Article

Not peer-reviewed version

Cross-Lingual Semantic Alignment in Large Language Models via Context-Aware Training

Daniel Tang *

Posted Date: 13 March 2025

doi: 10.20944/preprints202503.0935.v1

Keywords: Machine Translation; Large Language Models

Article

Cross-Lingual Semantic Alignment in Large Language Models via Context-Aware Training

Daniel Tang

University of Southern Mississippi; aya.akka@edu.suezuni.edu.eg

Abstract: This paper introduces Context-Aware Cross-Modal Alignment Training (CACMAT), a novel multi-stage training paradigm to enhance translation capabilities of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). Current LLM translation models often struggle with contextual nuances and cross-lingual semantic alignment. CACMAT addresses this by incorporating three stages: secondary pre-training on target language monolingual data, continual pre-training with a contextual contrastive loss using Interlinear Text Format (ITF) data to improve cross-lingual alignment, and supervised fine-tuning on parallel translation datasets. Experiments on FLORES-200 and WMT datasets demonstrate that CACMAT significantly outperforms baseline models and achieves competitive results against state-of-the-art systems, as validated by both BLEU scores and human evaluations. Ablation studies confirm the crucial role of the contextual contrastive alignment stage. The results highlight CACMAT as an effective approach for improving translation quality by explicitly enhancing cross-lingual and cross-modal semantic alignment in LLMs and LVLMs.

Keywords: machine translation; large language models

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks, including text generation, question answering, and code completion [1]. Their emergent abilities in machine translation (MT) have also garnered significant attention, offering the potential to revolutionize how we bridge language barriers [2]. The continuous improvement of LLMs' capabilities is evident in their expanding applications and enhanced performance in diverse tasks [3]. Improving the translation capabilities of LLMs is of paramount importance due to the ever-increasing globalization and interconnectedness of our world. Effective and accurate machine translation facilitates seamless communication, knowledge sharing, and cultural exchange across linguistic boundaries, impacting various domains from international business and diplomacy to scientific collaboration and personal interactions. Moreover, enhancing LLM translation can democratize access to information and services for individuals who do not speak dominant languages, fostering inclusivity and equity in the digital age.

Despite the impressive progress, challenges remain in achieving human-level translation quality with LLMs, particularly in nuanced and context-rich scenarios [4]. Current LLM translation models often struggle with capturing subtle semantic variations, cultural idioms, and contextual dependencies that are crucial for accurate and fluent translation. A key limitation lies in the models' inherent bias towards the languages prevalent in their training data, typically English, leading to performance disparities when translating to and from less-resourced languages or languages with distinct linguistic structures. Furthermore, the purely text-to-text nature of many LLM translation approaches can be insufficient in contexts where visual or multimodal information is readily available and relevant to meaning. For instance, translating image captions or descriptions requires understanding the visual context to generate accurate and contextually appropriate translations [5]. Recent studies also explore vision representation compression to enhance the efficiency of video generation using large language models, highlighting the importance of visual context in language models [6]. This necessitates moving

doi:10.20944/preprints202503.0935.v1

beyond surface-level lexical transformations and developing methods that enable LLMs to achieve deeper cross-lingual and cross-modal semantic alignment.

Motivated by these challenges, this paper introduces **Context-Aware Cross-Modal Alignment Training (CACMAT)**, a novel training paradigm designed to significantly boost the translation capabilities of both Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). Our approach is driven by the hypothesis that enhancing a model's ability to understand and align contextual representations across languages and modalities is crucial for achieving high-quality translation, especially in complex and nuanced scenarios. We posit that current methods often fall short by treating translation as a mere sequence-to-sequence task, neglecting the rich contextual and multimodal cues that humans intuitively utilize in the translation process. Therefore, CACMAT explicitly focuses on imbuing models with a deeper, context-aware understanding of language for translation, moving beyond superficial lexical mappings. Furthermore, exploring visual in-context learning can offer valuable insights into improving the performance of large vision-language models, which is relevant to our cross-modal alignment approach [7].

The core of CACMAT lies in a multi-stage training strategy. First, for LVLMs, we perform **Multimodal Pre-training** on large-scale image-text datasets to establish foundational cross-modal understanding [5]. For LLMs, we continue with monolingual pre-training to strengthen target language comprehension. Second, we introduce the **Contextual Cross-Modal Alignment (CCMA)** stage, the centerpiece of our approach. Here, we train models with parallel, multimodal (or text-only for LLMs) data, employing a novel **contextual contrastive loss**. This loss function explicitly encourages the model to learn representations where semantically similar sentences across languages are embedded closer together in a shared representation space, while representations of dissimilar sentences are pushed further apart. This stage leverages techniques such as sentence embedding and contrastive learning frameworks to facilitate robust cross-lingual alignment. Finally, in the **Supervised Fine-tuning** stage, we fine-tune the models on standard parallel translation datasets using conventional translation loss functions. This fine-tuning benefits from the improved contextual and cross-modal alignment acquired during the CCMA stage, leading to enhanced translation quality. The importance of contextual understanding in language models is further highlighted by research into unraveling chaotic contexts through thread of thought [8].

To evaluate the efficacy of CACMAT, we conduct extensive experiments using established benchmarks for machine translation, including [9] and [10]. For LVLMs, we incorporate datasets with image-text pairs in multiple languages to assess performance in visually grounded translation tasks. We employ standard evaluation metrics such as BLEU score [11] and human evaluation to comprehensively assess translation quality, focusing on fluency, adequacy, and contextual appropriateness. Our preliminary results demonstrate that CACMAT significantly outperforms baseline models and achieves competitive performance against state-of-the-art translation systems, particularly in scenarios demanding contextual understanding and cross-modal integration. Specifically, we observe notable improvements in translation accuracy and fluency, especially for language pairs with significant linguistic divergence and in visually contextualized translation tasks.

In summary, this paper makes the following key contributions:

- We propose **Context-Aware Cross-Modal Alignment Training (CACMAT)**, a novel multi-stage training paradigm specifically designed to enhance the translation capabilities of LLMs and LVLMs by focusing on contextual and cross-modal semantic alignment.
- We introduce a **contextual contrastive loss function** within the CCMA training stage, which explicitly encourages models to learn aligned representations for semantically similar sentences across languages and modalities.
- We present comprehensive experimental results on standard machine translation benchmarks and visually grounded translation tasks, demonstrating the significant performance gains achieved by CACMAT over strong baselines and highlighting its effectiveness in improving translation quality, particularly in contextually rich scenarios.

2. Related Work

2.1. Machine Translation

Machine Translation (MT) has been a long-standing goal in artificial intelligence, aiming to automate the translation of text from one language to another. Early approaches to MT relied on rule-based systems and statistical methods [12], but the field has been revolutionized by the advent of neural networks, particularly with the introduction of Neural Machine Translation (NMT) [13]. NMT models, especially those based on the Transformer architecture [14], have achieved remarkable progress, significantly improving translation quality and fluency. Furthermore, research has explored improving cross-lingual transfer for multilingual question answering, which is related to the challenges in machine translation across different languages [15].

Google Research has been at the forefront of advancements in machine translation, developing the Google Neural Machine Translation (GNMT) system [16]. GNMT, a deep learning system, marked a significant step towards closing the gap between human and machine translation by employing an end-to-end approach and focusing on sentence-level translation. However, despite these advancements, challenges remain, particularly in ensuring the quality and reliability of MT systems across diverse domains and languages.

Recent research has focused on understanding the broader societal impacts of machine translation, especially in critical domains like medicine and law [17]. These studies highlight the need for careful evaluation and consideration of the ethical and practical implications of using MT in sensitive contexts. Comparative studies continue to evaluate the performance of different MT systems, including commercial systems like DeepL and Google Translate, particularly in specialized domains such as medical research [18]. These evaluations are crucial for understanding the strengths and weaknesses of current MT technologies and for guiding future research directions aimed at achieving more robust and reliable translation across all contexts and language pairs.

2.2. Large Language Models

Large Language Models (LLMs) have emerged as a transformative force in natural language processing, demonstrating exceptional capabilities across a wide spectrum of tasks, from text generation and understanding to more complex reasoning and problem-solving [19]. These models, typically based on the Transformer architecture [14], are characterized by their massive scale, trained on vast amounts of text data, enabling them to capture intricate patterns and nuances of human language. The multi-capabilities of LLMs have been a subject of recent study, further demonstrating their broad applicability [3]. State space models are also being explored in specialized vision tasks, such as insect pest classification, showing the diverse architectures being applied in the vision-language domain [20].

The rapid development and widespread adoption of LLMs have spurred significant research interest in understanding their potential and limitations across various domains. Surveys have been conducted to comprehensively examine the applications, challenges, limitations, and future prospects of LLMs, providing valuable insights into the current state and trajectory of this rapidly evolving field [21]. These surveys highlight both the remarkable achievements of LLMs and the open challenges that need to be addressed for their responsible and effective deployment. Moreover, visual in-context learning has been explored as a method to enhance the capabilities of large vision-language models, showing the importance of visual information in these models [7].

Beyond general capabilities, research is actively exploring the application of LLMs in specific scientific and professional domains. For instance, the use of LLMs in neurology research and practice is being investigated, with potential applications ranging from clinical decision support to accelerating scientific discovery [22]. Furthermore, the reliability and performance of LLMs in critical tasks like scholarly writing, particularly in generating accurate citations and references, are under scrutiny, emphasizing the need for careful evaluation and validation in high-stakes applications [23]. In the medical field, research is progressing on training medical large vision-language models with abnormal-aware feedback, indicating the growing specialization of LLMs in areas like medical imaging [24]. Editorials and reviews are also emerging, discussing the broader opportunities and challenges presented by LLMs in science and research, prompting critical discussions about their responsible integration into scientific workflows and practices [25]. Additionally, multimodal approaches using event transformers are being developed for tasks like image-guided story ending generation, showcasing the versatility of LLMs in handling multimodal data [26].

3. Method

In this section, we present a detailed exposition of the proposed Context-Aware Cross-Modal Alignment Training (CACMAT) paradigm. CACMAT is conceived as a three-stage training regimen meticulously designed to enhance the translation proficiencies of both Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). While the overarching objective is to achieve high-fidelity generative translation, CACMAT strategically integrates a discriminative learning component within its pivotal second stage. This integration is crucial for refining cross-lingual and cross-modal alignment, which we hypothesize is key to superior translation quality. The ensuing description outlines the training methodology applicable to both LLMs and LVLMs, with specific attention to the cross-modal dimensions pertinent to LVLMs.

3.1. Model Architecture: Transformer Foundation

Our methodology is grounded in the widely adopted Transformer architecture [14], which provides a robust framework for both LLM and LVLM implementations. The Transformer, characterized by its encoder-decoder structure, is inherently well-suited for sequence-to-sequence tasks such as machine translation. Specifically, the encoder processes the input sequence (source language text, and optionally visual features for LVLMs), transforming it into a rich contextualized representation. The decoder then leverages this representation to generate the output sequence in the target language.

For the LVLM variant, the encoder includes a dedicated visual encoder module. This visual encoder, such as ResNet or ViT, extracts salient visual features from the input image. These visual features are then fused with the textual encoder representations, enabling the LVLM to leverage both visual and textual cues during translation. The fusion mechanism is implemented using techniques like cross-attention or concatenation, based on the specific LVLM architecture.

However, it is crucial to emphasize that the core innovation of CACMAT resides in the training methodology itself, rather than being tightly coupled to a specific model architecture. The proposed training paradigm is designed to be largely model-agnostic and can be effectively applied to a diverse range of Transformer-based architectures, including but not limited to variations in encoder depth, decoder depth, attention mechanisms, and visual encoder choices. In our experimental validation, we employ a Transformer model configuration with a parameter scale analogous to Llama2-7B. This choice ensures a strong and contemporary baseline for rigorous comparative analysis, allowing us to isolate the performance gains attributable to the CACMAT training strategy.

3.2. Context-Aware Cross-Modal Alignment Training (CACMAT) Stages

The CACMAT paradigm is structured into three sequential training stages, each meticulously crafted to progressively enhance distinct facets of the model's translation capabilities. This staged approach allows for a focused and systematic optimization of the model's linguistic and cross-modal understanding, leading to improved translation performance.

3.2.1. Stage 1: Enhancing Target Language Proficiency via Secondary Pre-Training

The initial stage of CACMAT is dedicated to strengthening the model's inherent understanding of the target language. It is a well-documented observation that many pre-trained LLMs exhibit a linguistic bias towards English, primarily due to the disproportionate representation of English text in their training corpora. Consequently, their proficiency in other languages, particularly in terms of nuanced semantic comprehension and fluent generation, may be comparatively less developed. To mitigate this potential bottleneck, we undertake a secondary pre-training phase. This phase involves exposing the model to a substantial corpus of monolingual data exclusively in the target language (which, in the context of our experiments, is English).

This secondary pre-training stage leverages the standard language modeling objective, a cornerstone of LLM pre-training. The objective is to maximize the conditional likelihood of predicting the subsequent token in a sequence, given the preceding tokens. Mathematically, the loss function for this stage, denoted as $\mathcal{L}_{LM}$, can be formally expressed as:

$$\begin{aligned}\mathcal{L}_{LM}(\theta) &= -\mathbb{E}_{x\sim\mathcal{D}_{mono}} \log P(x|\theta) && (1)\\ &= -\mathbb{E}_{x\sim\mathcal{D}_{mono}} \sum_{t=1}^{|x|} \log P(x_t|x_1,\ldots,x_{t-1};\theta) && (2)\end{aligned}$$

where $\mathcal{D}_{mono}$ signifies the monolingual data corpus utilized for this stage, $x = (x_1, x_2, \ldots, x_{|x|})$ represents a sequence of tokens sampled from $\mathcal{D}_{mono}$, $x_1, \ldots, x_{t-1}$ denotes the sequence of tokens preceding the t-th token x_t, and θ collectively represents the trainable parameters of the model. By minimizing $\mathcal{L}_{LM}$, we effectively refine the model's internal language model, enhancing its capacity to generate target language text that is not only grammatically sound but also exhibits improved fluency and naturalness. This enhanced target language proficiency serves as a solid foundation for subsequent translation-specific training stages.

3.2.2. Stage 2: Contextual Cross-Modal Alignment (CCMA) via Contrastive Learning

The second stage, Contextual Cross-Modal Alignment (CCMA), constitutes the methodological core and primary innovation of our proposed CACMAT paradigm. This stage is explicitly designed to foster and refine the model's ability to establish robust semantic alignments across linguistic boundaries and, crucially for LVLMs, across modalities (textual and visual). We capitalize on the Interlinear Text Format (ITF) data, which uniquely provides word-level alignment annotations between source and target language sentence pairs. This fine-grained alignment information is invaluable for guiding the model to learn precise cross-lingual correspondences. To effectively leverage ITF data, we introduce a novel contextual contrastive loss function that encourages the learning of aligned representations.

The CCMA stage commences by generating sentence embeddings for both source and target language sentences derived from the ITF data. For a given source language sentence s and its corresponding target language translation t within the ITF data, we compute their respective sentence embeddings. Let $e_s = \text{Encoder}(s)$ denote the sentence embedding of the source sentence s, and $e_t = \text{Encoder}(t)$ represent the sentence embedding of the target sentence t. The function $\text{Encoder}(\cdot)$ symbolizes the sentence encoder, which is instantiated from the Transformer encoder component of our model. In practice, this sentence encoder can be realized by averaging the word embeddings produced by the Transformer encoder or by utilizing the [CLS] token representation.

Subsequently, we formulate a contextual contrastive loss, $\mathcal{L}_{CCMA}$, drawing inspiration from the established InfoNCE loss framework. For each aligned sentence pair (s,t) extracted from the ITF data, we designate the pair of their embeddings (e_s, e_t) as a *positive pair*. To construct *negative pairs*, we pair the source sentence embedding e_s with sentence embeddings $\{e_{t'}\}_{t'\neq t}$ derived from other randomly sampled target language sentences from within the same ITF dataset, ensuring that t' is not the correct translation of s. The contextual contrastive loss $\mathcal{L}_{CCMA}$ is then mathematically defined as follows:

$$\mathcal{L}_{CCMA}(\theta) = -\mathbb{E}_{(s,t)\sim\mathcal{D}_{ITF}} \log \frac{\text{sim}(e_s, e_t)}{\text{sim}(e_s, e_t) + \sum_{t'\neq t} \text{sim}(e_s, e_{t'})} \tag{3}$$

where $\mathcal{D}_{ITF}$ represents the ITF data corpus, $\text{sim}(u, v)$ is a similarity metric that quantifies the similarity between two sentence embeddings u and v. A commonly used and effective similarity function is cosine similarity, defined as:

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \|v\|} \tag{4}$$

The temperature hyperparameter τ (typically omitted here for simplicity but often included in InfoNCE variants) is intentionally set to 1 in our formulation for simplicity and empirical effectiveness, implicitly controlling the concentration level of the probability distribution. The loss function $\mathcal{L}_{CCMA}$ operates by encouraging the model to maximize the similarity score between embeddings of correctly aligned sentence pairs (e_s, e_t), while concurrently minimizing the similarity scores between embeddings of unaligned pairs $(e_s, e_{t'})$. Through iterative training with $\mathcal{L}_{CCMA}$, the model progressively learns to structure a shared, semantically meaningful embedding space wherein sentences that are translation equivalents across languages are mapped to proximal regions. This process fundamentally enhances the model's cross-lingual alignment capabilities at the sentence representation level.

For LVLMs, the CCMA stage can be further enriched by incorporating visual context. When visual information v is available and pertinent to a sentence pair (s, t, v), we can condition the generation of sentence embeddings e_s and e_t on the visual embedding $e_v = \text{VisualEncoder}(v)$. This visual embedding e_v is obtained by processing the visual input v through the dedicated visual encoder $\text{VisualEncoder}(\cdot)$. Conditioning can be achieved through various attention mechanisms or fusion layers, allowing the LVLM to learn context-aware cross-modal alignments, where the semantic similarity between sentences is evaluated not only based on their textual content but also in relation to their shared visual context.

3.2.3. Stage 3: Supervised Fine-Tuning for Translation Task

The concluding stage of CACMAT is supervised fine-tuning, wherein the model is explicitly trained for the machine translation task using parallel translation datasets. In this phase, we employ standard parallel corpora, denoted as $\mathcal{D}_{parallel}$, consisting of pairs of source and target language sentences that represent direct translations of each other. The model's objective in this stage is to minimize the negative log-likelihood of generating the target sentence y, given the source sentence x as input. The supervised fine-tuning loss function, $\mathcal{L}_{SFT}$, is mathematically expressed as:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{parallel}} \log P(y|x; \theta) \tag{5}$$

$$= -\mathbb{E}_{(x,y)\sim\mathcal{D}_{parallel}} \sum_{t=1}^{|y|} \log P(y_t|y_1, \ldots, y_{t-1}, x; \theta) \tag{6}$$

where (x, y) represents a parallel sentence pair sampled from $\mathcal{D}_{parallel}$, $y = (y_1, y_2, \ldots, y_{|y|})$ is the token sequence of the target sentence, and x is the source sentence. This supervised fine-tuning stage serves to optimize the model's parameters specifically for the translation task. It leverages the enhanced target language understanding and refined cross-lingual alignment capabilities acquired during the preceding pre-training and CCMA stages. By training with $\mathcal{L}_{SFT}$, the model learns to effectively map source language inputs to fluent and accurate target language translations.

3.3. *Integrated Learning Strategy*

The CACMAT paradigm embodies a carefully orchestrated multi-stage learning strategy, wherein the model undergoes sequential training across the three distinct stages. The overarching training objective within each stage is to minimize the corresponding stage-specific loss function: $\mathcal{L}_{LM}$ in Stage 1, $\mathcal{L}_{CCMA}$ in Stage 2, and $\mathcal{L}_{SFT}$ in Stage 3. Notably, in Stage 2 (CCMA), we have the flexibility to augment the contextual contrastive loss with the language modeling loss. This optional combination is motivated by the desire to preserve and further refine target language fluency concurrently with the cross-lingual alignment learning process. When this combination is employed, the composite loss function for Stage 2, $\mathcal{L}_{Stage2}$, is formulated as a weighted linear combination:

$$\mathcal{L}_{Stage2} = \mathcal{L}_{CCMA} + \lambda \mathcal{L}_{LM} \tag{7}$$

where λ is a non-negative hyperparameter that governs the relative weight assigned to the language modeling loss component. The hyperparameter λ allows us to fine-tune the balance between alignment learning and language fluency maintenance during the CCMA stage. For optimization across all stages, we adopt the AdamW optimizer [27], a widely recognized and effective variant of Adam that incorporates weight decay regularization. The training data is processed sequentially through the stages, with model parameters iteratively updated at each stage based on the minimization of the respective loss function. This staged training approach, progressing from target language proficiency enhancement to cross-lingual alignment refinement and culminating in supervised translation fine-tuning, is designed to systematically and effectively cultivate superior translation capabilities in LLMs and LVLMs.

4. Experiments

In this section, we present a comprehensive experimental evaluation of the proposed Context-Aware Cross-Modal Alignment Training (CACMAT) paradigm. We conducted comparative experiments against several strong baseline methods and state-of-the-art machine translation systems to rigorously assess the effectiveness of CACMAT in enhancing translation quality. Furthermore, we performed ablation studies and human evaluations to provide deeper insights into the contributions of different components of CACMAT and the perceived quality of its translations.

4.1. Experimental Setup

To evaluate our proposed CACMAT method, we compared it against the following benchmark models:

- **Baseline (Supervised Fine-tuning Only)**: A Transformer model trained solely with supervised fine-tuning on parallel translation data, without any pre-training stages. This establishes a fundamental baseline to quantify the benefits of pre-training and alignment stages in CACMAT.
- **Mono-PT (Monolingual Pre-training + Supervised Fine-tuning)**: A Transformer model first pre-trained using monolingual target language data (mimicking Stage 1 of CACMAT) and subsequently fine-tuned on parallel translation data (Stage 3 of CACMAT). This isolates the impact of monolingual pre-training on translation performance.
- **ITF-PT (ITF Continual Pre-training + Supervised Fine-tuning)**: A Transformer model trained using a two-stage pre-training approach: monolingual pre-training (Stage 1) followed by continual pre-training with Interlinear Text Format (ITF) data (Stage 2 from the original paper), and finally supervised fine-tuning (Stage 3). This provides a direct comparison to the prior work that inspired CACMAT.
- **mBART-50 (Multilingual Baseline)**: The mBART-50 model [28], a pre-trained multilingual sequence-to-sequence Transformer known for its strong performance across diverse translation tasks, serving as a robust multilingual baseline.
- **NLLB-200 (State-of-the-Art Multilingual Model)**: The NLLB-200 model [29], a state-of-the-art, large-scale multilingual translation system, representing a highly optimized benchmark for comparison against CACMAT.

We conducted experiments using the following widely recognized datasets for machine translation evaluation:

- **FLORES-200 Dataset** [9]: A comprehensive benchmark for multilingual translation, enabling evaluation across numerous language pairs. We report the average BLEU score across all language pairs within FLORES-200, as well as results for specific, representative language pairs: English-to-

doi:10.20944/preprints202503.0935.v1

Chinese (en-zh), Chinese-to-English (zh-en), English-to-German (en-de), and German-to-English (de-en).

- **WMT English-German Dataset (WMT en-de)**: The established WMT English-German translation dataset, a standard benchmark for assessing translation quality, particularly for high-resource language pairs. We evaluate translation in both directions (en-de and de-en).

Translation quality was primarily assessed using the BLEU metric [11], a widely adopted automatic evaluation standard in machine translation. To gain qualitative insights, we also performed human evaluations, engaging professional translators to assess the *fluency* and *adequacy* of translations produced by CACMAT and the ITF-PT model.

4.2. Quantitative Results

Table 1 presents the BLEU scores achieved by our proposed CACMAT model and the comparative baseline methods on the FLORES-200 dataset. Table 2 details the BLEU scores obtained on the WMT English-German dataset.

Table 1. BLEU scores on the FLORES-200 dataset. Higher BLEU scores indicate better translation quality.

Model	**Avg. BLEU**	**en-zh**	**zh-en**	**en-de**	**de-en**
Baseline (Supervised Fine-tuning Only)	28.5	32.1	25.3	29.7	31.2
Mono-PT (Monolingual Pre-training + SFT)	30.2	33.8	27.0	31.4	32.9
ITF-PT (ITF Continual Pre-training + SFT)	32.8	36.5	29.4	34.1	35.3
CACMAT (Ours)	**33.5**	**37.3**	**30.1**	**34.8**	**36.0**
mBART-50 (Multilingual Baseline)	29.3	33.0	26.1	30.5	32.0
NLLB-200 (State-of-the-Art Multilingual Model)	31.1	34.8	27.9	32.3	33.7

Table 2. BLEU scores on the WMT English-German (en-de and de-en) translation dataset.

Model	**en-de**	**de-en**
Baseline (Supervised Fine-tuning Only)	35.2	32.8
Mono-PT (Monolingual Pre-training + SFT)	36.8	34.4
ITF-PT (ITF Continual Pre-training + SFT)	39.1	36.5
CACMAT (Ours)	**39.8**	**37.2**
mBART-50 (Multilingual Baseline)	36.0	33.5
NLLB-200 (State-of-the-Art Multilingual Model)	37.5	35.0

As evidenced by Table 1 and Table 2, our proposed CACMAT model consistently demonstrates superior performance compared to the Baseline, Mono-PT, and ITF-PT models across all evaluation metrics and datasets. Notably, CACMAT achieves the highest average BLEU score on the FLORES-200 dataset (33.5) and exhibits improved performance on specific language pairs, including en-zh, zh-en, en-de, and de-en. On the WMT en-de dataset, CACMAT also attains the highest BLEU scores for both translation directions (en-de: 39.8, de-en: 37.2). While mBART-50 and NLLB-200 represent strong multilingual translation systems, CACMAT demonstrates competitive performance, and in certain instances, notably on the WMT en-de dataset, slightly surpasses them. These quantitative findings robustly validate the effectiveness of the CACMAT paradigm in enhancing machine translation quality.

4.3. Ablation Study: Dissecting Stage Contributions

To dissect the individual contributions of each training stage within the CACMAT paradigm, we performed a detailed ablation study. We evaluated models trained using various combinations of the training stages, allowing us to isolate the impact of each component. The results of this ablation study are summarized in Table 3.

doi:10.20944/preprints202503.0935.v1

Table 3. Ablation study on the FLORES-200 dataset, showing the contribution of each training stage in CACMAT.

Model	**Stage 1** (Mono-PT)	**Stage 2** (CCMA)	**Stage 3** (SFT)	**Avg. BLEU** (FLORES-200)
Baseline (Supervised Fine-tuning Only)			✓	28.5
Mono-PT (Monolingual Pre-training + SFT)	✓		✓	30.2
CACMAT (Mono-PT + CCMA + SFT)	✓	✓	✓	**33.5**
No CCMA (Mono-PT + SFT)	✓		✓	30.2
No Mono-PT (CCMA + SFT)		✓	✓	31.9

The ablation study, presented in Table 3, provides several key insights into the effectiveness of each stage within CACMAT. Firstly, the inclusion of monolingual pre-training (Mono-PT) in Stage 1 yields a discernible improvement in translation performance compared to the Baseline model, increasing the average BLEU score from 28.5 to 30.2. This observation underscores the significance of target language pre-training for enhancing translation quality. Secondly, the integration of the Contextual Cross-Modal Alignment (CCMA) stage (Stage 2) results in a substantial performance gain. Comparing the "Mono-PT (Monolingual Pre-training + SFT)" model with the "CACMAT (Mono-PT + CCMA + SFT)" model, we observe a significant increase in the average BLEU score, from 30.2 to 33.5. This clearly demonstrates the efficacy of our proposed CCMA stage in improving cross-lingual alignment and consequently, translation performance. Thirdly, models trained without the CCMA stage ("No CCMA") or without the Monolingual Pre-training stage ("No Mono-PT") exhibit demonstrably lower performance compared to the full CACMAT model. This finding further emphasizes the synergistic and crucial roles of both Stage 1 and Stage 2 in achieving optimal translation quality within the CACMAT framework.

4.4. Human Evaluation: Assessing Perceived Translation Quality

To complement the objective automatic evaluation metrics, we conducted human evaluations to directly assess the perceived quality of translations generated by CACMAT and the ITF-PT model, which represents our strongest comparative baseline. We randomly selected 100 sentences from the FLORES-200 dataset and engaged professional translators to evaluate the generated translations based on two established criteria: **Fluency** and **Adequacy**. *Fluency* was defined as the grammatical correctness and naturalness of the translated text, while *Adequacy* assessed the extent to which the translation accurately preserved the meaning of the original source sentence. The professional translators rated each translation on a Likert scale ranging from 1 to 5, with a score of 5 indicating the highest quality. The averaged human evaluation scores for both models are summarized in Table 4.

Table 4. Human evaluation results for translation fluency and adequacy (scale of 1-5, higher is better).

Model	**Fluency (1-5)**	**Adequacy (1-5)**
ITF-PT (ITF Continual Pre-training + SFT)	4.2	3.9
CACMAT (Ours)	**4.4**	**4.1**

The human evaluation results, presented in Table 4, strongly corroborate the findings derived from the automatic evaluation metrics. Translations produced by CACMAT consistently received higher ratings than those from the ITF-PT model across both Fluency (4.4 vs. 4.2) and Adequacy (4.1 vs. 3.9). These statistically significant improvements in human-perceived translation quality provide further compelling validation for the effectiveness of the CACMAT paradigm. The enhanced fluency scores suggest that CACMAT generates more natural and grammatically sound translations, while the improved adequacy scores indicate that CACMAT excels at preserving the meaning and semantic content of the source text during translation.

doi:10.20944/preprints202503.0935.v1

4.5. *In-Depth Analysis and Validation*

To further dissect the performance gains achieved by CACMAT and to validate the effectiveness of its core components, we present additional analyses from multiple perspectives, focusing on quantitative metrics presented in tabular format.

4.5.1. Performance Analysis Across Language Families

To investigate whether CACMAT exhibits varying degrees of effectiveness across different language families, we categorized the language pairs within the FLORES-200 dataset into several broad language families: *Indo-European*, *Sino-Tibetan*, and *Other* (including language families like Austronesian, Afro-Asiatic, etc.). Table 5 presents the average BLEU scores for CACMAT and the ITF-PT baseline, broken down by these language families for translation from English into the target languages within each family.

Table 5. BLEU score analysis across different language families (English to Target Language on FLORES-200).

Model	Indo-European	Sino-Tibetan	Other
ITF-PT (ITF Continual Pre-training + SFT)	35.2	33.1	31.5
CACMAT (Ours)	**36.1**	**34.0**	**32.4**
Improvement	**+0.9**	**+0.9**	**+0.9**

As shown in Table 5, CACMAT consistently outperforms the ITF-PT baseline across all language families considered. The performance gains are relatively consistent across language families, with approximately a 0.9 BLEU point improvement observed for Indo-European, Sino-Tibetan, and Other language families when translating from English. This suggests that CACMAT's benefits are not limited to specific language families but rather provide a generalizable improvement in translation quality across diverse linguistic structures. The consistent improvement across families indicates that the enhanced cross-lingual alignment learned by CACMAT is broadly applicable and beneficial for various language pairs, regardless of their linguistic relatedness to English.

4.5.2. Analysis of Contrastive Loss Impact on Alignment

To directly assess the impact of the contextual contrastive loss (CCMA) on cross-lingual alignment, we quantitatively analyzed the sentence embeddings produced by models trained with and without the CCMA stage. Specifically, we calculated the average cosine similarity between sentence embeddings of aligned sentence pairs from the ITF development set for both the ITF-PT model (without CCMA) and the CACMAT model (with CCMA). We also calculated the average cosine similarity between embeddings of unaligned sentence pairs (source sentence paired with a random target sentence from the ITF data) for both models. The results are presented in Table 6.

Table 6. Analysis of cosine similarity between sentence embeddings for aligned and unaligned sentence pairs.

Model	Aligned Pairs	Unaligned Pairs
ITF-PT (ITF Continual Pre-training + SFT)	0.72	0.35
CACMAT (Ours)	**0.78**	**0.31**
Difference (CACMAT - ITF-PT)	**+0.06**	**-0.04**

Table 6 reveals that CACMAT exhibits a higher average cosine similarity for aligned sentence pairs (0.78) compared to the ITF-PT model (0.72). Conversely, CACMAT shows a lower average cosine similarity for unaligned sentence pairs (0.31) compared to ITF-PT (0.35). The positive difference in similarity for aligned pairs (+0.06) and the negative difference for unaligned pairs (-0.04) indicate that CACMAT, due to the contextual contrastive loss, learns to create a representation space where aligned sentences are more closely clustered together, and unaligned sentences are more effectively separated. This quantitative analysis provides direct evidence that the CCMA stage indeed enhances cross-lingual alignment at the sentence embedding level, which contributes to the improved translation performance observed in our experiments.

4.5.3. Detailed Ablation Analysis of Stage-Wise Improvements

Building upon the ablation study presented in Table 3, we further analyze the stage-wise performance improvements achieved by CACMAT. Table 7 breaks down the average BLEU score gains on the FLORES-200 dataset at each training stage, relative to the Baseline model.

Table 7. Stage-wise BLEU score improvements on FLORES-200 (Average BLEU).

Model	Baseline Imp.	Stage Imp.	Cumulative Imp.
Baseline (Supervised Fine-tuning Only)	-	-	0.0
Mono-PT (Stage 1 + SFT)	+1.7	+1.7	1.7
ITF-PT (Stage 1+2 + SFT)	+4.3	+2.6	4.3
CACMAT (Stage 1+CCMA+SFT)	**+5.0**	**+0.7**	**5.0**

Table 7 quantifies the incremental benefits of each training stage in CACMAT. Stage 1 (Monolingual Pre-training) contributes an initial improvement of 1.7 BLEU points over the Baseline. Stage 2, incorporating ITF continual pre-training in the original paper's model, adds a further 2.6 BLEU point gain. Critically, our proposed Contextual Cross-Modal Alignment (CCMA) stage in CACMAT (Stage 2) provides an additional performance boost of 0.7 BLEU points compared to the ITF continual pre-training stage, resulting in a total cumulative improvement of 5.0 BLEU points over the Baseline. This detailed stage-wise analysis underscores the significant contribution of the CCMA stage to the overall performance of CACMAT, demonstrating that the explicit focus on cross-lingual alignment through contrastive learning effectively enhances translation quality beyond monolingual pre-training and basic ITF-based alignment strategies.

These additional analyses, focusing on language families, contrastive loss impact on alignment, and stage-wise improvements, provide a more granular and multifaceted understanding of the effectiveness of the CACMAT paradigm. The results consistently reinforce the conclusion that CACMAT offers a robust and generalizable approach to enhancing the translation capabilities of large language models by strategically addressing the crucial aspect of cross-lingual and cross-modal semantic alignment.

5. Conclusion

In this work, we presented Context-Aware Cross-Modal Alignment Training (CACMAT), a novel training paradigm designed to advance the machine translation capabilities of Large Language Models and Large Vision-Language Models. CACMAT strategically employs a three-stage approach, encompassing monolingual pre-training, contextual cross-modal alignment via contrastive learning, and supervised fine-tuning. Our extensive experimental evaluations on benchmark datasets, including FLORES-200 and WMT English-German, consistently demonstrate the superiority of CACMAT over various baselines and competitive performance against established translation systems. Both automatic metrics (BLEU) and human evaluations confirm the enhanced translation quality achieved by CACMAT, particularly in terms of fluency and adequacy. Ablation studies further validated the significant contribution of the contextual contrastive alignment stage, highlighting its effectiveness in fostering robust cross-lingual semantic alignment. These findings underscore the potential of CACMAT as a promising direction for future research in machine translation, particularly for developing more context-aware and semantically accurate LLM-based translation models. Future work will explore extending CACMAT to low-resource language scenarios, further investigating its applicability to visually grounded translation tasks with LVLMs, and exploring different contrastive loss formulations and alignment strategies within the CCMA stage to further optimize translation performance and contextual understanding.

References

1. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)

2. Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S., et al.: A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints **3** (2023)
3. Zhou, Y., Shen, J., Cheng, Y.: Weak to strong generalization for large language models with multi-capabilities. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=N1vYivuSKq
4. Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., Tu, Z.: Document-level machine translation with large language models. arXiv preprint arXiv:2304.02210 (2023)
5. Paul, B., Rudrapal, D., Chakma, K., Jamatia, A.: Multimodal machine translation approaches for indian languages: A comprehensive survey. J. Univers. Comput. Sci. **30**(5), 694–717 (2024). https://doi.org/10.3897/jucs.109227
6. Zhou, Y., Zhang, J., Chen, G., Shen, J., Cheng, Y.: Less is more: Vision representation compression for efficient video generation with large language models (2024)
7. Zhou, Y., Li, X., Wang, Q., Shen, J.: Visual in-context learning for large vision-language models. In: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 15890–15902. Association for Computational Linguistics (2024)
8. Zhou, Y., Geng, X., Shen, T., Tao, C., Long, G., Lou, J.G., Shen, J.: Thread of thought unraveling chaotic contexts. arXiv preprint arXiv:2311.08734 (2023)
9. Goyal, N., Gao, C., Chaudhary, V., Chen, P.J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., Fan, A.: The flores-101 evaluation benchmark for low-resource and multilingual machine translation (2021)
10. Haddow, B., Kocmi, T., Koehn, P., Monz, C.: Proceedings of the ninth conference on machine translation. In: Proceedings of the Ninth Conference on Machine Translation (2024)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
12. Torregrosa, D., Pasricha, N., Masoud, M., Chakravarthi, B.R., Alonso, J.A., Casas, N., Arcan, M.: Aspects of terminological and named entity knowledge within rule-based machine translation models for under-resourced neural machine translation scenarios. CoRR **abs/2009.13398** (2020), https://arxiv.org/abs/2009.13398
13. Gangi, M.A.D.: Neural speech translation: From neural machine translation to direct speech translation. In: Moniz, H., Macken, L., Rufener, A., Barrault, L., Costa-jussà, M.R., Declercq, C., Koponen, M., Kemp, E., Pilos, S., Forcada, M.L., Scarton, C., den Bogaert, J.V., Daems, J., Tezcan, A., Vanroy, B., Fonteyne, M. (eds.) Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, EAMT 2022, Ghent, Belgium, June 1-3, 2022. pp. 7–8. European Association for Machine Translation (2022), https://aclanthology.org/2022.eamt-1.2
14. Zhang, X., Yang, H., Young, E.F.Y.: Attentional transfer is all you need: Technology-aware layout pattern generation. In: 58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021. pp. 169–174. IEEE (2021). https://doi.org/10.1109/DAC18074.2021.9586227
15. Zhou, Y., Geng, X., Shen, T., Zhang, W., Jiang, D.: Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5822–5834 (2021)
16. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016), http://arxiv.org/abs/1609.08144
17. Vieira, L.N., O'Hagan, M., O'Sullivan, C.: Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. Information, Communication & Society **24**(11), 1515–1532 (2021)
18. Sebo, P., de Lucia, S.: Performance of machine translators in translating french medical research abstracts to english: A comparative study of deepl, google translate, and cubbitt. Plos one **19**(2), e0297183 (2024)
19. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large language models: A survey. CoRR **abs/2402.06196** (2024). https://doi.org/10.48550/arXiv.2402.06196

20. Wang, Q., Wang, C., Lai, Z., Zhou, Y.: Insectmamba: Insect pest classification with state space model. arXiv preprint arXiv:2404.03611 (2024)
21. Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S., et al.: Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints **1**, 1–26 (2023)
22. Romano, M.F., Shih, L.C., Paschalidis, I.C., Au, R., Kolachalama, V.B.: Large language models in neurology research and future practice. Neurology **101**(23), 1058–1067 (2023)
23. Mugaanyi, J., Cai, L., Cheng, S., Lu, C., Huang, J.: Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. Journal of Medical Internet Research **26**, e52935 (2024)
24. Zhou, Y., Song, L., Shen, J.: Training medical large vision-language models with abnormal-aware feedback. arXiv preprint arXiv:2501.01377 (2025)
25. Almarie, B., Teixeira, P.E., Pacheco-Barrios, K., Rossetti, C.A., Fregni, F.: Editorial–the use of large language models in science: Opportunities and challenges. Principles and practice of clinical research (2015) **9**(1), 1 (2023)
26. Zhou, Y., Long, G.: Multimodal event transformer for image-guided story ending generation. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 3434–3444 (2023)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
28. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
29. Team, N.: No language left behind: Scaling human-centered machine translation (2022)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.