GIIFT: Graph-guided Inductive Image-free Multimodal Machine Translation

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

Multimodal Machine Translation (MMT) has demonstrated the significant help of visual information in machine translation. However, existing MMT methods face challenges in leveraging the modality gap by enforcing rigid visual-linguistic alignment whilst being confined to inference within their trained multimodal domains. In this work, we construct novel multimodal scene graphs to preserve and integrate modality-specific information and introduce **GIIFT**, a two-stage **G**raph-guided **I**nductive **I**mage-**F**ree MM**T** framework that uses a cross-modal Graph Attention Network adapter to learn multimodal knowledge in a unified fused space and inductively generalize it to broader image-free translation domains. Experimental results on the Multi30K dataset of English-to-French, English-to-German, and English-to-Ukraine tasks demonstrate that our GIIFT surpasses existing MMT baselines even without images during inference. Results on the WMT benchmark show significant improvements over the image-free MMT translation baselines, demonstrating the strength of GIIFT towards inductive image-free inference [1].

1 Introduction

Multimodal machine translation (MMT) aims to improve traditional text-only neural machine translation (NMT) by incorporating multimodal data, particularly visual inputs (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Existing methods mostly focus on forcing the alignment between image and text to improve MMT, that they have demonstrated the effectiveness benefited from the aligned visual information in disambiguating text (Ive et al., 2019; Zhao et al., 2022b; Futeral et al., 2023). However, an aligned form of multimodal dataset in both the training and inference phases has disabled the MMT model to generalize further in the normal pure-text machine translation setup. Although the rich information of images can bring benefits to translation beyond the level of text, when aligned image information is indispensable for inference in the translation process, the application of MMT models will be severely limited. Therefore, the inability to get rid of aligned images in the inference phase is the critical bottleneck for the flexible application of MMT models.

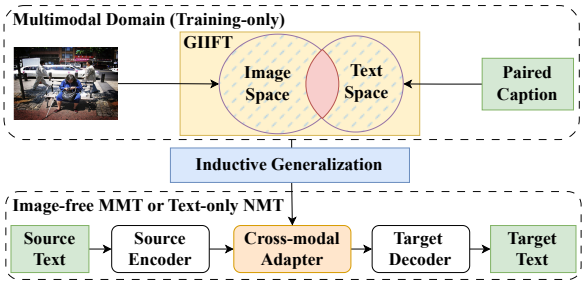

Figure 1: The inductive image-free generalization of GIIFT. GIIFT inductively learns from an entire multimodal domain, then enables image-free inference for MMT or text-only NMT via cross-modal generalization. In contrast, previous models can only learn from the limited overlap between image and text in red.

To address the bottleneck mentioned above, there have been a few methods that attempt to explore resolutions for achieving image-free inference in MMT models. For example, synthesizing visual hallucination from text or textual scene graph during training is used for the image-free inference (Li et al., 2022; Fei et al., 2023); transfer learning from the image-to-text captioning task to the text-to-text translation task (Gupta et al., 2023). However, none of these efforts has managed to consistently reach the performance of fully bridging the gap between multimodal and image-free inference. There are still critical challenges in advancing image-free inference in MMT.

First, previous models learn inadequately because they forcibly align the modality gap, typically the intrinsic information imbalance between im-

[1] Code is available at: https://github.com/xjiaf/GIIFT

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 98–112
November 8–9, 2025.

ages and texts (Schrodi et al., 2025). Consider the case in Figure 1, this constraint limits image-free inference generalized from only the red overlap in the multimodal domain. Recent work (Ramasinghe et al., 2024; Schrodi et al., 2025), however, shows that accepting this gap and exploiting the full information (the entire collection in Figure 1) substantially enhances multimodal learning. Second, current image-free MMT approaches fail to realize cross-modal generalization, resulting in a marked drop in image-free inference compared with traditional MMT with multimodal inference (Fei et al., 2023). Third, current MMT models are transductive (Sutskever et al., 2014), which struggle with inductively generalizing from multimodal domains to text-only domains for broader applications. Facing these challenges, we investigate the following research questions in this work: *RQ1: How can we fully represent different modalities to embrace the modality gap? RQ2: Can we effectively make image-free inference via cross-modal generalization without downgrading performance? RQ3: Can MMT have the inductive ability to generalize multimodal domains to broader text-only domains?*

We propose **GIIFT**, a two-stage **G**raph-guided **I**nductive **I**mage-**F**ree MM**T** framework. As shown in Figure 1, the GIIFT learns from the entire multimodal domain in the first stage, and aims to achieve inductive generalization for image-free MMT or text-only NMT in the second stage. Graph-structured novel Multimodal Scene Graph (MSG) and Linguistic Scene Graph (LSG) are proposed to represent the multimodal domain, in which each modality informs and enriches the other by graph representation in the unified space. Specifically, we extract visual relationships from images and linguistic relationships from text, then preserve and integrate them via global super nodes to construct MSGs. LSG is a linguistic version, preserving linguistic relationships with global super nodes. These relations' and nodes' features are mutually enriched and uniformly initialized via M-CLIP (Carlsson et al., 2022). To enable cross-modal generalization to image-free or broader text-only domains, GIIFT is designed with a cross-modal GAT adapter to inductively learn multimodal knowledge from the multimodal domain via MSGs in the first stage, and generalize it for image-free inference via LSGs in the second stage, based on a replaceable backbone, mBART (Liu et al., 2020).

Overall, the main contributions are:

(i) We build novel MSGs and LSGs to fully represent different modalities in a unified space for embracing the modality gap.

(ii) We propose the two-stage GIIFT framework with a novel lightweight GAT adapter to achieve inductive cross-modal generalization for image-free inference MMT via MSGs and LSGs.

(iii) Experimental results on En→{Fr, De, Uk} Multi30K show that GIIFT outperforms most of the existing MMT methods even without image during inference. On En→{Fr, De} WMT, GIIFT surpasses the best image-free baseline, CLIPTrans, by average gains of **+1.92 (8.00%) BLEU** and **+2.82 (4.80%) METEOR**, demonstrating effective induction from multimodal Multi30K to other text-only NMT domains.

(iv) Further analysis demonstrates that the proposed GIIFT can effectively embrace modality gaps via MSGs and achieve robust image-free inference via two-stage cross-modal generalization, and shows that the GIIFT can achieve consistent performance of fully bridging the gap between multimodal inference and image-free inference.

2 Related Work

2.1 Multimodal Machine Translation

MMT research integrates visual and textual information for machine translation with an increasing number of models (Grönroos et al., 2018; Huang et al., 2020; Zhao et al., 2022a; Cheng et al., 2024). Early approaches commonly adopted RNN-based architectures enhanced with attention mechanisms to incorporate global or spatial visual features (Calixto et al., 2017). Transformer variants soon supplanted RNNs, introducing tighter cross-modal fusion such as dynamic token re-weighting (Caglayan et al., 2018; Lin et al., 2020), double attention mechanisms (Zhao et al., 2020), gating mechanisms (Wu et al., 2021), multimodal adapters (Zhao and Calapodescu, 2022) or multi-granular fusion via graph structures (Krishna et al., 2017; Wang et al., 2018; Yin et al., 2020). Recent research leverages pre-trained resources such as CLIP (Gupta et al., 2023; Li et al., 2022) or BERT (Li et al., 2020). Within the widely adopted encoder-decoder framework, MMT research has progressed along two fronts in representation and inference:

(1) Visual-linguistic representation. Most MMT models enforce rigid visual-linguistic alignment. Disambiguation work (Ive et al., 2019; Zhao et al., 2022b; Futeral et al., 2023) links each textual token to a matching image region to resolve lexi-

cal ambiguity. UMMT(Fei et al., 2023) aligns every hallucinated visual scene graph node with textual counterpart. CLIPTrans (Gupta et al., 2023) trains sequentially on two stages, image captioning and the corresponding translation, enforcing alignment across both stages. Such alignments discard modality-specific information by merely learning the overlap between modalities.

(2) Image-free inference. Previous MMT relies on access to the paired image at test time. To mitigate this limitation, researchers have pursued three lines of work. First, retrieval-based models replace the missing picture with visual features fetched from an indexed caption-image bank into the decoder (Zhang et al., 2020). Second, hallucination methods (Johnson et al., 2018; Li et al., 2022; Fei et al., 2023) synthesize visual inputs from texts. Third, using transfer learning to train the NMT models with images (Gupta et al., 2023).

2.2 Graph Neural Networks

GNN is powerful in modeling relational structures by leveraging message-passing mechanisms (Wu et al., 2020; Liang et al., 2022), which iteratively aggregates and updates nodes' representation with information from their neighbors, capturing both local and global relational patterns. Some GNNs, such as Graph Convolutional Network (GCN) (Kipf and Welling, 2017) and its variants, are only transductive, while others, including GAT (Veličković et al., 2018), and GraphSAGE (Hamilton et al., 2017), also enable inductive learning (Battaglia et al., 2018; Xiong et al., 2025) to handle previously unseen nodes (Zhou et al., 2020). GNNs also use hierarchical or global pooling techniques (Ying et al., 2018; Lee et al., 2019; Gao and Ji, 2019) to capture subgraph-level or graph-level embeddings (Zhou et al., 2020).

3 Methodology

In this work, we construct unified MSGs and LSGs to generalize multimodal knowledge for image-free inference. We design the GIIFT framework with two-stage continuous learning via a lightweight GAT adapter for inductive cross-modal generalization. Our methodology is structured as follows: 1) We introduce the problem definition of our inductive image-free inference (Subsection 3.1). 2) The details of the MSG and LSG scene graph generation (Subsection 3.2). 3) The description of the GIIFT framework (Subsection 3.3).

3.1 Problem Definition

Let $\mathcal{D}_m$ be a multimodal multilingual dataset of triplets (i, c_s, c_t), where i is an image and (c_s, c_t) are its source and target captions, and let $\mathcal{D}_l$ be a text-only parallel corpus of pairs (t_s, t_t). Traditional MMT methods align i with c_s during training and then perform image-free inference based on that alignment, but the visual knowledge remains tied to c_s and $\mathcal{D}_m$. Our inductive image-free inference instead learns multimodal knowledge from i and (c_s, c_t) in $\mathcal{D}_m$ that cross-modally generalizes to both $\mathcal{D}_m$ or translation pairs $(t_s, t_t) \in \mathcal{D}_l$.

3.2 Scene Graph Generation

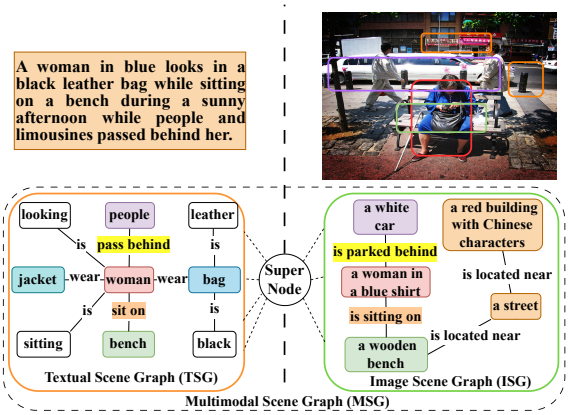

Figure 2: Representation of textual and visual information via MSG. Corresponding objects, attributes, and relations across both Textual Scene Graphs and Image Scene Graphs are depicted using identical colors.

To preserve modality-specific information and extract complex relations (e.g., spatial relations, environmental scene, and event states of people and objects) during data preprocessing, we use a multimodal Large Language Model (LLM) as the image parser and an off-the-shelf text parser to build the MSG and LSG, respectively. Figure 2 shows that the MSG comprises an Image Scene Graph (ISG) and a Textual Scene Graph (TSG), and a visual super node to bridge ISG and TSG included in a unified space, whereas the LSG replaces the visual super node with a textual one and omits the ISG.

(1) Image Scene Graph. ISGs are obtained from images i using LLaVA-34B (Liu et al., 2023). To obtain coherent and well-structured outputs, we set the sampling temperature to 0 for deterministic generations and raise it to 0.4 for more challenging cases requiring exploratory outputs. Our prompts (details in Appendix A) comprise: **(i) Task Description**, specifying how to formulate relations and produce structured scene graph content; **(ii)**

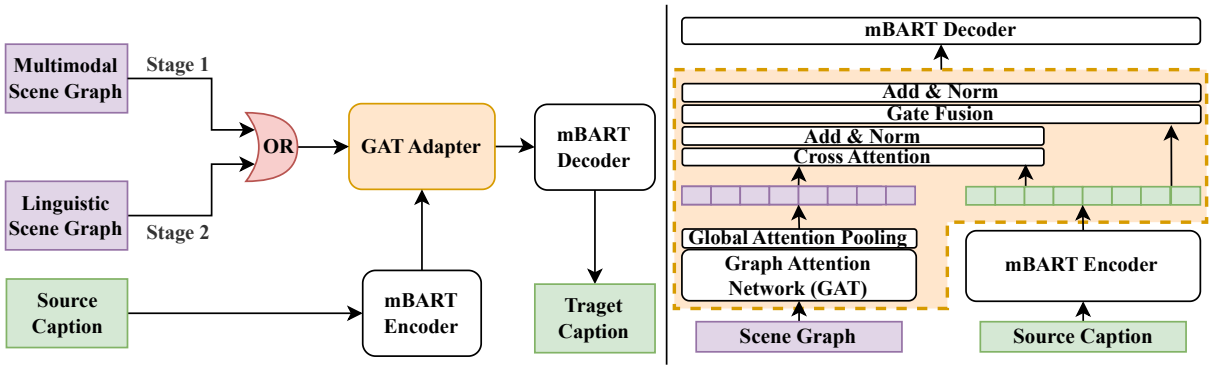

Figure 3: Left: Overview of the two-stage GIIFT framework. Stage 1: multimodal learning via MSGs. Stage 2: cross-modal generalization via LSGs. Right: Overview of the architecture of the cross-modal GAT adapter, which inductively learns and fuses the multimodal knowledge for the backbone, mBART.

Negative Examples, illustrating common errors in output formatting and how to rectify them; **(iii) Format Examples**, providing abstract but well-formed scene graph templates without using specific objects, thus preventing prompt contamination. Therefore, we generate ISGs with unique and relational visual information, such as event states.

(2) Textual Scene Graph. TSGs are parsed by FACTUAL (Li et al., 2023) from texts c_s or t_s. FACTUAL is lighter and more efficient than LLMs for large-scale corpora. TSGs encode entities and their relations as the textual analogue to ISGs. Thus, we obtain TSGs with structured linguistic relationships and unique semantic information.

(3) Multimodal Scene Graph. For each pair (i, c_s), we merge the ISG and TSG into an MSG by introducing a super node that encodes holistic image embeddings from the M-CLIP image encoder. This super node connects to all ordinary ISG and TSG nodes to unite modality-specific information and deliver diverse granular information, serving as a critical bridge to accept the modality gap and build multimodal relationships. We embed all ordinary node and relation features by the M-CLIP text encoder, enabling a unified representation of multimodal relationships and inductive foundations.

(4) Linguistic Scene Graph. For cross-modal generalization, we construct the LSG for text pairs (t_s, t_t) by retaining only the TSG with a super node representing the entire text embedding via the M-CLIP text encoder, which shares the unified hidden space with MSG's. Similarly, we embed all ordinary nodes and edges using the M-CLIP text encoder. The super node connects to all ordinary TSG nodes, enabling multi-granular textual information for image-free inference.

3.3 GIIFT Framework

Harnessing scene graphs as modality-bridges, GIIFT uses MSGs and LSGs, which are inductively learned for multimodal knowledge and generalized for image-free translation, respectively, via an essential cross-modal GAT adapter to guide the backbone mBART in two-stage training pipelines. The proposed lightweight GAT adapter is completely decoupled from the pre-trained mBART backbone (as shown in Figure 3), which can be flexibly incorporated with other language models for deployment.

3.3.1 Two-stage Training Framework

Stage 1: Multimodal Learning via MSG. As shown in Figure 3 (Left), Stage 1 trains a shared GAT adapter on MSGs from paired images and captions, inducing multimodal knowledge that guides image-free translation. Within the MSG, $\mathcal{N}_i^{\mathrm{SG}}$ is the neighbors of the ordinary node i from ISG or TSG, reflecting the structured relationships from image or text. The embedding $\mathbf{Z}_i^{(l)}$ of node i at layer $l, (0 \le l \le L)$ in GAT is calculated recursively as:

$$\begin{aligned}\mathbf{Z}_i^{(l)} = \sigma\Big(\sum_{j\in\mathcal{N}_i^{\mathrm{SG}}} \alpha_{ij}^{(l)}\, \phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)} \| \mathbf{Z}_j^{(l-1)} \| \mathbf{E}_{ij}]) \\ + \alpha_{i,\mathrm{SN}}^{(l)}\, \phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)} \| \mathbf{Z}_{\mathrm{SN}}^{(l-1)} \| \mathbf{E}_{i,\mathrm{SN}}])\Big).\end{aligned} \tag{1}$$

Here, $\mathbf{Z}_i^{(l-1)}$ is the embedding of node i at the previous layer (with initial node embedding $\mathbf{Z}_i^{(0)}$ via M-CLIP text encoder), $\mathbf{Z}_{\mathrm{SN}}^{(l-1)}$ is the global multimodal embedding from the super node at layer $l-1$ which is passed to all the ordinary nodes, and $\mathbf{E}_{ij}$ or $\mathbf{E}_{i,\mathrm{SN}}$ denotes edge embedding of the relationships in the scene graph via M-CLIP text encoder, the $\alpha_{ij}^{(l)}$ and $\alpha_{i,\mathrm{SN}}^{(l)}$ are attention weights, $\sigma(\cdot)$ is

an activate function and $\phi(\cdot)$ is the LeakyReLU. $\mathbf{E}_{ij}$ is the embedding of the relation content in the scene graph, obtained with the M-CLIP text encoder, whereas $\mathbf{E}_{i,\mathrm{SN}}$ is the embedding of the edge between the super-node and each ordinary node i. Because these super-node edges have no textual content, we initialize $\mathbf{E}_{i,\mathrm{SN}}$ as a ones vector whose dimensionality matches that of the embeddings produced by the M-CLIP text encoder.

From Eq. (1), we can observe that the shared weight $\mathbf{W}$ enables learning of multi-granular multimodal relationships by jointly processing local textual embeddings from ISG or TSG nodes ($\mathcal{N}_i^{\mathrm{SG}}$) and global multimodal context from the super node. This shared weight $\mathbf{W}$ is crucial in capturing these multimodal relationships and will be leveraged for the cross-modal generalization process in Stage 2.

The initial MSG super node provides the global image embedding $\mathbf{Z}_{\mathrm{SN}}^{\mathrm{MSG}(0)}$ and aggregates information from all ordinary nodes as follows:

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

where $\mathcal{N}_{\mathrm{SN}}^{\mathrm{MSG}}$ contains all ordinary nodes in MSG.

Through the Global Attention Pooling (Li et al., 2015) layer in the GAT adapter, we then obtain the multimodal graph representation $\mathbf{Z}_g^{\mathrm{MSG}} \in \mathcal{D}_m$:

$$\mathbf{Z}_g^{\mathrm{MSG}} = \mathrm{AttnPool}(\{\mathbf{Z}_i^{\mathrm{MSG}} : i \in \mathcal{V}_{\mathrm{MSG}}\}), \tag{3}$$

where $\mathcal{V}_{\mathrm{MSG}}$ denotes the set of nodes in the MSG, including ordinary nodes and the super node.

$\mathbf{Z}_g^{\mathrm{MSG}}$ is then fed into the decoder for translation. The encoder remains frozen and the decoder learns a balanced representation to generate translations based on the gate mechanism (see Eq. (8)) between multimodal representations $\mathbf{Z}_g^{\mathrm{MSG}}$ and the source embedding $\mathbf{H}$.

Stage 2: Cross-modal Generalization via LSG. As shown in Figure 3 (Left), Stage 2 inputs the same GAT adapter with LSGs built from texts, allowing multimodal knowledge to be cross-modally generalized to broader image-free domains. In the LSG, each ordinary node i represents a textual entity with the initial embedding $\mathbf{Z}_i^{(0)}$ and a super node of the global text embedding, $\mathbf{Z}_{\mathrm{SN}}^{\mathrm{LSG}(0)}$ by the M-CLIP text encoder. The embedding $\mathbf{Z}_i^{(l)}$ of node i at the layer l for an ordinary node is:

$$\begin{aligned}\mathbf{Z}_i^{(l)} = \sigma(&\sum_{j\in\mathcal{N}_i^{\mathrm{LSG}}}\alpha_{ij}^{(l)}\phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)}\|\mathbf{Z}_j^{(l-1)}\|\mathbf{E}_{ij}]) \\ &+\alpha_{i,\mathrm{SN}}^{(l)}\phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)}\|\mathbf{Z}_{\mathrm{SN}}^{\mathrm{LSG}(l-1)}\|\mathbf{E}_{i,\mathrm{SN}}])).\end{aligned} \tag{4}$$

The LSG super node is updated as:

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

with $\mathcal{N}_{\mathrm{SN}}^{\mathrm{LSG}}$ denoting all LSG ordinary nodes. The LSGs help the generalization of shared multimodal knowledge weight $\mathbf{W}$ from $\mathcal{D}_m$ in Stage 1 to the image-free domain $\mathcal{D}_l$ in Stage 2.

Similar to Eq. (3), we obtain the graph representation of LSG $\mathbf{Z}_g^{\mathrm{LSG}}$. The mBART decoder is enhanced by generalized knowledge $\mathbf{Z}_g^{\mathrm{LSG}}$ from $\mathcal{D}_m$ and adapted to the image-free domain $\mathcal{D}_l$ with unfrozen mBART encoder hidden state.

3.3.2 Cross-modal GAT Adapter

We employ a multi-layer GAT with residual connections (He et al., 2015; Veličković et al., 2018; Li et al., 2019) to learn multimodal knowledge and cross-modally generalize it to a broader image-free domain. Figure 3 (Right) shows the architecture of the GAT adapter fusing the output from the mBART encoder and enhancing the mBART decoder. We denote $\mathbf{Z}_g$ as the graph representation of MSG or LSG by the Global Attention Pooling. The fusion output $\mathbf{O}$ between the graph representation $\mathbf{Z}_g$ and mBART encoder hidden states $\mathbf{H}$ is performed via the cross attention as:

$$\mathbf{A} = \mathrm{MultiHeadAttn}(\mathbf{H}, \mathbf{Z}_g, \mathbf{Z}_g) + \mathbf{H} \tag{6}$$

$$\mathbf{O} = \mathrm{LayerNorm}(\mathrm{Dropout}(\mathbf{A})) \tag{7}$$

A gate mechanism $\mathbf{g}$ balances the embedding flow:

$$\mathbf{g} = \sigma(\mathbf{W}_2\mathrm{ReLU}(\mathbf{W}_1[\mathbf{O}\|\mathbf{H}])) \tag{8}$$

$$\mathbf{H}' = \mathrm{LayerNorm}(\mathbf{g}\odot\mathbf{O} + (1-\mathbf{g})\odot\mathbf{H}) \tag{9}$$

where $\odot$ denotes element-wise multiplication and $[\cdot\|\cdot]$ represents concatenation, $\mathbf{H}'$ is the input of mBART decoder. This two-stage process demonstrates the central inductive role of the cross-modal GAT adapter in representing and generalizing structured multimodal relationships.

4 Experiments

Datasets. We conduct experiments on two benchmarks: Multi30K (Elliott et al., 2016) and WMT2014 (Bojar et al., 2014). Multi30K is a widely used MMT benchmark as a multilingual extension of the Flickr30k. WMT is a text-only multilingual NMT dataset. For evaluation, we conduct EN→{DE, FR} translation tasks on Multi30K's

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., 2020)	39.80	32.20	28.70	60.90	53.90	-	-5.14
Gated Fusion* (Wu et al., 2021)	42.00	33.60	29.00	61.70	54.80	44.90	-4.13
WRA-MNMT (Zhao et al., 2022b)	39.30	32.30	28.50	61.70	54.10	43.40	-5.02
UMMT# (Fei et al., 2023)	37.40	-	-	56.90	-	-	-7.68
Soul-Mix (Cheng et al., 2024)	**44.24**	37.14	34.26	64.75	57.47	49.25	-0.61
GIIFT (with image)	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 (image-free)	44.04	**38.41**	**34.94**	**65.61**	58.05	49.72	

Table 1: BLEU on the Multi30K. Δ is gap vs. "GIIFT (image-free)". "GIIFT (with image)" is trained and tested with images and texts in only one stage with unfrozen mBART. * represent ensembled models, # denotes the model trained and tested with images and texts.

Model	EN → DE			EN → FR			Mean Δ
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
mBART (NMT backbone)	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., 2020)	57.60	51.90	47.60	74.90	69.30	-	-9.72
Gated Fusion* (Wu et al., 2021)	67.80	61.90	56.10	81.00	76.30	70.50	-3.48
WRA-MNMT (Zhao et al., 2022b)	58.30	52.80	48.50	76.30	70.60	63.80	-8.92
UMMT# (Fei et al., 2023)	57.20	-	-	70.70	-	-	-13.42
Soul-Mix (Cheng et al., 2024)	69.93	63.59	59.94	83.24	78.23	73.48	-1.01
GIIFT (with image)	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 (image-free)	**71.08**	**65.88**	**61.66**	**83.65**	78.36	73.86	

Table 2: METEOR on the Multi30K. Δ is gap vs. "GIIFT (image-free)". "GIIFT (with image)" is trained and tested with images and texts in only one stage with unfrozen mBART. * represent ensembled models, # denotes the model trained and tested with images and texts.

Model	EN → UK					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
mBART	53.43	74.67	46.05	68.86	44.05	66.75
CLIPTrans	54.01	74.76	46.30	69.23	44.11	66.87
GIIFT (image-free)	**55.10**	**75.18**	**47.85**	**70.01**	**44.93**	**67.05**

Table 3: BLEU and METEOR on EN → UK (Ukraine) task of the Multi30K.

Model	EN → DE		EN → FR	
	BLEU	METEOR	BLEU	METEOR
mBART	15.58	41.18	26.50	52.06
CLIPTrans	16.63	42.13	26.78	51.76
(*w/o.* Stage 1)	17.60	42.81	27.71	53.38
GIIFT (image-free)	**18.10**	**43.88**	**28.70**	**54.58**
(*w/o.* Stage 1)	17.79	43.01	27.89	53.45

Table 4: Comparison of domain generalization on text-only WMT. "(*w/o.* Stage 1)" denotes model trained without image involvement from Multi30K.

three standard test splits: Test2016, Test2017, and MSCOCO. We also train and test EN→{DE, FR} on WMT with multimodal knowledge from Multi30K to evaluate the inductive image-free inference. We downsample WMT train set matching Multi30K's size, while keeping the validation and test sets unchanged.

Implementation Details. We train GIIFT on an A100 GPU, with AdamW optimizer (polynomial decay). The batch size is 64, learning rate is $2e^{-5}$ (Stage 1) and $1e^{-5}$ (Stage 2). The GAT Adapter is 9 layers with the same 1024 dimensions as M-CLIP. Text decoding is beam search with size 5. Implementations are in PyTorch and Huggingface

Transformers library. We report BLEU (Papineni et al., 2001) and METEOR via SacreBLEU (Post, 2018) and the evaluate library (Banerjee and Lavie, 2005), respectively. Results are three-run averages with early-stop patience 5 on BLEU. We chose mBART as our backbone to ensure a fair comparison with baselines such as CLIPTrans (Gupta et al., 2023) and to highlight our improvements. All tables bold the best and underline the second-best. Baseline figures are derived from papers or repositories. All the numbers of each model reported are from their fine-tuned version on the corresponding dataset.

4.1 Results on Multi30K

Table 1 and 2 contain translation performance comparison of BLEU and METEOR on EN→{DE, FR} tasks of Multi30K. Table 3 shows the BLEU and METEOR on EN → UK (Ukraine) task of Multi30K compared with baselines. "GIIFT (with image)" is a single-stage variant where both training and testing use paired images and texts, and the mBART backbone remains fully unfrozen throughout. "GIIFT (image-free)" is the full image-free model trained in two-stage framework as Subsection 3.3.

From Table 1 and 2, we observe that GIIFT (image-free) surpass the strongest baseline, Soul-Mix (even tested with images), with an average gain of +0.61 (1.37%) BLEU and +1.01 (1.55%) METEOR, and over the best scene-graph-based image-free baseline, UMMT, by +13.525 (42.27%) BLEU and +18.865 (36.07%) METEOR. It shows the effectiveness of GIIFT by preserving the entire information from images and texts.

Moreover, the GIIFT (image-free) model with cross-modal generalizing multimodal knowledge for image-free inference and GIIFT (with image) with multimodal inference secure top-two ranks on 5 out of 6 benchmarks with overall parity. GIIFT (image-free) even surpasses GIIFT (with image) by +0.21 (0.63%) BLEU and +0.1 (0.17%) METEOR on average. In contrast, UMMT degrades dramatically without images, scoring on average -5.4 (-12.76%) BLEU and -5.45 (-8.53%) METEOR below UMMT$^{\#}$ with multimodal inference. This demonstrates that the GIIFT has the robustness of cross-modal generalization for image-free inference, which can mitigate the gap between multimodal inference and image-free inference.

From Table 3, we can find that GIIFT (image-free) consistently outperforms both mBART and CLIPTrans baselines on the EN→{UK} task of Multi30K. We can observe that GIIFT achieves an average of +1.45 higher than mBART and +1.15 higher than CLIPTrans in BLEU; and an average of +0.65 higher than mBART and +0.46 higher than CLIPTrans in METEOR. It shows the effectiveness of GIIFT by preserving and learning multimodal information from images and texts for cross-modal generalization.

4.2 Results on WMT

In Table 1, CLIPTrans outperforms other image-free inference baselines, so we adopt it as our primary baseline and use its official repository for experiments. Like GIIFT (image-free), this two-stage mBART-based model is trained with Stage 1 on Multi30K and Stage 2 on WMT. The results of CLIPTrans and GIIFT in Table 4 are obtained under the same model parameter setup.

Table 4 shows that GIIFT achieves the highest BLEU and METEOR scores overall, while also significantly outperforming GIIFT (*w/o.* Stage 1) trained without images from Multi30K on both metrics. This validates GIIFT's inductive ability to achieve robust image-free inference via cross-modal generalization.

The difference between CLIPTrans and GIIFT stems from how each model handles the modality gap and the resulting impact on cross-modal generalization. CLIPTrans attempts to align images to textual captions for transferring the Stage 1 visual features into Stage 2 caption translations via a mapping network, which limits the multimodal correlation for cross-modal generalization. By contrast, GIIFT can effectively embrace the modality gap via graph-guided fusion and achieve inductive generalization via a cross-modal GAT adapter. In Stage 1, all the modalities are learned and represented in a *unified multimodal knowledge space* via multimodal scene graphs (MSGs). In Stage 2, the proposed cross-modal GAT adapter generalizes that knowledge into the text-only domain via the assistance of linguistic scene graphs (LSGs). Benefit from a two-stage inductive learning via MSGs or LSGs, GIIFT shows better performance in achieving robust image-free inference performance.

4.3 Human Evaluation

Table 5 presents human evaluation on the Multi30K EN→FR test set using a 10-point Likert scale for completeness, ambiguity, and fluency, alongside

Model	BLEU	Completeness ↑	Ambiguity ↓	Fluency ↑
mBART	63.37	7.0	7.3	8.1
CLIPTrans	64.55	8.0	5.5	8.8
GIIFT (image-free)	**65.61**	**9.3**	**4.6**	**9.5**
(*w/o.* Stage 1)	64.91	8.8	5.1	9.0

Table 5: Human evaluation metrics (10-point Likert scale) and BLEU scores on Multi30K EN→FR task.

BLEU scores for reference.

Results in Table 5 show that our GIIFT substantially enhances key translation dimensions that correlate with human satisfaction. Compared with the baselines, GIIFT significantly outperforms in completeness, indicating that the model can utilize more multimodal context information. GIIFT achieves the highest completeness, fluency ratings, and the lowest ambiguity among all baselines, demonstrating practically meaningful gains even when BLEU is already high. These gains highlight that our method delivers practical improvements across dimensions that BLEU alone cannot capture.

4.4 Comparison with Multimodal LLMs

Table 6 shows the experimental comparison of BLEU and METEOR on EN→{DE, FR} tasks of Multi30K with multimodal LLMs.

We use an RTX A6000 GPU to employ LlaVA-7B, LlaVA-34B (Liu et al., 2023) and Llama3-70B (Grattafiori et al., 2024) based on Ollama. We adopt LLaVA's few-shot inference paradigm (Brown et al., 2020) by prompting the model with a small set of (image, source caption, target translation) examples drawn from Multi30K and then asking it to translate a new image's caption into the target language.

From Table 6, we observe that our GIIFT achieves the highest BLEU and METEOR scores on EN→{DE, FR} benchmarks, with significant improvements over mBART and LLMs. These results confirm the effectiveness of our GIIFT model and show that compact, domain-specialized multimodal MMT models can outperform larger general-purpose LLMs on the translation task.

Considering the parameter complexity compared with multimodal LLMs, the GIIFT only introduced the GAT adapter in the framework, which is lightweight and efficient for training. For instance, the mBART backbone has approximately 600M parameters, while our GIIFT framework adds only about 33.2M parameters in total, approximately 27M in the GAT adapter layers and 6.2M in the gated fusion layer. Although LLaVA has far more parameters for supporting broad capabilities, our GIIFT's parameters are wholly devoted to yielding greater efficiency and accuracy for the translation.

4.5 Ablation Study

To further verify the effectiveness of the different components in the GIIFT framework, we showed the experimental results of the ablated versions in Table 7 on EN→{DE, FR} Multi30K.

As shown in Table 7, the GIIFT (image-free) with two-stage continuous learning by the proposed GAT adapter achieves the best performance across all benchmarks. Removing the gating mechanism, GIIFT (*w/o.* gate), results in a more significant reduction in BLEU and METEOR scores, underscoring the critical balancing role of the gating mechanism to fuse multimodal knowledge and mBART information. Additionally, omitting the multimodal learning stage 1, GIIFT (*w/o.* Stage 1), leads to a decrease in performance, highlighting the importance of learning generalizable multimodal knowledge from MSGs. The performance of GIIFT (unfrozen) is significantly lower than GIIFT (image-free), and closer to the backbone model mBART. This underscores freezing the mBART encoder in Stage 1 to maintain its stable embedding. Compared with the backbone mBART, there is a substantial drop in BLEU and METEOR, which demonstrates the effectiveness of the GIIFT framework in learning multimodal knowledge and cross-modal generalization for image-free inference.

5 Case Study

To investigate the advantages of GIIFT, we examine cases in comparing: the GIIFT (image-free), which learns multimodal knowledge from MSGs for cross-modal image-free generalization via LSGs; GIIFT (*w/o.* Stage 1), which only learns linguistic relationships from LSGs; and the text-only mBART.

(i) Environmental scene. In Figure 4 (left), GIIFT (image-free) and GIIFT (*w/o.* Stage 1) both correctly capture the spatial preposition "on" through MSG or LSG, respectively, despite the source English caption omitting the spatial information. But lacking the scene graph, mBART produces imprecise translations without "on", which highlights the functions of LSGs to guide the learning and generalize the spatial relation knowledge. This case shows that LSGs guide GIIFT to better generalize spatial relation information from MSGs' multimodal knowledge in Stage 2 for image-free translation inference. Additionally, as shown in

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
Llama3-70B	40.49	69.09	36.12	65.06	34.38	60.86	50.95	77.13	49.39	74.54	49.38	73.39
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 (image-free)	**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 6: Comparison with Multimodal LLM on Multi30K.

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 (image-free)	**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 7: Ablation study on Multi30K. "GIIFT (*w/o.* freezing)" has an unfrozen mBART encoder in Stage 1. "GIIFT (*w/o.* Stage 1)" is trained without images. "GIIFT (*w/o.* gate)" is trained in two stages without gate fusion.

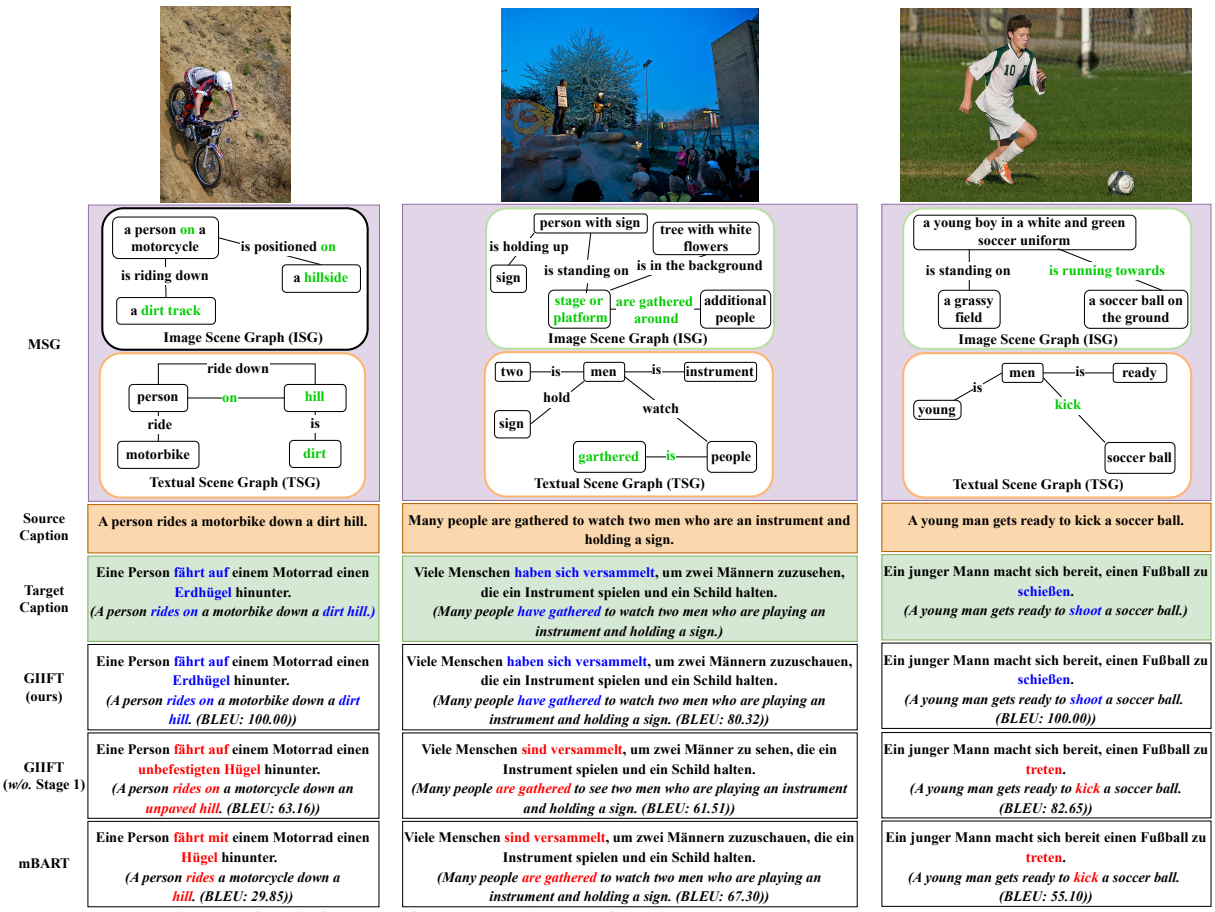

Figure 4: Under image-free inference, full GIIFT (image-free) is compared to GIIFT (*w/o.* Stage 1) and mBART on Multi30K validation set. The italicized bracketed translations of the German caption mark the differences in red.

Figure 4 (left), the environmental scene "dirt hill", is only accurately translated by full GIIFT (image-free) with multimodal knowledge from MSGs. GIIFT (*w/o.* Stage 1) and mBART, although "dirt" in the source caption, produce imprecise translations, reflecting overlooking of the modality-specific information. This case shows the effectiveness of MSGs that can well embrace modality gaps by multimodal graph fusion and fully preserve multimodal information for improving translation. Other cases from Test2016 are shown in Appendix B.

(ii) Temporal states. In Figure 4 (middle), the MSG captures a scene with "are gathered around the stage or platform", enabling full GIIFT to recognize it as a completed state and generate the appropriate perfect tense in German. In contrast,

both GIIFT (*w/o.* Stage 1) and mBART, limited by modality gap, cannot capture the temporal state from an aligned visual-linguistic space, which eliminates the unique temporal state from the image. So they translate English "are gathered" directly into the German present tense.

(iii) Action states. Figure 4 (right) shows the MSG's action state "is running towards" guides GIIFT (image-free) to correctly translate the action as "shoot" rather than "kick". The visual information, available only through MSG, enables the correct translation from multimodal knowledge. GIIFT (*w/o.* Stage 1) and mBART, however, can only literally translate to "kick" in German, which also shows the importance of accepting different modality-specific information rather than alignment.

6 Conclusion

This work introduces GIIFT, a two-stage graph-guided inductive MMT framework, along with novel Multimodal and Linguistic Scene Graphs. GIIFT outperforms existing MMT models on Multi30K even with image-free inference, demonstrating the effectiveness of learning multimodal knowledge in a unified space via MSGs and achieving cross-modal generalization via LSGs. Its image-free performance remains robust, matching its multimodal-inference counterpart. GIIFT also outperforms both the text-only NMT backbone and leading image-free MMT baselines on WMT, showing effective induction of multimodal knowledge to broader text-only domains. Further analysis highlights the advantages of multimodal graph-guided generalization for image-free inference and confirms the effectiveness of the two-stage framework with a lightweight but efficient GAT adapter for cross-modal inductive learning.

Acknowledgments

This work was supported by a Grant-in-Aid for Early-Career Scientists #24K20841, JSPS.

References

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.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *WMT*, pages 304–323.

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*.

Ondřej 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 Aleš 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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th international conference on neural information processing systems*, Nips '20, Red Hook, NY, USA. Curran Associates Inc. Number of pages: 25 Place: Vancouver, BC, Canada tex.articleno: 159.

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, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

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.

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. 2014. 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.

Petar Veličković, 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.

Jiafeng Xiong, Ahmad Zareie, and Rizos Sakellariou. 2025. A Survey of Link Prediction in Temporal Networks. *arXiv preprint*.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. 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*.

Yuting Zhao and Ioan Calapodescu. 2022. Multimodal robustness for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 8505–8516.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *Conference of the European Association for Machine Translation*, pages 105–114.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022a. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022b. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

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.

A Scene Graph Prompts for LLaVA

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?

We use an RTX A6000 GPU to employ Llava-34B. For each query, the system info explains the task description and provides guidelines for scene graph generation in JSONL format. We also have a quality assessment script to discard any ISG data that fails to meet our generation task description presented in the prompts. We take a temperature of 0 as the default for multimodal large language models (MLLMs) to have a relatively robust performance. If the MLLM can't generate scene graphs that meet the requirements with a temperature of 0, we'll switch to a temperature of 0.4. We exclusively use MLLM for image preprocessing to generate scene graphs according to our task description, with quality ensured by the script.

Source Caption	**Several women are performing a dance in front of a building**	**A crowd gathered around a park water fountain in the rain.**
Target Caption	**Mehrere Frauen führen vor einem Gebäude einen Tanz auf.** *(Several women perform a dance in front of a building.)*	**Eine Menschenmenge hat sich im Regen um einen Springbrunnen im Park versammelt.** *(A crowd has gathered in the rain around a fountain in the park.)*
GIIFT (ours)	**Mehrere Frauen führen vor einem Gebäude einen Tanz auf.** *(Several women perform a dance in front of a building. (BLEU: 100.00))*	**Eine Menschenmenge hat sich im Regen um einen Springbrunnen im Park versammelt.** *(A crowd has gathered in the rain around a fountain in the park. (BLEU: 100.00))*
GIIFT (*w/o.* Stage 1)	**Mehrere Frauen führen vor einem Gebäude einen Tanz vor.** *(Several women present a dance in front of a building. (BLEU: 78.25))*	**Eine Menschenmenge versammelt sich im Regen um einen Springbrunnen im Park.** *(A crowd is gathering in the rain around a fountain in the park. (BLEU: 64.56))*
mBART	**Mehrere Frauen führen einen Tanz vor einem Gebäude auf.** *(Several women perform a dance in-front-of a building up. (BLEU: 33.03))*	**Eine Menschenmenge hat sich um einen Springbrunnen im Regen versammelt.** *(A crowd has gathered around a fountain in the rain ~~in the park~~. (BLEU: 45.52))*

Figure 5: Case study of GIIFT on image-free inference when compared to GIIFT (*w/o.* Stage 1) and the mBART. Data points are drawn from the Test2016 set of Multi30K. The gold sentence represents the ground truth. The italicised sentence in the bracket presents the English translation of the German text, while red words highlight the crucial translation differences.

B Case Study on Multi30K Test2016 Set

As shown in Figure 5 (Left), our full model, GIIFT (image-free), correctly associates the text with visual scene context to translate the separable verb accurately, using "auf" to express the "perform" action, whilst GIIFT (*w/o.* Stage 1) incorrectly translates it as "vor" to express "present". mBART fails to properly understand the scene, resulting in disordered word arrangement.

As shown in Figure 5 (Right), our full model, GIIFT (image-free), correctly generalizes the tense from textual to visual information, accurately translating the crowd's gathered state as "has gathered" rather than directly translating "gathered". GIIFT (*w/o.* Stage 1) incorrectly translates it as the progressive tense "is gathered". Although mBART uses the correct tense, it still fails to properly understand the scene context, omitting the location "in the park" and misattributing the modifier "in the rain", leading to overall semantic confusion.