61.37	83.32	78.95	73.98	-0.10
MMT Model with Image-free Inference							
ImagiT (Long et al., 2021)	55.70	52.40	48.80	74.00	68.30	65.00	-11.72
VALHALLA (Li et al., 2022)	68.80	62.50	57.00	81.40	76.40	70.90	-2.92
VALHALLA* (Li et al., 2022)	69.30	62.80	57.50	81.80	77.10	71.40	-2.43
UMMT (Fei et al., 2023)	52.30	-	-	64.70	-	-	-18.87
CLIPTrans (Gupta et al., 2023)	70.22	65.43	61.26	82.48	77.82	72.78	-0.75
GIIFT (ours)	71.08	65.88	61.66	83.65	78.36	73.86	

Table 2: Multi30K 上的 METEOR。Δ 表示与 “GIIFT (ours)” 相比的差距。* 表示集成模型，# 表示使用图像和文本进行训练和测试的模型。“GIIFT (ours#)” 在一个阶段内使用未冻结的 mBART 进行训练。

Model	EN → DE		EN → FR	
	BLEU	METEOR	BLEU	METEOR
mBART (backbone) (Liu et al., 2020)	15.58	41.18	26.50	52.06
CLIPTrans (Gupta et al., 2023)	16.63	42.13	26.78	51.76
(w/o. Stage 1)	17.60	42.81	27.71	53.38
GIIFT (ours)	18.10	43.88	28.70	54.58
(w/o. Stage 1)	17.79	43.01	27.89	53.45

Table 3: 在仅包含文本的 WMT 数据上的领域泛化比较。“(无阶段 1)” 表示模型在没有 Multi30K 图像参与的情况下训练。

了在阶段 1 中冻结 mBART 编码器以保持其稳定嵌入的必要性。NMT 骨架模型 mBART 在 BLEU 和 METEOR 上表现出显著下降，体现了 GIIFT 框架在学习多模态知识和跨模态泛化以进行无图像推理方面的有效性。

3 案例研究

为了研究第二阶段中 (1) 仅文本的图引导泛化的具体优势和 (2) 接受模态差距，我们在图 3 中比较几个案例：完整模型；GIIFT (我们的)，它通过 LSGs 从 MSGs 中学习多模态知识以进行跨模态无图像泛化 (类似于图 3 中的 TSGs)；GIIFT (没有阶段 1)，通过具有结构的语言关系进行 LSGs 翻译；以及仅使用文本的 NMT 主干 mBART。

(1) LSGs 引导 GIIFT 在第二阶段将 MSGs 的多模态知识更好地泛化空间关系信息。在图 3 (左图) 中，GIIFT (我们的) 和 GIIFT (没有阶段 1) 都通过 MSG 或 LSG 正确地捕获了空间介词 “on”，尽管源英文标题省略了空间信息。但由于缺乏场景图，mBART 产生不精确的翻译，没有 “on”，这突出了 LSGs 用于引导学习和泛化空间关系知识的功能。

(2) MSGs 接受模态差距，并保留一般对齐或仅文本翻译常常忽略的信息，从而改善翻译。

(i) 环境上下文。在图 3 (左图) 中，环境特征 “dirt hill” 只有通过完整的 GIIFT (我们的) 利用来自 MSGs 的多模态知识才能被准确翻译。GIIFT (没有阶段 1) 和 mBART 尽管在源标题中有 “dirt”，但产生不精确的翻译，反映出对模态特定信息的忽视。

(ii) 时间状态。在图 3 (中间) 中，MSG 捕

Model	EN → DE						EN → FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
mBART (backbone)	41.12	69.59	36.63	65.07	32.89	60.15	63.37	82.40	57.01	77.63	47.28	71.58
GIIFT (w/o. Stage 1)	43.63	70.95	37.76	65.49	34.47	60.81	64.91	83.07	57.71	78.01	48.95	73.20
GIIFT (w/o. freezing)	42.84	70.37	37.24	65.35	34.38	60.69	63.74	82.62	56.52	77.23	48.92	72.54
GIIFT (w/o. gate)	43.50	70.95	37.96	65.11	33.85	60.21	64.14	82.61	57.58	78.00	48.59	72.90
GIIFT (ours)	44.04	71.08	38.41	65.88	34.94	61.66	65.61	83.65	58.05	78.36	49.72	73.86

Table 4: 在 Multi30K 上的消融研究。“GIIFT（无冻结）”在阶段 1 中拥有未冻结的 mBART 编码器。“GIIFT（无阶段 1）”是在没有图像的情况下训练的。“GIIFT（无门控）”是在没有门控融合的两阶段下进行训练的。

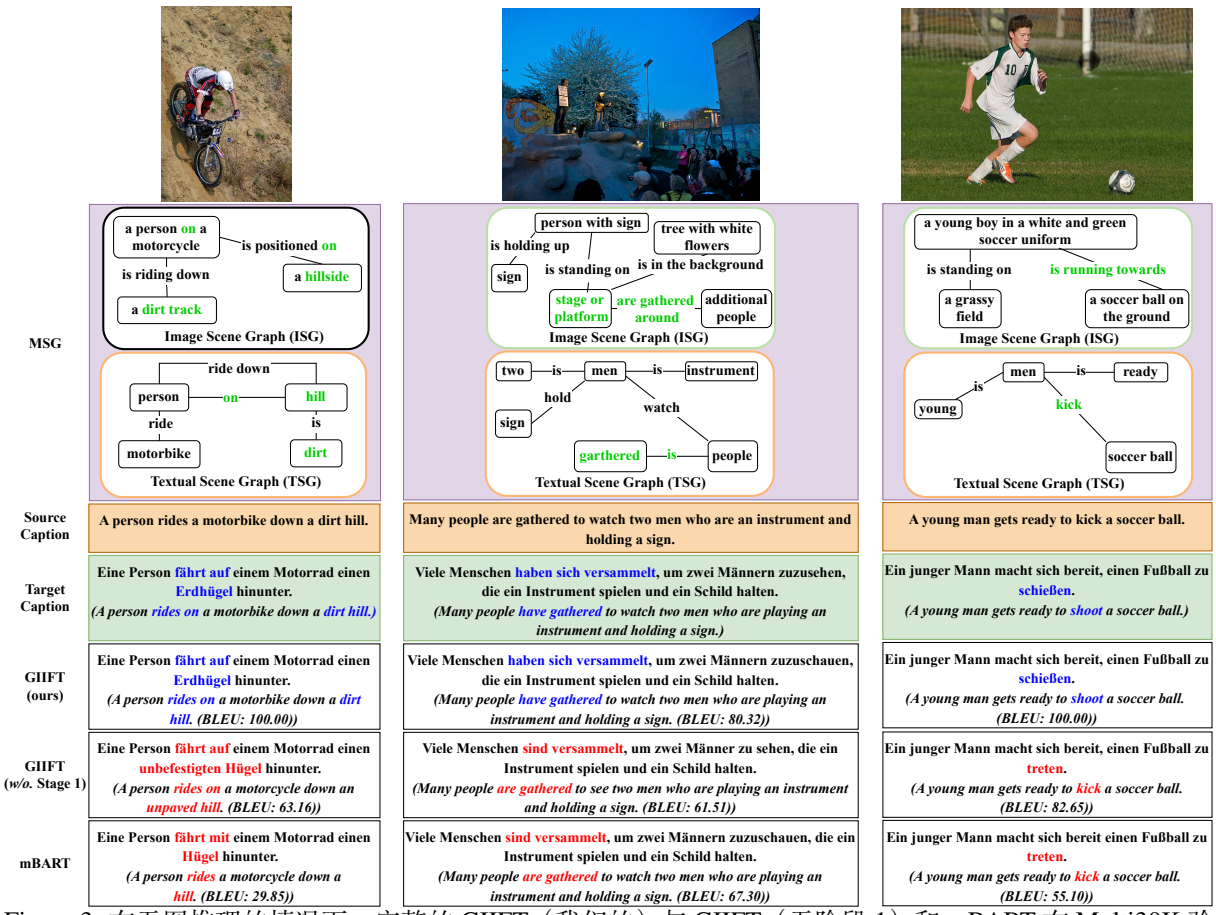

Figure 3: 在无图推理的情况下，完整的 GIIFT（我们的）与 GIIFT（无阶段 1）和 mBART 在 Multi30K 验证集上进行了比较。德文标题的斜体括号翻译用红色标记了差异。

捉了一个场景“聚集在舞台或平台周围”，使得完整的 GIIFT 能够识别其为一个完成状态，并在德语中生成适当的完成时态。相比之下，受到模型间隔限制的 GIIFT（不含 Stage 1）和 mBART 无法从对齐的视觉-语言空间中捕捉时间状态，从而消除了图像中独特的时间状态。因此，他们将英文中的“are gathered”直接翻译成德语的现在时。

(iii) 动作状态。图 3 (右) 展示了 MSG 的动作状态“正在向前跑”指导 GIIFT（我们的）正确地将动作翻译为“射击”而不是“踢”。仅通过 MSG 提供的视觉信息，使得从多模态知识中进行正确翻译成为可能。而 GIIFT（不含 Stage 1）和 mBART 却只能字面翻译为德语中的“踢”，这也显示了接受不同模态特定信息而不仅是对齐的重要性。来自 Test2016 的其他案例在附录 **??** 中。

4 相关工作

4.1 多模态机器翻译

MMT 研究汇集了视觉和文本信息用于机器翻译，并拥有越来越多的模型 (Specia et al., 2016; Li et al., 2021; Grönroos et al., 2018; Huang et al., 2020; Tayir et al., 2024a; Cheng et al., 2024) 。早期的方法通常采用基于 RNN 的架构，并加强了注意力机制以整合全球或空间视觉特征 (Calixto et al., 2017) 。随着时间的推移，变体结构的 Transformer 很快取代了 RNN，引入了更紧密的跨模态融合，例如动态标记重加权 (Caglayan et al., 2018; Lin et al., 2020) 、门控机制 (Wu et al., 2021) 或通过图结构实现的

多粒度融合（Krishna et al., 2017; Wang et al., 2018; Yin et al., 2020b）。近期的研究利用了预训练资源，例如 CLIP（Gupta et al., 2023; Li et al., 2022）或 BERT（Li et al., 2020）。在广泛采用的编码器-解码器框架内，MMT 研究在表示和推理两个方面取得了进展：

(1) 视觉语言表示。大多数 MMT 模型强制执行严格的视觉语言对齐。去歧义工作（Ive et al., 2019; Futeral et al., 2023）链接每个文本标记与相符的图像区域以解决词汇歧义。UMMT（Fei et al., 2023）将每个虚构的视觉场景图节点与文本对应部分对齐。CLIPTrans（Gupta et al., 2023）在两个阶段顺序训练，即图像描述和相应的翻译，强制在两个阶段的对齐。这种对齐通过仅学习模态之间的重叠来丢弃模态特定的信息。

(2) 无图像推理。先前的多模态翻译（MMT）在测试时依赖于获取配对的图像。为了缓解这一限制，研究人员进行了三方面的工作。首先，基于检索的模型将索引的字幕图像库中的视觉特征提取到解码器中，以替代缺失的图片（Zhang et al., 2020）。其次，幻觉方法（Johnson et al., 2018; Li et al., 2022; Fei et al., 2023）从文本中合成视觉输入。第三，使用迁移学习来训练带有图像的神经机器翻译（NMT）模型（Gupta et al., 2023）。

4.2 图神经网络

GNN 在建模关系结构方面非常强大，通过利用消息传递机制（Wu et al., 2020; Liang et al., 2022），它迭代地聚合和更新节点的表示，结合来自其邻居的信息，捕捉局部和全局的关系模式。一些 GNN，例如图卷积网络（GCN）(Kipf and Welling, 2017) 及其变体，只能用于推断，而其他的，包括 GAT（Velikovi et al., 2018）和 GraphSAGE（Hamilton et al., 2017），也可以进行归纳学习（Battaglia et al., 2018），以处理以前未见过的节点（Zhou et al., 2020）。GNN 还使用层次化或全局池化技术（Ying et al., 2018; Lee et al., 2019; Li et al., 2015; Gao and Ji, 2019）来捕获子图级别或图级别的嵌入（Zhou et al., 2020）。

5 结论

这项工作介绍了 GIIFT，一个基于图引导的两阶段 MMT 框架，以及新颖的多模态和语言场景图。GIIFT 在 Multi30K 上的表现优于现有的 MMT 模型，即使在无图像推理的情况下也表现出色，展示了通过 MSGs 在统一空间中学习多模态知识的有效性，并通过 LSGs 实现跨模态泛化。其无图像性能依然稳健，与多模态推理对应物相匹配。在 WMT 上，GIIFT 也胜过仅文本的 NMT 骨干网和领先的无图像 MMT 基准，显示了将多模态知识扩展到更广泛的仅文本领域的有效引导。我们的案例研究突显了图引导翻译和弥合模态差距的优势。这些发现为在现实应用和场景中开发可更普遍适用的 MMT 系统的未来研究奠定了坚实的基础。

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. CoRR .

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, and 8 others. 2018. Relational inductive biases, deep learning, and graph networks. arXiv preprint .

Ondej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation , pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kyle Buettner and Adriana Kovashka. 2024. Quantifying the Gaps Between Translation and Native Perception in Training for Multimodal, Multilingual Retrieval. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , pages 5863–5870, Miami, Florida, USA. Association for Computational Linguistics.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC Submissions for WMT18 Multimodal Translation Task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers , pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and Multilingual CLIP. In Proceedings of the Thirteenth Language Resources and Evaluation Conference , pages 6848–6854, Marseille, France. European Language Resources Association.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. SoulMix: Enhancing Multimodal Machine Translation with Manifold Mixup. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 11283–11294, Bangkok, Thailand. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In Proceedings of the 5th Workshop on Vision and Language , pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 5980–5994, Toronto, Canada. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Hongyang Gao and Shuiwang Ji. 2019. Graph U-Nets. In International Conference on Machine Learning , pages 2083–2092.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD Submission to the WMT18 Multimodal Translation Task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers , pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. CLIPTrans: Transferring Visual Knowledge with Pre-trained Models for Multimodal Machine Translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision .

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems , NIPS'17, pages 1025–1035, Red Hook, NY, USA. Curran Associates Inc. Event-place: Long Beach, California, USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 8226–8237, Online. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling Translations with Visual Awareness. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image Generation from Scene Graphs. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 1219–1228, Salt Lake City, UT. IEEE.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint .

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision , 123(1):32–73.

Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-Attention Graph Pooling. In Proceedings of the 36th International Conference on Machine Learning .

Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. 2019. DeepGCNs: Can GCNs Go as Deep as CNNs? In The IEEE International Conference on Computer Vision (ICCV) .

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What Does BERT with Vision Look At? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 5265–5275, Online. Association for Computational Linguistics.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu (Richard) Chen, Rogerio Feris, David Cox, and Nuno Vasconcelos. 2022. VALHALLA: Visual Hallucination for Machine Translation. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) .

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated Graph Sequence Neural Networks. arXiv preprint . Version Number: 4.

Zhifeng Li, Yu Hong, Yuchen Pan, Jian Tang, Jianmin Yao, and Guodong Zhou. 2021. Feature-level Incongruence Reduction for Multimodal Translation. In Proceedings of the Second Workshop on Advances in Language and Vision Research , pages 1–10, Online. Association for Computational Linguistics.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023. FACTUAL: A Benchmark for Faithful and Consistent Textual Scene Graph Parsing. In Findings of the Association for Computational Linguistics: ACL 2023 , pages 6377–6390, Toronto, Canada. Association for Computational Linguistics.

Fan Liang, Cheng Qian, Wei Yu, David Griffith, and Nada Golmie. 2022. Survey of Graph Neural Networks and Applications. Wireless Communications and Mobile Computing , 2022:1–18.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic Context-guided Capsule Network for Multimodal Machine Translation. In Proceedings of the 28th ACM International Conference on Multimedia , MM '20, pages 1320–1329, New York, NY, USA. Association for Computing Machinery.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In Advances in Neural Information Processing Systems , volume 36, pages 34892–34916. Curran Associates, Inc.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2022. Gumbel-Attention for Multi-modal Machine Translation. _eprint: 2103.08862.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics , 8:726–742. 01870 Place: Cambridge, MA Publisher: MIT Press.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative Imagination Elevates Machine Translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pages 5738–5748, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02 , page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers , pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. 2024. Accept the modality gap: An exploration in the hyperbolic space. In 2024 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR) , pages 27253–27262.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. Preprint , arXiv:2404.07983.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers , pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014a. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 , NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014b. Sequence to sequence learning with neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 , NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024a. Encoder–Decoder Calibration for Multimodal Machine Translation. IEEE Transactions on Artificial Intelligence , 5(8):3965–3973.

Turghun Tayir, Lin Li, Xiaohui Tao, Mieradilijiang Maimaiti, Ming Li, and Jianquan Liu. 2024b. Visual Pivoting Unsupervised Multimodal Machine Translation in Low-Resource Distant Language Pairs. In Findings of the Association for Computational Linguistics: EMNLP 2024 , pages 5596–5607, Miami, Florida, USA. Association for Computational Linguistics.

Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. arXiv preprint .

Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. 2018. Scene Graph Parsing as Dependency Parsing. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) , pages 397–407, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In 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) , pages 6153–6166, Online. Association for Computational Linguistics.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Yu Philip S. 2020. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems , 32(1):4–24.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020a. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 3025–3035, Online. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020b. A novel graph-based multi-modal fusion encoder for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , pages 3025–3035, Online. Association for Computational Linguistics.

Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. In Proceedings of the 32nd International Conference on Neural Information Processing Systems , NIPS'18, pages 4805–4815, Red Hook, NY, USA. Curran Associates Inc. Event-place: Montréal, Canada.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural Machine Translation with Universal Visual Representation. In International Conference on Learning Representations .

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. AI Open , 1:57–81.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating BERT into Neural Machine Translation. In International Conference on Learning Representations .

A LLaVA 的场景图提示

Please analyze the image provided and construct a structured scene graph, adhering to the following guidelines, and represent it in a JSONL (JSON Lines) format:
1. Entities: List all significant objects or subjects visible in the image, which may include things, animals, or people. Describe each entity in detail, noting their quantities, colors, and any distinctive features. Each description should be distinct and consistent across the document to ensure clarity.
2. Relations: Define all pivotal relationships between the entities using tuples. Each tuple must maintain the exact terminology used in the entities' descriptions. These relationships should be expressed as triplets: [subject entity, predicate, object entity]. Importantly, ensure that the scene graph forms a connected structure. Every entity appearing as a subject or object in one relation must connect to another entity in a different relation, preventing any isolated nodes or subgraphs within the graph. In cases involving an entity related to multiple others, such as being 'between' or 'consist of' them, express this by dividing the relationship into distinct tuples using descriptors like 'is positioned between' and 'and also between' to maintain clarity. Generate triplets with a subject, an active verb or relational word, and a distinct object. Each triplet should clearly describe an action or relationship, avoiding states or implied conditions.
Avoid focusing on too detailed or minor elements that do not significantly contribute to the scene's overall understanding. Use active verbs that show a clear action or relationship. Avoid state or possession verbs like "have" that imply a condition without a distinct action. Incorrect Relations Examples to Avoid:

1.["one person in red shirt", "one dog", "one cat"] (lacks clear action)
2.… … … …
Correct Relations Examples of the above, the number of the example is the same as the number of the incorrect example:

1.["one person in red shirt", "is holding", "a book"]
2.… … … …
Key Point: Ensure every triplet uses an active verb or distinct relational word to connect the subject and object, clearly describing a specific action or relationship and forming a triplet.

This structure ensures that the scene graph is comprehensive and interconnected, accurately reflecting the dynamics and layout of the scene. The response must strictly follow the JSONL format specified here and not include any extraneous text.

This is a scene graph JSONL example response of the Example Image, the entity_descriptions1, entity_descriptions2, entity_descriptions3, entity_descriptions4 and entity_descriptions5 need to be replaced by specific entities in the image. The relation word1, relation word2, and relation word3 are also need to be replaced by the specific action or relation you observe in the given image. Also, the number of entities and relations is not fixed. It should depend on the given image. The following scene graph JSONL is just an example. You need to describe the real relations based on your given image.

```
{ "entities": ["entity_descriptions1", "entity_descriptions2", "entity_descriptions3", "entity_descriptions4", entity_descriptions5], "relations": [["entity_descriptions1", "relation word1", "entity_descriptions3"], ["entity_descriptions2", "relation word2", "entity_descriptions4"], ["entity_descriptions1", "relation word3", "entity_descriptions5"]]
}
```

You must not include the word 'image' in the scene graph JSONL. You must not copy the example above! You must describe the entities and their relationships in the given image. Now, you must respond to the scene graph based on the image provided! Straitly follow my instructions. Now what is the scene graph of the image?

我们使用 RTX A6000 GPU 来运行 Llava-34B。针对每个查询，系统信息解释了任务描述，并提供了以 JSONL 格式生成场景图的指导方针。我们还有一个质量评估脚本，用于剔除任何未能满足提示中提供的生成任务描述的 ISG 数据。我们将温度设置为 0，作为多模态大型语言模型（MLLMs）的默认设置，以获得相对稳健的性能。如果 MLLM 在温度为 0 的情况下无法生成满足要求的场景图，我们将切换到温度为 0.4。我们专门使用 MLLM 进行图像预处理，以根据我们的任务描述生成场景图，并通过脚本确保质量。

B 实验比较和分析与 MLLMs

Model	EN → DE						EN → FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
LlaVa-7B	27.15	58.54	23.70	52.05	19.54	47.77	35.67	65.57	34.79	62.94	35.00	62.87
LlaVa-34B	25.30	58.76	25.16	55.58	22.04	51.53	40.25	69.47	38.95	67.36	39.99	68.82
mBART	41.12	69.59	36.63	65.07	32.89	60.15	63.37	82.40	57.01	77.63	47.28	71.58
GIIFT (ours)	44.04	71.08	38.41	65.88	34.94	61.66	65.61	83.65	58.05	78.36	49.72	73.86

Table 5: 与多模态 LLM 的比较。

我们使用 RTX A6000 GPU 来应用基于 Ollama 的 Llava-34B 和 Llava-7B。我们采用 LLaVA 的少样本推理范式，通过使用从 Multi30K 提取的一小组 (图片、来源标题、目标翻译) 示例来提示模型，然后要求其将新图片的标题翻译成目标语言。GIIFT 和 mBART 骨干网均包含约 6 亿个参数，远少于 LLaVA 的参数量一个数量级，证明架构设计可以超越参数规模。GIIFT 在 EN → { DE, FR } 基准测试中取得了最高的 BLEU 和 METEOR 分数，显著优于 mBART。这些结果证实了我们模型的有效性，并表明紧凑且领域专用的多模态模型可以胜过更大的通用 LLM。

LLaVA 之所以落后于 mBART，主要是因为它的训练和架构是为通用视觉语言指令而非翻译优化的。它依赖于简单的特征投影来注入视觉上下文，而基于 mBART 的模型使用专门为多模态翻译调优的门控融合机制。在少量样本提示下，LLaVA 必须将示例和图像放入有限的上下文窗口，这可能导致格式过拟合和风格漂移。最后，尽管 LLaVA 的参数多得多，但它们支持广泛的功能；而 mBART 的参数完全用于序列到序列的翻译，从而在这一任务上带来了更高的效率和准确性。

如图 **??**（左）所示，我们的完整模型 GIIFT（我们的方法）正确地将文本与视觉场景上下文关联起来，准确翻译了可分动词，使用 “auf” 来表达 “执行” 动作，而 GIIFT（没有阶段 1）错误地将其翻译为 “vor” 来表达 “展示”。mBART 未能正确理解场景，导致单词排列混乱。

如图 **??**（右）所示，我们的完整模型 GIIFT (ours) 正确地将时态从文本信息推广到视觉信息，准确地将人群聚集的状态翻译为 “has gathered” 而不是直接翻译 “gathered”。GIIFT (w/o. Stage 1) 将其错误地翻译为进行时态 “is gathered” 。虽然 mBART 使用了正确的时态，但它依然未能正确理解场景背景，遗漏了 “in the park” 这个位置信息，并错误地把修饰语 “in the rain” 归于其他，导致整体语义混乱。