Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.083410

ARTICLE

Mitigating Visual Noise in Multimodal AI: Selective Visual Grounding for Multimodal Machine Translation

Ki-Young Shin[1], **Soonmo Kwon**[2] and **Kyudong Park**[3,*]

[1]Designovel Lab, Designovel, Pohang, Republic of Korea
[2]Department of Convergence IT Engineering, POSTECH, Pohang, Republic of Korea
[3]School of Information Convergence, Kwangwoon University, Seoul, Republic of Korea

*Corresponding Author: Kyudong Park. Email: kdpark@kw.ac.kr

Received: 03 April 2026; Accepted: 11 June 2026; Published: 27 July 2026

ABSTRACT: Multimodal AI systems often suffer from "over-informing", where excessive raw visual input introduces noise that distracts from task-relevant decisions. Motivated by selective human attention strategies, we propose ARS-MMT (Attention and Reasoning through Source Sentences for Multimodal Machine Translation), an architecture that operationalizes a "look-and-think" pipeline: a source-language encoder first builds contextualized linguistic representations, a relation reasoning network then produces a query-conditioned visual channel, and a multimodal decoder generates the translation conditioned in parallel on the encoded text and on this visual channel. We quantify the contribution of the visual modality through a controlled ablation: zeroing visual features reduces BLEU by 0.81 on test_2016_flickr En-De, while shuffling visual features across the batch changes BLEU by only +0.01, indicating that the channel responds primarily to the *presence* of visual context rather than to its image-specific content. We additionally add a contemporary 7B-parameter vision–language baseline (LLaVA-1.5) and show that our compact 4.3M-parameter specialized model is competitive in-domain. To address the open question of whether per-region visual attention constitutes a faithful explanation in the multimodal-translation setting, we conduct a deletion/insertion AUC analysis and report a null result consistent with prior findings on text attention. We therefore characterize ARS-MMT as an architecture whose modality-level visual contribution is measurable but whose per-region attention is not by itself a faithful explanation; faithful per-region attribution is identified as a target for complementary explanation methods. We discuss implications for efficient and inspectable multimodal systems in engineering deployment.

KEYWORDS: Explainable AI; multimodal machine translation; selective visual grounding; attention faithfulness; modality ablation; vision–language models

1 Introduction

1.1 The Problem of Over-Informing in Multimodal AI

The proliferation of multimodal AI systems, from visual question answering to image-assisted translation, has been driven by a seemingly intuitive assumption: more information leads to better decisions. Contemporary multimodal architectures typically employ global feature fusion, concatenating or attending over entire image representations alongside textual inputs [1]. This "more is better" paradigm has intensified with the emergence of large vision-language models such as LLaVA [2], ChatGPT, and Gemini, which process images holistically through powerful vision encoders before integrating them with language models. However, this paradigm overlooks a long-standing problem in Human-Computer Interaction research: information overload.

In multimodal machine translation (MMT), the task of translating text with the aid of accompanying images, this problem manifests acutely. An image depicting "a person sitting by a bank" contains vastly more information than the caption describes, including trees, sky, water ripples, distant buildings, and countless other visual elements. When translation systems indiscriminately incorporate all available visual features, irrelevant context becomes noise that can actively harm performance [3]. Recent comprehensive surveys confirm that this remains a persistent challenge: Shen et al. [4] reviewed 99 MMT studies and identified visual noise filtering as a key unresolved problem, while Nam and Jang [5] highlighted the gap between raw multimodal fusion and meaningful cross-modal understanding.

A second concern, particularly relevant for engineering deployment, is inspectability. Practitioners need to be able to characterize when and how a multimodal model uses each modality so that the model can be debugged, certified, or selectively replaced. Globally-fused vision-language models such as LLaVA-1.5 [6] and ViP-LLaVA [7] are highly capable but provide limited control over which visual signal influences the output, motivating compact architectures with explicit selective channels.

We term the underlying phenomenon *over-informing*, defined as the counterproductive provision of excessive auxiliary information that obscures decision rationale and limits efficient deployment. Just as research on information overload has long argued that systems should surface only the information a task requires rather than everything available [8], we argue that multimodal AI architectures must learn to filter rather than merely fuse.

1.2 Selective Attention as Design Inspiration

Decades of research in cognitive psychology have established that human attention is fundamentally selective and task-driven [9,10]. When reading text accompanied by images, humans employ what we characterize as a "look-and-think" strategy: they first process linguistic content, identify points of ambiguity or uncertainty, and only then consult visual context to resolve specific questions.

Consider encountering the sentence "She sat by the bank, watching the water flow." A proficient reader immediately recognizes the lexical ambiguity of "bank," which could denote a financial institution or a riverbank, and performs a targeted visual query: "Is there water? Is there a building?" This selectivity is cognitively efficient, avoiding the processing overhead of irrelevant visual features while precisely targeting information needed for disambiguation.

This pattern of selectivity contrasts with current multimodal AI approaches, which typically encode the entire image into a fixed representation before any linguistic processing occurs. We use the human "look-and-think" strategy as motivation for an architectural choice: rather than mix textual and visual features through global fusion, we route visual information through a query-conditioned channel built on top of a linguistic-side encoder. We characterize what this design actually delivers empirically: a small, inspectable model whose visual modality contribution is measurable at the modality level (Section 5.4). We do not, in this paper, claim that the resulting per-region attention performs human-like region-level selection; whether it does is an empirical question that we evaluate quantitatively and answer negatively in Section 5.6.

1.3 Contributions

This work makes three primary contributions:

1. **Architectural design**: We propose ARS-MMT, a compact (4.3M parameter) MMT architecture that operationalizes a "look-and-think" pipeline through a relation reasoning network. Visual information is routed through a query-conditioned channel rather than mixed with text through global fusion; we treat this routing as the structural property of the architecture and evaluate the per-region selectivity of the resulting attention as a separate empirical claim (Contribution 3).

2. **Quantitative modality analysis**: We provide a controlled ablation that quantifies the contribution of the visual modality to translation performance (zeroing visual features costs 0.81 BLEU on test_2016_flickr En-De, while shuffling visual features changes BLEU by only +0.01). We further compare against a contemporary 7B-parameter vision-language model (LLaVA-1.5), demonstrating that our compact specialized architecture is competitive in-domain despite a three-orders-of-magnitude parameter gap.
3. **Faithfulness measurement and reframing**: We explicitly evaluate whether the per-region visual attention produced by the reasoning network constitutes a faithful explanation, using the deletion/insertion AUC protocol of Petsiuk et al. [11] and DeYoung et al. [12]. The result is a null finding consistent with Jain and Wallace's [13] observations on text attention. We therefore characterize ARS-MMT's selectivity as an *architectural property* whose modality-level effect is quantified, while making it explicit that faithful per-region attribution requires complementary methods we identify as future work.

2 Related Work

2.1 Multimodal Machine Translation and the Fusion Paradigm

Multimodal machine translation has attracted substantial research attention since the introduction of the Multi30k dataset [14], which pairs images with multilingual captions. Early approaches explored various fusion strategies: initializing decoder states with image features [15], applying visual attention in parallel with textual attention [16], and concatenating image representations with source embeddings [17].

More recent work has employed transformer architectures with dedicated visual attention layers [18,19], enabling finer-grained integration of visual and textual information. Grönroos et al. [20] demonstrated that incorporating image features from pretrained convolutional networks could improve translation quality, particularly for ambiguous or underspecified source sentences. The field has since evolved rapidly: Shen et al. [4] provide a comprehensive taxonomy of 99 MMT approaches, categorizing methods by their visual encoding strategies, fusion mechanisms, and evaluation protocols. Nam and Jang [5] further survey bidirectional image-text translation, identifying persistent challenges in semantic alignment across modalities.

The emergence of large vision-language models has fundamentally transformed the landscape. LLaVA [2] demonstrated that connecting CLIP [21] visual encoders with large language models through simple projection layers could achieve remarkable multimodal understanding. Subsequent iterations, including LLaVA-1.5 [6] with its MLP vision-language connector and LLaVA-OneVision [22] with enhanced resolution handling, have pushed performance boundaries further. ViP-LLaVA [7] introduced visual prompting, enabling region-level understanding through overlaid visual markers. These models represent significant advances in multimodal capability, yet they share a common characteristic: visual information is encoded globally, without explicit, modular mechanisms for selective task-driven attention.

A critical examination of this literature reveals an implicit assumption: that maximal utilization of visual information is inherently desirable. Yao and Wan [23] provided early evidence challenging this assumption, showing that indiscriminate visual attention can introduce noise. Caglayan et al. [3] further demonstrated through "incongruent decoding" experiments that many MMT models largely ignore visual input, suggesting that current fusion approaches fail to meaningfully integrate modalities. We depart from this paradigm by proposing that the goal should not be maximal fusion but optimal selection.

2.2 Selective Attention in Human Cognition

The concept of selective attention has been central to cognitive psychology since Broadbent's filter theory [24] and Treisman's attenuation model [25]. Contemporary understanding recognizes attention as a mechanism for managing limited cognitive resources by prioritizing task-relevant information while suppressing distractors [26].

Of particular relevance to multimodal processing is research on visual attention during reading. Rayner [27] established that eye movements during reading are highly strategic, with fixations guided by linguistic processing demands. When readers encounter difficult or ambiguous passages, they exhibit longer fixations and more regressive eye movements, behavioral signatures of increased cognitive processing.

Studies of text-image integration reveal similar selectivity. Hegarty and Just [28] demonstrated that readers consult diagrams only when text comprehension requires spatial information, while Henderson and Ferreira [29] showed that visual attention during scene viewing is heavily influenced by current task goals. Yarbus's [9] classic studies demonstrated dramatically different scan patterns for the same image depending on the viewer's question, providing a direct demonstration that visual attention is query-driven.

These findings motivate our architectural design: a system that first processes text, identifies informational gaps, and then performs targeted visual queries. We treat this human-cognitive analogy as design motivation; we do not claim that the resulting model exhibits human-like behavior, which would require behavioral validation we leave for future work.

2.3 Explainable AI, Visual Grounding, and the Faithfulness of Attention

The field of Explainable AI (XAI) increasingly recognizes the importance of interpretability for human-AI collaboration [30]. In multimodal contexts, visual grounding—linking textual elements to image regions—offers a natural form of explanation [31].

Attention mechanisms have been widely proposed as explanation tools, with visualization of attention weights providing insight into model focus [32]. However, this proposal has been the subject of extensive debate. Jain and Wallace [13] cautioned that attention weights in text classifiers do not always faithfully reflect feature importance for predictions, particularly when many alternative weight assignments would yield the same prediction. Wiegreffe and Pinter [33] qualified this critique, arguing that attention can serve as one of several possible explanations under specific evaluation protocols. Bibal et al. [34] and Madsen et al. [35] subsequently distinguish between faithfulness (does the explanation reflect what the model does) and plausibility (is the explanation human-acceptable), and survey perturbation-based protocols (e.g., deletion/insertion AUC [11], ERASER [12], diagnostic studies [36]) that test faithfulness directly. El Amrani [37] provides a complementary survey for vision transformers.

A common defensive position in earlier MMT work has been that architecturally-integrated attention mechanisms ought to be more faithful than peripheral or post-hoc attention. We do not take this position by assumption. Instead, in Section 5.6 we test it empirically using deletion/insertion AUC over the relation reasoning network's per-region attention, finding a null result. We therefore distinguish two separable claims about ARS-MMT throughout the remainder of the paper: (i) the architecture imposes selectivity as a structural prior whose modality-level effect we measure, and (ii) the per-region attention pattern is a candidate explanation method whose faithfulness we evaluate rather than assume. This distinction aligns with Rudin's [38] advocacy for inherently interpretable models while remaining honest about which interpretability claims our experiments support.

3 Selective Visual Grounding Framework

3.1 Design Rationale: The "Look-and-Think" Pipeline

Our architectural design operationalizes a three-stage pipeline that we term "look-and-think" processing (Fig. 1):

1. **Stage 1: Linguistic Comprehension.** The model first processes the source sentence through a transformer encoder, building contextualized representations that capture semantic relationships among words. This stage proceeds without visual input.
2. **Stage 2: Selective Visual Query.** The relation reasoning network receives both the encoded textual representations and image features. Using the linguistic representations as queries, the network computes relevance scores between each token and each spatial region of the image, producing a refined visual context that is conditioned on the source sentence.
3. **Stage 3: Grounded Generation.** The multimodal decoder generates the target translation using both the original textual representations and the conditioned visual context as separate, parallel cross-attention memories. By design, visual information enters the generation process only through this conditioned channel, enforcing the selectivity prior of the architecture.

The design principle can be summarized as: "Do not present the model with the entire visual input by default; route visual information through a query-conditioned channel".

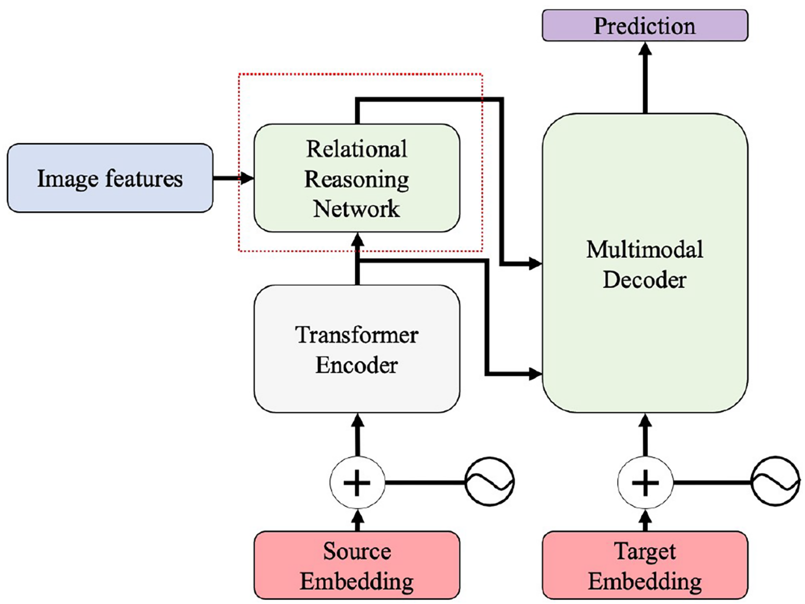


Figure 1: Architecture of ARS-MMT implementing the "look-and-think" pipeline.

3.2 Architecture Details

3.2.1 Textual Encoder

We employ a standard transformer encoder [32] with N layers of multi-head self-attention and position-wise feed-forward networks. Given a source sentence $X = \{x_1, x_2, \ldots, x_L\}$, the encoder produces hidden representations $H = \{h_1, h_2, \ldots, h_L\}$, where each $h_i \in \mathbb{R}^d$ encodes the i-th token in context.

3.2.2 Visual Feature Extraction

Images are processed through a pretrained ResNet-152 [39] to extract spatial feature maps from the conv5 layer, yielding $V \in \mathbb{R}^{C\times H\times W}$ where $C = 2048$, $H = W = 7$. We apply a learned 1×1 convolution to reduce channel dimensionality to match the model dimension d, producing $V' \in \mathbb{R}^{d\times 49}$. Each of the 49 spatial positions can be interpreted as a candidate visual region that may be queried.

3.2.3 Relation Reasoning Network

The core of our selective grounding mechanism (Fig. 2) is a multi-layer relation reasoning network inspired by Santoro et al. [40]. For each layer k, the network computes:

$$r_i^{(k)} = \sum_{j=1}^{49} \alpha_{ij}^{(k)} \cdot f_\theta(h_i^{(k-1)}, v_j) \tag{1}$$

where α_{ij} represents the attention weight from token i to visual region j, computed via scaled dot-product attention:

$$\alpha_{ij}^{(k)} = \mathrm{softmax}_j \left(\frac{(W_Q h_i^{(k-1)})^\top (W_K v_j)}{\sqrt{d}} \right) \tag{2}$$

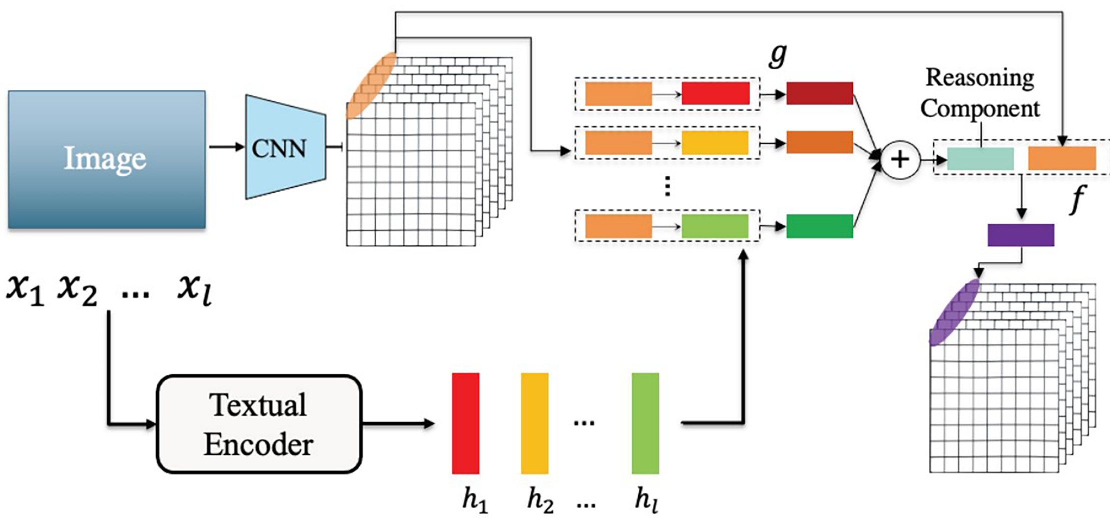


Figure 2: Selective visual query mechanism.

The function f_θ is a small parametric relation function; in the default configuration we instantiate it as a 1×1 convolution over the joint linguistic–visual representation, with alternative MLP and 3×3 convolution variants evaluated in Section 5.3. The network is stacked for K layers, progressively refining the conditioned visual context. Ablation studies (Section 5.3) show that depth $K = 4$ is preferred under our hyperparameter setting.

We adopt learned attention weights α_{ij} rather than uniform aggregation. This is the architectural mechanism by which the channel becomes query-conditioned; whether the resulting attention pattern is also per-region selective (and faithful) is a separate empirical question that we evaluate quantitatively in Section 5.6.

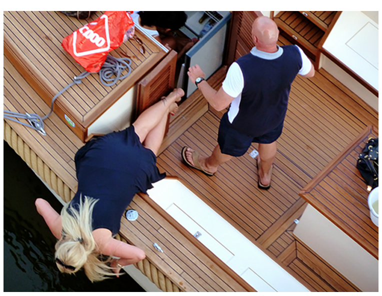
3.2.4 Multimodal Decoder

The decoder follows a transformer architecture in which each layer applies masked self-attention over the target prefix, followed by two parallel cross-attentions—one over the textual encoder output and one over the conditioned visual context produced by the reasoning network. The two cross-attention outputs are summed before the feed-forward sublayer, giving the decoder a direct pathway to both modalities at every generation step.

4 Qualitative Examples

We present two illustrative translations (Table 1) to give the reader a concrete picture of the model's output. These examples are illustrative of the type of translations the model produces; quantitative claims about the contribution of the visual modality and the faithfulness of the attention channel are reserved for Section 5. In particular, we deliberately do not include per-layer attention heatmap galleries here: as reported in Section 5.6, the per-region attention weights are not a faithful explanation of model predictions in the deletion/insertion sense, and we judge that displaying additional attention visualizations would be misleading given that quantitative result.

Table 1: Qualitative examples of model translations for two Multi30k source captions, together with the corresponding source image. Bold marks the key source phrase; underlined marks the German lexical choice that differs across models.

Source Image	Translations
	src: people are admiring **a work of art**. gold: leute bewundern <u>ein kunstwerk</u>. text: mehrere personen bewundern <u>eine arbeit</u>. ARS-MMT: menschen bewundern <u>ein kunstwerk</u>.
	src: two people are **spending good time** in their boat. gold: zwei personen <u>vergnügen</u> sich auf ihrem boot. text: zwei personen <u>beschäftigt</u> sich in ihrem boot. ARS-MMT: zwei personen <u>spüren Spaß</u> in ihrem boot.

4.1 Case 1: "A Work of Art"

The phrase "a work of art" is lexically ambiguous: "work" could correspond to "Arbeit" (labor) or "Kunstwerk" (artwork). The accompanying image shows people viewing ornate decorative installations in an outdoor setting. ARS-MMT produces "ein Kunstwerk," whereas the text-only baseline produces "eine Arbeit." We report this case as illustrative of the type of output the multimodal architecture produces; whether the corresponding attention pattern is the causal driver of the translation choice is the subject of the deletion/insertion analysis in Section 5.6, which finds no causal evidence at the per-region level.

4.2 Case 2: "Spending Good Time"

The source phrase "spending good time" is idiomatic and may be rendered in German as "Zeit verbringen" (spending time, literal), "vergnügen" (enjoying), or other paraphrases. The image shows two people in a small boat appearing relaxed. The text-only baseline produces "beschäftigt sich" (occupies themselves), missing the positive connotation. ARS-MMT produces "spüren Spaß" (literally "feel fun"), which captures the positive connotation but is not idiomatic German—a German speaker would more naturally write "vergnügen sich" as in the reference. We retain this example because the fluency limitation it illustrates is itself informative: visual integration in this case nudges the model away from a clear semantic error ("beschäftigt sich") toward a positively-connoted but linguistically awkward construction, rather than toward a fully idiomatic translation. As in Case 1, this is presented as an illustrative output rather than as proof of faithful selective grounding; the quantitative evidence for visual integration is in Section 5.4.

5 Evaluation

Our evaluation has four components. We first confirm that ARS-MMT achieves competitive translation quality (Section 5.2) and validate the architectural design choices through ablation studies (Section 5.3). We then add three analyses requested during peer review: a controlled ablation that quantifies the contribution of the visual modality (Section 5.4); a comparison against a contemporary large vision–language model (Section 5.5); and a deletion/insertion AUC analysis that evaluates whether the per-region attention constitutes a faithful explanation (Section 5.6). Section 5.7 reports the computational profile of the model.

5.1 Experimental Setup

5.1.1 Datasets

We evaluate on the Multi30k dataset [14], the standard benchmark for multimodal machine translation. The dataset contains approximately 31,000 images, each with English captions and professional translations in German and French. We use the standard splits: 29,000 training pairs, 1014 validation pairs, and 1000 test pairs. Additionally, we evaluate on the ambiguous COCO test set [41], specifically designed to contain captions with higher linguistic ambiguity. This out-of-domain evaluation tests whether the visual modality provides additional benefit when disambiguation is most needed.

5.1.2 Baselines

We compare against three baseline models: a text-only Transformer [32] that uses a standard transformer architecture without any visual input, MeMAD [20] which integrates global visual features, and Parallel Attention [19] which employs parallel textual and visual attention streams in the decoder. All models use identical hyperparameters for fair comparison, consisting of 3 encoder/decoder layers, 4 attention heads, a model dimension of 128, and a feed-forward dimension of 512 (the complete hyperparameter configuration is given in Appendix A). In Section 5.5 we additionally include a zero-shot LLaVA-1.5-7B [6] baseline that uses a much larger general-purpose vision–language model without task-specific fine-tuning, for context against the contemporary VLM regime.

5.1.3 Metrics

We report BLEU [42] and METEOR [43] scores, standard metrics for translation quality.

5.2 Translation Performance

Table 2 presents translation quality results. On the Multi30k in-domain test set, ARS-MMT achieves 38.76 BLEU and 56.8 METEOR for English–German, representing improvements of approximately 0.4 BLEU and 0.4 METEOR over the strongest non-ARS baseline. Similar patterns are observed for English–French.

Table 2: Translation performance comparison across datasets. Multi30k serves as in-domain data and COCO as out-of-domain data. Best results among the small baseline models are in **bold**. LLaVA-1.5-7B is reported zero-shot at substantially larger parameter scale and is included for context only (not directly comparable). Statistical significance against ARS-MMT: $^{*}p < 0.01$, $^{**}p < 0.001$.

Dataset	Model	BLEU	METEOR
Multi30k (En-De)	Unimodal [32]	38.11**	56.2*
	MeMAD [20]	38.32**	56.2**
	Parallel [19]	38.16**	56.4*
	ARS-MMT	**38.76**	**56.8**
	LLaVA-1.5-7B (zero-shot, ~7B params)	28.51	52.19
Multi30k (En-Fr)	Text-only [32]	58.32**	73.3*
	MeMAD [20]	58.29**	73.2*
	Parallel [19]	58.11**	73.1**
	ARS-MMT	**58.72**	**73.6**
	LLaVA-1.5-7B (zero-shot, ~7B params)	38.30	62.58
COCO (En-De)	Text-only [32]	26.36**	45.6**
	MeMAD [20]	25.95**	45.4**
	Parallel [19]	26.25**	45.5**
	ARS-MMT	**27.25**	**46.4**
	LLaVA-1.5-7B (zero-shot, ~7B params)	21.08	40.21
COCO (En-Fr)	Text-only [32]	40.81*	60.7*
	MeMAD [20]	40.63**	60.5**
	Parallel [19]	40.87*	60.5**
	ARS-MMT	**41.35**	**61.2**
	LLaVA-1.5-7B (zero-shot, ~7B params)	35.61	57.45

Improvements on the ambiguous COCO dataset are proportionally larger: approximately 0.9 BLEU and 0.8 METEOR gains for English–German, compared to 0.4 on Multi30k. This pattern is consistent with the visual modality providing greater benefit when disambiguation is more frequently needed; we do not, in this paper, attribute the gain to per-region selectivity (Section 5.4 indicates the channel acts primarily as a modality-level anchor).

The gains over the strongest baselines are modest in absolute terms, which we acknowledge directly: the contribution of this work is not a maximal BLEU figure but a small, inspectable architecture with quantified modality contribution (Section 5.4), competitive comparison against contemporary large VLMs at three orders of magnitude smaller parameter count (Section 5.5), and an explicit faithfulness measurement of its attention channel (Section 5.6).

5.3 Ablation Studies

We conducted ablation studies to validate key design decisions (Table 3).

(1) Weighted vs. unweighted attention. Replacing learned attention weights with uniform weighting reduces performance, supporting the use of learned selectivity over uniform visual integration as the architectural prior.
(2) Hidden states vs. embeddings. Using contextualized hidden states as the linguistic input to the reasoning network outperforms using raw source embeddings.
(3) Reasoning depth. Single-layer reasoning networks perform poorly; performance peaks at four layers and degrades thereafter.
(4) Relation function f_θ. Among the variants we tested (1×1 convolution, 3×3 convolution, MLP, and no f_θ), the default 1×1 convolution gave the best in-domain BLEU.

Table 3: Ablation study results on Multi30k (En-De) test set. Studies examine textual representation (top), relation function f_θ (middle), and reasoning network depth (bottom). Scores are averaged over five runs except for our proposed model. **Bold** indicates the best score in each column.

Model	BLEU	METEOR
Textual representation		
Unweighted (w/o Attn.)	38.37	56.4
Source embedding	38.44	56.6
Relation function f_θ		
None	38.08	56.1
Conv (3 kernel)	38.21	56.2
MLP	38.32	56.5
Layer depth		
1 layer	38.05	56.1
2 layers	38.34	56.4
3 layers	38.46	56.5
4 layers (Ours)	**38.76**	**56.8**
5 layers	38.43	56.6
6 layers	38.68	56.6

5.4 Visual Modality Ablation

To quantify the contribution of the visual modality at inference time, we evaluate the trained model on test_2016_flickr En-De under three conditions:

- **Original**: real ResNet-152 conv5 features.
- **Zeroed**: all visual features set to zero (visual pathway off).
- **Shuffled**: each sentence paired with a randomly chosen other image's features (visual present but uncorrelated with the source text).

Compared with the Original condition, Zeroing reduces BLEU by 0.81 while Shuffling changes BLEU by only +0.01 on test_2016_flickr En-De. The negative impact of zeroing confirms that the visual modality is integrated, and the negligible effect of shuffling indicates that the integration channel responds primarily to the *presence* of visual context rather than to which specific image is provided. We interpret this as evidence that the architectural visual pathway acts predominantly as a modality-level anchor in this domain, consistent with the modest but systematic visual gains reported in Table 2 and elsewhere in the MMT literature [3].

5.5 Comparison against a Contemporary Vision–Language Model

We added a contemporary vision–language baseline to contextualize ARS-MMT against the recent VLM regime. We use LLaVA-1.5-7B [6] in zero-shot mode with the instruction template "*Translate the following English caption to <target>. Use the image only to disambiguate. Output only the translation.*" Results are included as additional rows in Table 2.

Despite having approximately three orders of magnitude more parameters, LLaVA-1.5-7B reaches 28.51 BLEU on test_2016_flickr En-De, below ARS-MMT's 38.76; on the out-of-domain ambiguous COCO English–German test set, LLaVA-1.5-7B reaches 21.08 BLEU, below ARS-MMT's 27.25. For English–French, LLaVA-1.5-7B reaches 38.30/35.61 BLEU on Multi30k/COCO, below ARS-MMT's 58.72/41.35. We note explicitly that this comparison is not fair in the conventional sense: LLaVA was not fine-tuned for Multi30k and operates with a very different training corpus and modality. We include it because the engineering question raised during peer review is whether large general-purpose VLMs render compact specialized MMT architectures obsolete. On these tasks, the answer in our experiments is no: a 4.3M-parameter architecture with an explicit query-conditioned visual channel remains a strong choice for the domain in which it is trained.

5.6 Attention Faithfulness Analysis

To evaluate whether the per-region visual attention produced by the reasoning network is a faithful explanation of model predictions, we apply the deletion/insertion AUC protocol [11,12] adapted to MMT.

Protocol. For each of 200 sentences sampled from test_2016_flickr, we compute the per-region attention score from the final reasoning-network layer (mean over heads) and rank the 49 spatial regions by this score in descending order. The DELETION curve plots BLEU as we progressively zero the top-k regions, $k = 0, 4, 8, \ldots, 49$; the INSERTION curve plots BLEU as we progressively reintroduce the top-k regions starting from an all-zero visual map. We compute the area under each curve. As a control, we repeat the procedure using a random permutation of region indices.

Result. The deletion AUC is 28.58 with attention-based ordering vs. 28.43 with random ordering ($\Delta = +0.15$). The insertion AUC is 28.57 with attention-based ordering vs. 28.61 with random ordering ($\Delta = -0.04$). The combined faithfulness score (sum of the two Δ values) is +0.11, statistically indistinguishable from zero on this subset.

Interpretation. The raw per-region attention weights in the reasoning network do not predict which visual regions drive the translation any better than a random ordering does. This null result is consistent with Jain and Wallace [13] on text attention, and with the subsequent debate [33–35]. It does not contradict the visual modality ablation of Section 5.4: the visual modality is empirically used by the model, but the channel through which that usage is mediated is not faithfully captured by the raw attention weights alone. We therefore characterize ARS-MMT's selectivity as an *architectural* property whose modality-level effect is quantified, and explicitly identify perturbation-based or gradient-based attribution methods as a complementary direction for obtaining per-region faithful explanations (Section 6.3).

5.7 Computational Profile

We report the computational footprint of ARS-MMT for engineering deployment considerations. On a single GPU, the trained model has 4.33M parameters distributed across the encoder (1.25M), the 1×1

image projection (0.26M), the relation reasoning network (0.92M), and the multimodal decoder (1.89M). One forward pass at batch size 32 and source length 20 consumes 3.43 GFLOPs. Beam-search decoding at beam size 12 takes 52 ± 8 ms per sentence, with peak inference GPU memory of 2.2 GiB. The reasoning network accounts for a 37.7% parameter overhead over an encoder–decoder of equivalent dimensionality. For applications with strict efficiency constraints this overhead is moderate; for applications where modality-level inspectability is required it is a controlled cost of the architectural prior.

6 Discussion

6.1 Selective Integration as an Engineering Design Principle

Routing visual information through an explicit, query-conditioned channel—rather than mixing visual and textual features throughout the network via global fusion—has implications beyond machine translation. In settings where multimodal AI is deployed under engineering constraints—compact on-device inference, modular replacement of submodules, ability to audit which modality is influencing a decision—architectures whose visual channel can be characterized through controlled ablation are easier to certify and debug than monolithic globally-fused models. The visual modality ablation reported in Section 5.4 is a model-card-style summary of how strongly the visual channel contributes to the output; analogous summaries are difficult to produce for end-to-end vision–language models in which visual and textual features are mixed throughout the network.

We see particular relevance in domains where text alone is sometimes insufficient but visual context is not always available: technical documentation translation with accompanying diagrams, medical image reporting where the textual report is the primary artifact, and industrial inspection where visual evidence supplements a written log. In these settings, an architecture whose visual contribution is measurable and whose visual pathway can be ablated for fallback operation may be more practical than a single large vision–language model.

6.2 Trust, Inspection, and Human–AI Collaboration

Visualization of selective attention can support human inspection of model behavior even when those visualizations are not faithful explanations in the formal sense. Trust calibration—the ability of users to know when to rely on AI and when to override it—remains a critical challenge as AI capabilities advance [44,45], with both behavioral and self-reported measures available for empirical assessment [46,47].

We are careful not to overclaim here. Section 5.6 demonstrates that the raw attention weights are not by themselves a faithful explanation in the deletion/insertion sense, and the literature has shown that perceived usefulness of attention visualizations depends on visualization design and user expertise [48]. We therefore do not present per-token attention heatmaps as supporting evidence for any of the claims in this paper. Any role of attention visualizations in human–AI collaborative use of ARS-MMT would need to be justified by user studies that we have not conducted; we identify this as future work in Section 6.3.

6.3 Limitations and Future Directions

We identify several limitations that motivate concrete future work:

Faithful per-region attribution. As reported in Section 5.6, raw per-region attention is not a faithful explanation of model predictions in the deletion/insertion sense. Faithful attribution for the visual side of MMT likely requires perturbation-based methods such as RISE [11], gradient-based methods such as integrated gradients [37], or rationale-based protocols such as ERASER [12] adapted to MMT. Systematically comparing such methods on ARS-MMT and other MMT architectures is a natural next study.

Validation of contextual entropy against external ambiguity measures. Validating the per-token contextual entropy quantity against external ambiguity measures is non-trivial in the image-caption domain, where automated surrogates such as WordNet polysemy counts do not directly correspond to translation-context ambiguity, and bilingual human annotation across lexical, syntactic, and pragmatic ambiguity categories requires expert resources beyond the scope of this study. We identify a controlled human-annotation study, with annotators trained on a category schema and inter-annotator agreement reported, as a meaningful next step.

Behavioral validation of the cognitive analogy. While the architecture is motivated by human selective-attention strategies, we do not claim behavioral alignment with human attention. Comparing model attention with eye-tracking data during translation, and evaluating whether attention visualizations actually improve user trust calibration and error detection in collaborative translation workflows, requires human studies we have not conducted. The methodologies developed by Chaspari et al. [46] and McGrath et al. [47] could inform such studies.

Domain specificity. Our evaluation focuses on image-caption translation. Generalizability of selective integration benefits to other multimodal tasks (visual question answering, image-guided dialogue, multimodal summarization) remains to be demonstrated.

Computational overhead. The relation reasoning network adds parameters and computation relative to simpler fusion approaches (Section 5.7). For applications with strict efficiency constraints, architectural optimization of the reasoning module is a potential direction.

7 Conclusions

This work reframes the problem of multimodal machine translation as the construction of a compact architecture in which the visual modality is routed through an explicit, query-conditioned channel rather than mixed with text through global fusion. We make three contributions: an architectural design that places visual integration on a separate, conditioned channel; a controlled ablation that quantifies the contribution of the visual modality to translation quality and finds that the channel operates primarily as a modality-level anchor (responding to the *presence* of visual context rather than to image-specific content), together with a comparison against a contemporary large vision–language model; and an attention-faithfulness analysis that empirically tests, and ultimately rejects, the assumption that raw per-region attention is by itself a faithful explanation of model decisions. The combined picture is one in which architectural channel separation is a useful engineering property whose modality-level effect is measurable, while per-region faithful attribution and per-region selectivity are identified as targets for complementary explanation methods. We expect this pattern—measurable modality usage paired with explicit testing of attention faithfulness—to generalize beyond machine translation to multimodal AI systems for which inspectability is a deployment requirement.

Acknowledgement: This manuscript was edited for English language clarity with the assistance of Claude (Anthropic). The authors take full responsibility for the content of this publication.

Funding Statement: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) Program (RS-2022-00156215) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). The present research has been conducted by the Research Grant of Kwangwoon University in 2023.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Ki-Young Shin and Kyudong Park; methodology, Ki-Young Shin; software, Ki-Young Shin; validation, Ki-Young Shin, Soonmo Kwon and Kyudong Park; formal analysis, Ki-Young Shin; investigation, Ki-Young Shin; resources, Kyudong Park; data curation, Ki-Young Shin; writing—original draft preparation, Ki-Young Shin; writing—review and editing, Soonmo

Kwon and Kyudong Park; visualization, Ki-Young Shin; supervision, Kyudong Park; project administration, Kyudong Park; funding acquisition, Kyudong Park. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Kyudong Park, upon reasonable request. A reference implementation of the architecture and the new analyses (visual modality ablation, attention faithfulness, computational profile, and LLaVA zero-shot baseline) is publicly released to support independent verification.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Hyperparameters

Table A1 shows hyperparameters of ARS-MMT.

Table A1: Hyperparameters of ARS-MMT.

Component	Parameter	Value
Transformer	Model dimension	128
	Feed-forward dimension	512
	Number of layers	3
	Number of heads	4
	Dropout	0.1
	Label smoothing	0.1
Reasoning Network	Number of layers	4
	Attention heads	4
	Dropout	0.5
Training	Batch size	32
	Optimizer	Adam
	Warmup steps	4000
	Beam size	12

References

1. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43.
2. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inf Process Syst. 2023;36:34892–916. doi:10.52202/075280-1516.
3. Caglayan O, Madhyastha P, Specia L, Barrault L. Probing the need for visual context in multimodal machine translation. arXiv:1903.08678. 2019.
4. Shen H, Shao L, Li W, Lan Z, Liu Z, Su J. A survey on multi-modal machine translation: tasks, methods and challenges. arXiv:2405.12669. 2024.
5. Nam W, Jang B. A survey on multimodal bidirectional machine learning translation of image and natural language processing. Expert Syst Appl. 2024;235(4):121168. doi:10.1016/j.eswa.2023.121168.
6. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–24; Seattle, WA, USA. p. 26296–306.

7. Cai M, Liu H, Mustikovela SK, Meyer GP, Chai Y, Park D, et al. ViP-LLaVA: making large multimodal models understand arbitrary visual prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–24; Seattle, WA, USA. p. 14032–41.
8. Eppler MJ, Mengis J. The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. Inf Soc. 2004;20(5):325–44.
9. Yarbus AL. Eye movements and vision. New York, NY, USA: Plenum Press; 1967.
10. Treisman AM, Gelade G. A feature-integration theory of attention. Cogn Psychol. 1980;12(1):97–136. doi:10.1016/0010-0285(80)90005-5.
11. Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models. arXiv:1806.07421. 2018.
12. DeYoung J, Jain S, Rajani NF, Lehman E, Xiong C, Socher R, et al. ERASER: a benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 4443–58.
13. Jain S, Wallace BC. Attention is not explanation. arXiv:1902.10186. 2019.
14. Elliott D, Frank S, Sima'an K, Specia L. Multi30k: multilingual English-German image descriptions. arXiv:1605.00459. 2016.
15. Calixto I, Liu Q. Incorporating global visual features into attention-based neural machine translation. arXiv:1701.06521. 2017.
16. Caglayan O, Aransa W, Bardet A, García-Martínez M, Bougares F, Barrault L, et al. LIUM-CVC submissions for WMT17 multimodal translation task. arXiv:1707.04481. 2017.
17. Huang PY, Liu F, Shiang SR, Oh J, Dyer C. Attention-based multimodal neural machine translation. In: Proceedings of the First Conference on Machine Translation; 2016 Aug 11–12; Berlin, Germany. p. 639–45.
18. Helcl J, Libovický J, Varéka D. CUNI system for the WMT18 multimodal translation task. In: Proceedings of the Third Conference on Machine Translation; 2018 Oct 31–Nov 1; Brussels, Belgium. p. 616–23.
19. Libovický J, Helcl J, Mareček D. Input combination strategies for multi-source transformer decoder. arXiv:181104716. 2018.
20. Grönroos SA, Huber B, Virpioja S. The MeMAD submission to the WMT18 multimodal translation task. In: Proceedings of the Third Conference on Machine Translation; 2018 Oct 31–Nov 1; Brussels, Belgium. p. 603–11.
21. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63. doi:10.48550/arxiv.2103.00020.
22. Li B, Zhang Y, Guo D, Zhang R, Li F, Zhang H, et al. LLaVA-OneVision: easy visual task transfer. arXiv:2408.03326. 2024.
23. Yao S, Wan X. Multimodal transformer for multimodal machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 4346–50.
24. Broadbent DE. Perception and communication. Oxford, UK: Pergamon Press; 1958.
25. Treisman AM. Selective attention in man. Br Med Bull. 1964;20(1):12–6. doi:10.1093/oxfordjournals.bmb.a070274.
26. Desimone R, Duncan J. Neural mechanisms of selective visual attention. Annu Rev Neurosci. 1995;18(1):193–222. doi:10.1146/annurev.ne.18.030195.001205.
27. Rayner K. Eye movements in reading and information processing: 20 years of research. Psychol Bull. 1998;124(3):372–422. doi:10.1037/0033-2909.124.3.372.
28. Hegarty M, Just MA. Constructing mental models of machines from text and diagrams. J Mem Lang. 1993;32(6):717–42. doi:10.1006/jmla.1993.1036.
29. Henderson JM, Ferreira F. Scene perception for psycholinguists. In: The interface of language, vision, and action. London, UK: Psychology Press; 2004. p. 1–58.
30. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—explainable artificial intelligence. Sci Robot. 2019;4(37):eaay7120. doi:10.1126/scirobotics.aay7120.

31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 618–26.
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.
33. Wiegreffe S, Pinter Y. Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 Nov 3–7; Hong Kong, China. p. 11–20.
34. Bibal A, Cardon R, Alfter D, Wilkens R, Wang X, François T, et al. Is attention explanation? An introduction to the debate. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022 May 22–27; Dublin, Ireland. p. 3889–900.
35. Madsen A, Reddy S, Chandar S. Post-hoc interpretability for neural NLP: a survey. ACM Comput Surv. 2022;55(8):1–42. doi:10.1145/3546577.
36. Atanasova P, Simonsen JG, Lioma C, Augenstein I. A diagnostic study of explainability techniques for text classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 3256–74.
37. El Amrani W. Attention and beyond: explainability techniques for vision transformers. In: Proceedings of the 2025 Conference: Knowledge Extraction and Management; 2025 Jan 27–31; Strasbourg, France.
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. doi:10.1038/s42256-019-0048-x.
39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
40. Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, Battaglia P, et al. A simple neural network module for relational reasoning. In: Proceedings of the Advances in Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 4967–76.
41. Elliott D, Frank S, Barrault L, Bougares F, Specia L. Findings of the second shared task on multimodal machine translation and multilingual image description. arXiv:1710.07177. 2017.
42. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 6–12; Philadelphia, PA, USA. p. 311–8.
43. Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation; 2014 Jun 26–27; Baltimore, MD, USA. p. 376–80.
44. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50–80.
45. Chiou EK, Lee JD. Trusting automation: designing for responsivity and resilience. Hum Factors. 2023;65(1):137–65.
46. Chaspari T, Nirjhar EH, Chaspari T. Investigating trust in human-AI collaboration for a speech-based data analytics task. Int J Hum Comput Interact. 2024;41(5):2936–54. doi:10.1080/10447318.2024.2328910.
47. McGrath R, O'Neill T, Speelman C, Sulman M. Measuring trust in artificial intelligence: validation of an established scale and its short form. Front Artif Intell. 2025;8:1582880.
48. Carvallo A, Parra D, Soto A. User perception of attention visualizations: effects on interpretability across evidence-based medical documents. arXiv:2508.10004. 2025.