MT^3: A Synergistic Multi-Task RL Framework for Specializing MLLMs in Text Image Machine Translation

Anonymous ACL submission

Abstract

Text Image Machine Translation (TIMT)—the task of translating textual content embedded in images—is critical for applications in accessibility, cross-lingual information access, and real-world document understanding. However, TIMT remains a complex challenge due to the need for accurate optical character recognition (OCR), robust visual-text reasoning, and high-quality translation, often requiring cascading multi-stage pipelines. Recent advances in large-scale Reinforcement Learning (RL) have improved reasoning in Large Language Models (LLMs) and Multimodal LLMs (MLLMs), but their application to end-to-end TIMT is still underexplored. To bridge this gap, we introduce **MT**^{3}, a novel **M**ulti-**T**ask RL framework to specialize **M**LLMs into end-to-end expert **TIMT** models. **MT**^{3} adopts a synergistic multi-task optimization paradigm targeting three key sub-skills: text recognition, context-aware reasoning, and translation. It is trained using a novel multi-mixed reward mechanism that provides fine-grained feedback, fostering a controllable and transparent optimization process. Furthermore, to facilitate the evaluation of TIMT in authentic cross-cultural and real-world social media contexts, we introduced XHSPost, the first social media TIMT benchmark. Our **MT**^{3}-7B-Zero achieves state-of-the-art results on the latest in-domain MIT-10M benchmark, outperforming strong baselines such as Qwen2.5-VL-72B and InternVL2.5-78B by notable margins across multiple metrics. Additionally, the model shows strong generalization to out-of-distribution language pairs and datasets. In-depth analyses reveal how multi-task synergy, reinforcement learning initialization, curriculum design, and reward formulation contribute to advancing MLLM-driven TIMT. Code and data will be released.

1 Introduction

Text Image Machine Translation (TIMT) is a crucial subfield of machine translation (MT) that focuses on translating source-language texts embedded in images into target-language texts (Ma et al., 2022; Lan et al., 2023; Liang et al., 2024). It has been widely applied in scenarios including photo translation, scanned document translation, and screenshot translation. Unlike traditional text-based machine translation (Xu et al., 2024; Feng et al., 2024), where input and output are purely textual, TIMT is inherently a cross-modal task. It requires systems to process images and generate corresponding textual translations, necessitating the comprehensive integration of fine-grained textual content and diverse visual elements such as layout, objects, and color schemes to produce accurate translations. However, current TIMT approaches, both cascade systems (Hinami et al., 2021; Sable et al., 2023; Zhang et al., 2023, 2025b) and end-to-end (E2E) models (Zhu et al., 2023; Lan et al., 2023; Ma et al., 2024; Liang et al., 2024), often do not explicitly capture fine-grained textual information or model the integration of visual element comprehension. Cascade systems typically use OCR output for a text-based MT model, thereby overlooking visual information, and suffer from error propagation. E2E models, while aiming for unified training, often lack explicit modeling of OCR and visual understanding, which is beneficial for translation quality (Niu et al., 2024).

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated impressive general-purpose capabilities across various cross-modal tasks, such as OCR and VQA (Bai et al., 2025; Chen et al., 2025). Concurrently, it has been shown that rule-based Reinforcement Learning (RL) training can enable these models to learn specific output and reasoning patterns, which can enhance their capabilities in tasks like mathematical reasoning or visual understanding (Team, 2024; Huang et al., 2025; Zhang et al., 2025a; Meng et al., 2025). However, while their performance on general-purpose tasks is remarkable, a critical

gap remains between this generalist ability and the expert-level proficiency required for complex, specialized tasks like TIMT. This work addresses a key question: how can we effectively **specialize** the powerful, generalist MLLMs into domain experts in a controllable and systematic manner?

In this work, we introduce **MT**^{3}, the first framework to apply **M**ulti-**T**ask RL to **M**LLMs for end-to-end **T**ext **I**mage **M**achine **T**ranslation. **MT**^{3} adopts a synergistic multi-task optimization paradigm targeting three key sub-skills: text recognition, context-aware reasoning, and translation. It is trained using a novel multi-mixed reward mechanism that adapts rule-based RL strategies to TIMT's intricacies, offering fine-grained feedback across tasks. Furthermore, recognizing the need for evaluations that reflect real-world social media interactions, we introduce XHSPost, the first social media TIMT benchmark, to evaluate TIMT in more authentic cross-cultural scenarios. Our experiments demonstrate the efficacy of this approach: our **MT**^{3}-7B-Zero model achieves state-of-the-art results on the latest in-domain MIT-10M benchmark (Li et al., 2024), outperforming strong MLLM baselines, including Qwen2.5-VL-72B (Bai et al., 2025) and InternVL2.5-78B (Chen et al., 2025), by notable margins (approximately 15-25 points average improvement across BLEU, chrF++, METEOR). Additionally, the model shows strong generalization to out-of-distribution (OOD) language pairs and datasets, including the newly introduced XHSPost. In-depth analyses reveal how multi-task synergy, reinforcement learning initialization, curriculum design, and reward formulation contribute to advancing MLLM-driven TIMT. Our core contributions are as follows:

- We propose MT^3, the first framework applying reinforcement learning to MLLMs for end-to-end TIMT, featuring a novel synergistic multi-task optimization paradigm and an effective multi-mixed reward mechanism. Extensive experiments validate MT^3's superior performance on standard TIMT benchmarks and its strong OOD generalization capability, significantly surpassing existing cascaded systems and advanced MLLMs.
- We present XHSPost, the first social media TIMT benchmark. This new resource is designed to facilitate research and enable practical evaluation of TIMT systems in authentic, real-world cross-cultural scenarios.
- We provide comprehensive analyses yielding key insights into multi-task RL for TIMT. We demonstrate that multi-task synergy is vital for RL success in TIMT, and zero-start RL from MLLMs outperforms SFT+long Chain-of-Thought (CoT) in performance and efficiency. Besides, we show that curriculum learning and reward metric selection critically impact training and results.

2 Method

In this section, we detail the MT^3 framework. MT^3 decomposes TIMT into distinct, explicit sub-tasks and employs a novel multi-mixed reward mechanism to guide the MLLM (see Figure 1). The RL training is performed using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm, selected for its efficiency and effectiveness.

2.1 Multi-Task Formulation for TIMT

Instead of treating TIMT as a monolithic, black-box task, MT^3 decomposes it into a structured, three-stage generative process. This explicit decomposition, guided by a formal task schema (see Appendix A), is designed to elicit and individually optimize the distinct capabilities required for high-quality TIMT: visual perception, multimodal reasoning, and context-aware translation. MT^3 transforms the generation process into a controllable and interpretable pipeline within a single MLLM.

Recognition. This step directly leverages the model's visual perception and text extraction capabilities, laying a key foundation for TIMT. The MLLM performs fine-grained OCR, transcribing image text in its natural reading order within `<recognize>` and `</recognize>` tags.

Reasoning. Then, MT^3 elicits and incentivizes the MLLM's advanced reasoning capabilities by considering the recognized text alongside the broader visual context (e.g., layout, objects, color schemes, spatial relationships) to ensure the subsequent translation is contextually appropriate. The MLLM engages in a reasoning process to generate an intermediate chain-of-thought analysis, outputting its analysis within `<think>` and `</think>` tags.

Translation. Finally, the MLLM performs the conditional generation of the target-language translation informed by the recognition and reasoning stages. This is not a simple classical source-to-target translation based solely on the recognized text. Instead, the generation process is conditioned

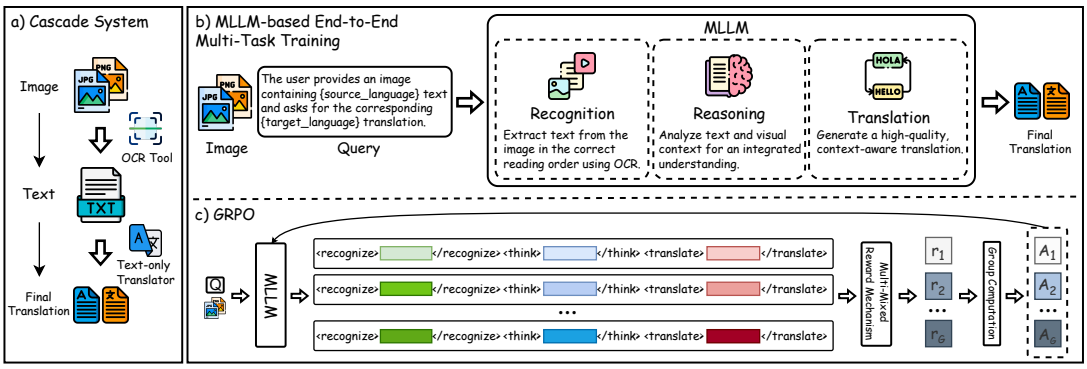

Figure 1: An illustration of MT^3 framework. MT^3 contrasts the traditional (a) Cascade System, which separates OCR and translation into a multi-stage pipeline, so the translator cannot utilize the visual components to help with translation. Our method unifies three key sub-skills—Recognition, Reasoning, and Translation—within a single MLLM. The model is then optimized using (c) Group Relative Policy Optimization (GRPO), where multiple structured outputs are sampled, evaluated by a multi-mixed reward mechanism, and used to update the policy model.

on the complete preceding context, including both the factual text from **Recognition** and the rich, synthesized understanding from the reasoning process in **Reasoning**. This output is presented within `<translate>` and `</translate>` tags.

This explicit multi-task formulation guides the model through distinct recognition, reasoning, and translation stages. It allows for monitoring and rewarding these intermediate steps, thereby fostering more robust and interpretable TIMT capabilities, as validated in our ablation studies (Section 3.1).

2.2 Multi-Mixed Reward Mechanism

The reward signal r is crucial in RL. While rule-based rewards suit tasks with verifiable answers like math (Guo et al., 2025), translation often lacks a single gold output and TIMT's multi-faceted quality (e.g., recognition accuracy, translation quality) necessitates a more nuanced approach. We propose a **multi-mixed reward mechanism**, adapting rule-based concepts by integrating format adherence checks with task-specific quality assessments.

Format Reward (R_{format}). We use regular expression extraction to enforce the structured response format illustrated in the MT^3 prompt. The format reward is computed as:

$$R_{format} = \begin{cases} 1, & \text{if format is correct} \\ -3, & \text{if format is incorrect} \end{cases} \quad (1)$$

This penalty for incorrect format strongly discourages deviations and encourages the model to rapidly learn the required output structure, as evidenced in Figure 3.

Task-Specific Rewards (R_{task}). If the output format is correct ($R_{format} = 1$), we calculate rewards for recognition and translation sub-tasks. While the MT community has developed various evaluation metrics (Freitag et al., 2022, 2023), Feng et al. (2025) and Ramos et al. (2025) noted that relying on a single metric might result in sub-optimal overall performance. Therefore, we propose a metric-mixed approach. We average multiple standard metrics for each task, providing a more robust, balanced, and nuanced reward signal:

- **Translation Reward ($R_{task-trans}$):** Assesses final translation quality by averaging scores from standard MT metrics: BLEU (Post, 2018), chrF++ (Popović, 2017), and METEOR (Banerjee and Lavie, 2005), against the reference. Formally, $R_{task-trans} = (S_{\text{BLEU}} + S_{\text{METEOR}} + S_{\text{chrF++}})/3$. The specific combination of metrics can be varied, as explored in our experiments on metric reward selection (Section 3.3). Furthermore, as highlighted by Feng et al. (2025), the quality of the reasoning process is implicitly encouraged through its positive influence on the final translation quality.

- **Recognition Reward ($R_{task-rec}$):** Evaluates recognized text quality against a ground-truth transcription by averaging suitable OCR metrics (Fu et al., 2024): BLEU, METEOR, F1-score, and normalized versions of Character Error Rate (CER) and Edit Distance. Formally, $R_{task-rec} = (S_{\text{BLEU}} + S_{\text{METEOR}} + S_{\text{F1-score}} + \min(\max(1 - S_{\text{EditDistance}}, 0), 1) + \min(\max(1 - S_{\text{CER}}, 0), 1))/5$.

Final Reward. The final reward r fed to GRPO combines R_{format} and R_{task}:

MODEL	MIT-10M (ZH-EN)				MIT-10M (EN-ZH)			
	BLEU	chrF++	METEOR	Avg.	BLEU	chrF++	METEOR	Avg.
Zero-shot Systems								
Cascade System								
EasyOCR + NLLB-3.3B	2.54	10.71	7.72	6.99	11.93	12.48	10.84	11.75
Advanced MLLMs								
InternVL2.5-8B	9.98	26.59	24.99	20.52	31.74	38.21	38.49	36.15
InternVL2.5-26B	13.29	27.82	24.60	21.90	33.22	42.43	46.55	40.73
InternVL2.5-38B	10.77	26.00	24.49	20.42	38.61	48.87	44.04	43.84
InternVL2.5-78B	12.50	26.85	27.11	22.15	36.97	48.96	46.19	44.04
Qwen2.5-VL-7B	15.59	31.34	33.44	26.79	31.66	44.20	48.41	41.42
Qwen2.5-VL-72B	11.33	32.82	41.26	28.47	35.55	50.78	52.76	46.36
Fine-tuned VLMs								
Qwen2.5-VL-7B (SFT, TIMT)	18.66	40.74	43.20	34.20	43.33	48.62	55.71	49.22
Qwen2.5-VL-7B (SFT, OCR + TIMT)	22.57	42.39	41.80	35.59	48.69	54.66	59.60	54.32
MT^3-7B-Zero (RL)	20.31	44.42	43.99	36.24	54.49	61.77	68.23	61.50

Table 1: In-domain (IND) performance comparison on the MIT-10M benchmark (ZH-EN and EN-ZH). Metrics reported are BLEU, chrF++, and METEOR, along with their average (Avg.). MT^3-7B-Zero (RL) is compared against open-source systems and supervised fine-tuned variants.

MODEL	OOD (LANGUAGE PAIRS)						OOD (DATASETS)							
	MIT-10M (EN-DE)		MIT-10M (ZH-FR)		MIT-10M (DE-FR)		OCRMT30K (ZH-EN)		DOTA (EN-ZH)		XHSPOST (EN-ZH)		XHSPOST (ZH-EN)	
	chrF++	METEOR	chrF++	METEOR	chrF++	METEOR	chrF++	METEOR	chrF++	METEOR	chrF++	METEOR	chrF++	METEOR
Strong Baseline														
Qwen2.5-VL-72B	48.37	40.60	27.29	30.13	48.82	45.06	**39.73**	**42.65**	**37.65**	**50.75**	**42.97**	**61.28**	**55.10**	**54.07**
InternVL2.5-78B	50.54	43.11	18.25	19.59	27.98	25.45	31.57	32.81	32.88	45.03	37.61	57.60	53.14	51.18
Same-size Baseline														
Qwen2.5-VL-7B	48.12	40.65	27.19	30.59	33.60	31.99	23.99	24.90	32.62	44.16	36.79	55.16	49.01	47.05
InternVL2.5-8B	42.42	36.67	19.32	14.71	23.17	20.41	21.95	21.12	24.75	34.89	28.32	46.20	42.19	38.93
MT^3-7B-Zero (RL)	**56.14**	**52.11**	**35.67**	**38.53**	**53.88**	**54.99**	37.42	38.17	34.87	46.60	39.95	58.58	53.26	52.04

Table 2: Out-of-distribution (OOD) generalization performance using chrF++ and METEOR. MT^3-7B-Zero is evaluated against strong and same-size MLLMs on unseen language pairs from MIT-10M (EN-DE, ZH-FR, DE-FR) and different datasets (OCRMT30K, DoTA, XHSPost). **Bold** indicates the best, underline indicates the second best.

$$r = \begin{cases} R_{\text{format}} + R_{\text{task-rec}} + R_{\text{task-trans}}, & \text{if format is correct} \\ -3, & \text{if format is incorrect} \end{cases} \tag{2}$$

This formulation extends the simple binary rule-based reward (a base score of 1 for correct format) by adding continuous, mixed-metric scores for sub-tasks. This provides more granular feedback than purely binary rewards, enabling the model to learn fine-grained improvements in both recognition and translation quality (see training curves in Figure 3).

2.3 RL Algorithm

We use the GRPO algorithm (Shao et al., 2024) to train the translation model with our rule-metric mixed reward. In each training step, for a given translational question q, we sample a group of candidate outputs $\{o_1, o_2, \cdots, o_G\}$ from the policy model $\pi_{\theta_{old}}$. $A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$ is the computed advantage using the group rule-metric mixed rewards $\{r_1, r_2, \cdots, r_G\}$. GRPO then maximizes the following objective to optimize π_θ:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q),\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \; \text{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon \Big) A_i \Big) - \beta \, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \Bigg], \tag{3}$$

where ε and β are hyperparameters controlling the PPO clipping threshold and the weight of the Kullback–Leibler (KL) divergence penalty (Schulman et al., 2017; Shao et al., 2024), respectively.

2.4 Experimental Setup

Datasets and Evaluation Metrics. Our primary experiments focus on English-Chinese (EN-ZH) and Chinese-English (ZH-EN) TIMT tasks. For RL training, we select 15K image-text pairs from the MIT-10M dataset (Li et al., 2024) for each translation direction. To comprehensively evaluate our MT^3 framework, we test its performance across three distinct settings: (1) **In-domain (IND):** utilizing the standard ZH-EN and EN-ZH test

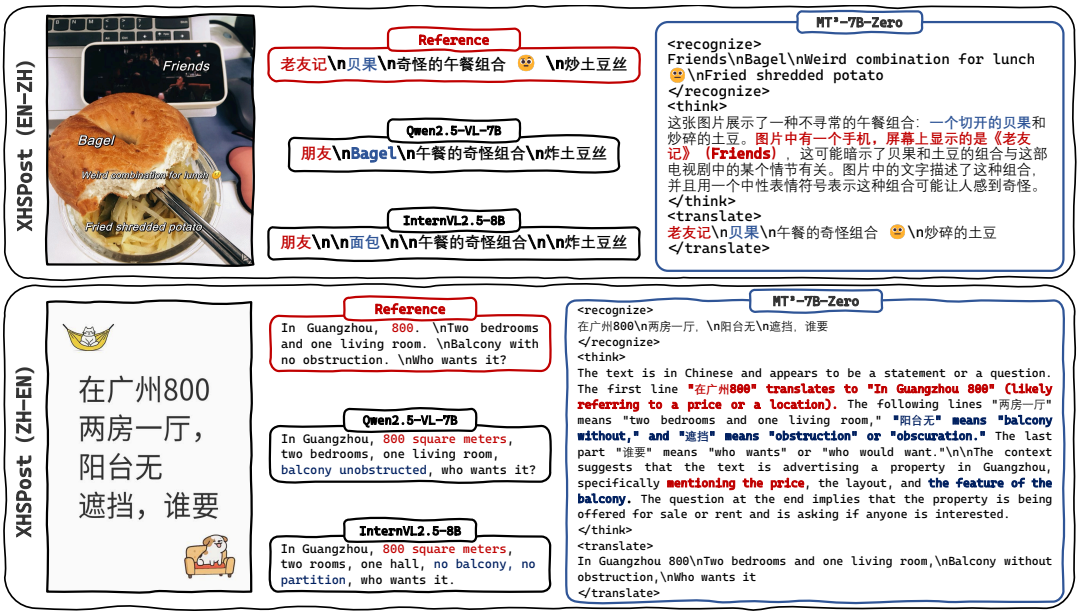

Figure 2: Illustrative TIMT examples from the XHSPost benchmark. MT^3 demonstrates superior contextual understanding by correctly translating '*Friends*' to '老友记' based on visual context (**Top**), and accurately interpreting a property advertisement, including price and layout details (**Bottom**).

sets from the MIT-10M dataset. (2) **Out-of-distribution (OOD) - Language Pairs:** employing unseen language pairs (EN-DE, ZH-FR, DE-FR) from within the MIT-10M dataset. (3) **Out-of-distribution (OOD) - Datasets:** using the test sets from the OCRMT30K dataset (Lan et al., 2023) and the document-level DoTA dataset (Liang et al., 2024). For all evaluations, we report SacreBLEU (Post, 2018), chrF++ (Popović, 2017), and METEOR (Banerjee and Lavie, 2005) scores as previous works (Li et al., 2024; Liang et al., 2024).

XHSPost Benchmark: A Real-World Social Media TIMT Scenario. Existing TIMT datasets primarily focus on general photos, academic documents, or advertisements (see Figure 9). To address the notable absence of benchmarks for the increasingly prevalent social media domain, we constructed XHSPost, a dataset of social media posts for ZH-EN and EN-ZH TIMT. After a meticulous selection and filtering process (detailed in Appendix M), we obtained 106 English and 109 Chinese posts. We anticipate XHSPost will foster real-world applications of TIMT.

Baselines. To comprehensively assess the performance of our MT^3-7B-Zero, we compare it against a diverse set of baselines. This includes a **Cascade System**, which first employs EasyOCR[1] to extract text from images, arranges the extracted text sequentially, and then translates it using the NLLB-3.3B model (Costa-Jussà et al., 2022). We also benchmark against several **Advanced MLLMs** used in a zero-shot prompting setup, specifically the InternVL2.5 series (Chen et al., 2025) and Qwen2.5-VL series (Bai et al., 2025); more evaluation details can be found in Appendix D. Furthermore, we compare against **Supervised Fine-Tuning (SFT)** variants of Qwen2.5-VL-7B. These SFT models were trained on the MIT-10M dataset using two distinct prompt templates: one focused solely on the end-to-end *TIMT* task, and another employing an *OCR* + *TIMT* multi-task format. The specific SFT prompts are provided in Appendix C.

Training Details. Our implementation is based on verl[2] framework. We select the Qwen2.5-VL-7B as starting models for MT^3 training. During training, we configure a batch size of 64 and utilize 16 rollouts per prompt within the GRPO algorithm. More training details are in Appendix B.

2.5 Experimental Results

In-Domain Performance. As detailed in Table 1, our MT^3-7B-Zero significantly outperforms compared systems on the in-domain MIT-10M benchmark. For ZH-EN, MT^3-7B-Zero achieved an average score of 36.24, surpassing Qwen2.5-VL-72B (Avg. 28.47) by 7.77 points and InternVL2.5-78B by 14.09 points. The improvements are more pronounced for EN-ZH, where our model (Avg. 61.50)

[1] https://github.com/JaidedAI/EasyOCR

[2] https://github.com/volcengine/verl

Ablation	Tasks			MIT-10M (ZH-EN)				Δ	MIT-10M (EN-ZH)				Δ	XHSPost (ZH-EN)				Δ
	Recognition	Reasoning	Translation	BLEU	chrF++	METEOR	Avg.		BLEU	chrF++	METEOR	Avg.		BLEU	chrF++	METEOR	Avg.	
Full Tasks	✓	✓	✓	20.31	44.42	43.99	36.24	—	54.49	61.77	68.23	61.50	—	26.94	53.26	52.04	44.08	—
w/o Reasoning	✗	✓	✓	22.21	45.06	44.63	37.30	+1.06	52.12	61.34	64.26	59.24	-2.26	25.01	51.50	49.73	42.08	-2.00
w/o Recognition	✓	✗	✓	16.99	40.08	42.65	33.24	-3.00	47.61	57.78	65.11	56.83	-4.67	25.06	52.01	50.67	42.58	-1.50
Only TIMT	✗	✗	✓	13.43	38.47	42.21	31.37	-4.87	44.04	55.87	61.77	53.89	-7.61	24.21	51.32	49.54	41.69	-2.39

Table 3: Multi-task ablation study for MT^3. Performance impact of removing Recognition, Reasoning, or both, compared to the full task setup. Δ denotes the average metric score difference compared to the *Full Tasks* setting.

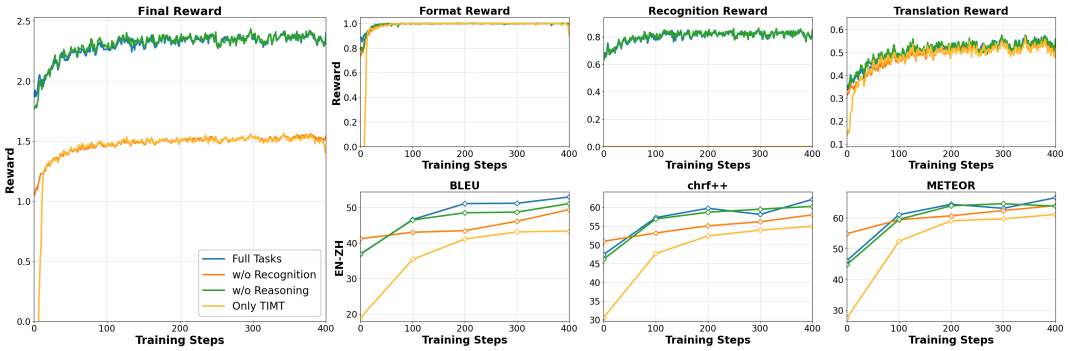

Figure 3: Impact of multi-task ablation on reward optimization and performance progression. **Top row**: Progression of *Final Reward* and individual task rewards (Format, Recognition, Translation). **Bottom row**: Progression of translation quality metrics (BLEU, chrF++, METEOR).

outperforms Qwen2.5-VL-72B by 15.14 points and InternVL2.5-78B by 17.46 points. Notably, our RL-trained model also surpasses SFT variants of Qwen2.5-VL-7B, including those trained with OCR+TIMT multi-task or single TIMT task setups. These results underscore the efficacy of our multi-task RL approach over supervised fine-tuning and zero-shot larger MLLMs for TIMT.

Out-of-Distribution Performance. We assessed MT^3-7B-Zero's generalization ability on diverse out-of-distribution (OOD) scenarios, including unseen language pairs from MIT-10M (EN-DE, ZH-FR, DE-FR), distinct TIMT datasets (OCRMT30K (Lan et al., 2023), the document-level DoTA (Liang et al., 2024)), and our newly introduced real-world social media benchmark, XHSPost. As shown in Table 2, our model consistently demonstrated strong OOD performance, generally outperforming comparable-sized models and rivaling larger MLLMs. For instance, on MIT-10M (DE-FR), MT^3-7B-Zero achieved 54.99 chrF++ and 53.88 METEOR, surpassing all listed baselines. On DoTA, our model's 46.60 chrF++ was considerably higher than Qwen2.5-VL-72B's 37.65, with a competitive METEOR score. These findings highlight the robust generalization fostered by the MT^3 framework. See BLEU scores in Appendix K.

3 Analyses

This section dissects the key components and training dynamics of the MT^3 framework through a series of targeted analyses. We first validate our core multi-task design via an in-depth ablation study (Section 3.1). We then examine crucial aspects of the training recipe, including the RL initialization strategy (Section 3.2) and the selection of reward metrics (Section 3.3). To further demonstrate the framework's robustness and scalability, we present additional experiments in the appendix, including an analysis of curriculum learning (Appendix E) and a multilingual training evaluation (Appendix F).

3.1 Multi-Task Ablation Analysis

To disentangle the contributions of the text recognition, reasoning, and translation components within our multi-task framework, we conduct an ablation study comparing the full-task MT^3-7B-Zero against variants where specific tasks are omitted during RL training: (1) *w/o Reasoning*; (2) *w/o Recognition*; and (3) *Only TIMT*. Detailed prompts for these settings are available in Appendix J. As shown in Table 3, the *Full Tasks* setting consistently achieves the best performance across the evaluated benchmarks. Removing the explicit text recognition step (*w/o Recognition*) leads to a significant performance drop across all settings, underscoring the criticality of accurate visual text extraction. Omitting the reasoning step (*w/o Reasoning*) generally results in a performance decrease, particularly on more nuanced datasets, although its utility can vary with task complexity or language pair characteristics. Training only the translation task (*Only TIMT*) results in the most substantial degradation.

Model	In-Domain						Out-of-Domain											
	MIT-10M (ZH-EN)			MIT-10M (EN-ZH)			OCRMT30K (ZH-EN)			DoTA (EN-ZH)			XHSPost (EN-ZH)			XHSPost (ZH-EN)		
	BLEU	chrF++	METEOR	BLEU	chrF++	METEOR	BLEU	chrF++	METEOR	BLEU	chrF++	METEOR	BLEU	chrF++	METEOR	BLEU	chrF++	METEOR
w/ "Cold Start"																		
MT^3-7B-QVQ-Distill	15.37	36.36	40.20	42.21	53.55	58.37	12.39	34.68	35.43	29.23	25.60	37.08	48.20	39.61	58.41	24.61	49.74	48.76
w/o "Cold Start"																		
MT^3-7B-Zero	20.31	44.42	43.99	54.49	61.77	68.23	14.21	37.42	38.17	40.46	34.87	46.60	49.41	39.95	58.58	26.94	53.26	52.04

Table 4: Impact of initialization strategies for RL. Comparison of MT^3-7B-Zero (RL from scratch) and MT^3-7B-QVQ-Distill (SFT on distilled data then RL) on in-domain and out-of-domain benchmarks.

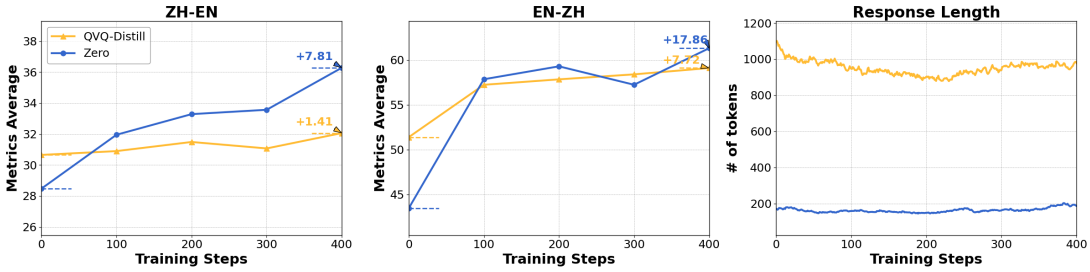

Figure 4: Training dynamics comparing Zero-start RL (MT^3-7B-Zero) vs. SFT initialization (MT^3-7B-QVQ-Distill). **Left and Center:** Average metric score progression on MIT-10M ZH-EN and EN-ZH test sets. **Right:** Average response length during RL training.

As illustrated in Figure 2, MT^3 correctly translates culturally nuanced terms and interprets ambiguous numerical information by integrating visual and textual cues, demonstrating how explicit recognition and reasoning foster deeper contextual understanding crucial for TIMT (see Appendix I for more detailed analysis). These cases demonstrate how the explicit modeling of recognition and reasoning within MT^3 fosters deeper contextual understanding crucial for accurate TIMT in diverse scenarios. The reward and performance progression curves (Figure 3) further support this: variants with recognition yield higher translation rewards, and the *Full Tasks* setting shows the most effective trajectories. This confirms the benefit of our multi-task formulation, emphasizing explicit text recognition and the synergy of jointly optimizing all three sub-tasks.

3.2 Impact of Initialization Strategy: Zero vs. QVQ-Distill

To assess the effect of the RL starting point, we compare MT^3-7B-Zero (RL directly from the general MLLM checkpoint, *Zero-start*) against MT^3-7B-QVQ-Distill (initialized via SFT on 10K high-quality TIMT examples distilled from the QVQ-72B (Team, 2024); details in Appendix G). Training dynamics (Figure 4) show that while the QVQ-Distill model benefits from an initial SFT performance boost, the Zero-start model exhibits a steeper learning curve, rapidly surpassing the cold-start variant and converging to a markedly higher performance ceiling. The Zero-start model also produces considerably shorter and more stable response lengths. A qualitative example in Appendix H further reveals differences in reasoning patterns: the QVQ-Distill model often simulates self-reflection patterns, whereas the Zero-start model can incentivize a more straightforward and less redundant reasoning path conducive to TIMT. Table 4 corroborates these findings, showing Zero-start yields significantly better final performance.

3.3 Metric Reward Selection for Translation

Our multi-mixed reward mechanism (Section 2.2) averages multiple metrics for task-specific rewards. Here, we analyze using individual metrics (BLEU, chrF++, METEOR) versus our proposed *Mixed Reward* (average of these three) for $R_{task-trans}$. Figure 5 presents Spearman and Kendall correlation matrices, which indicate positive correlations between the final rewards from individual metrics and the mixed-metric reward, suggesting the shared underlying signal of quality. Notably, BLEU shows the lowest correlation with chrF++, a divergence that is also reflected in their differing performance trajectories in Figure 6. These performance plots demonstrate that the *Mixed Reward* generally results in the most consistent and often the highest performance across all three evaluation metrics. This mitigates optimizing for a single metric's peculiarities, leading to holistically improved quality and validating our mixed-metric approach.

4 Related Works

Text Image Machine Translation. Text Image Machine Translation (TIMT) aims to translate texts

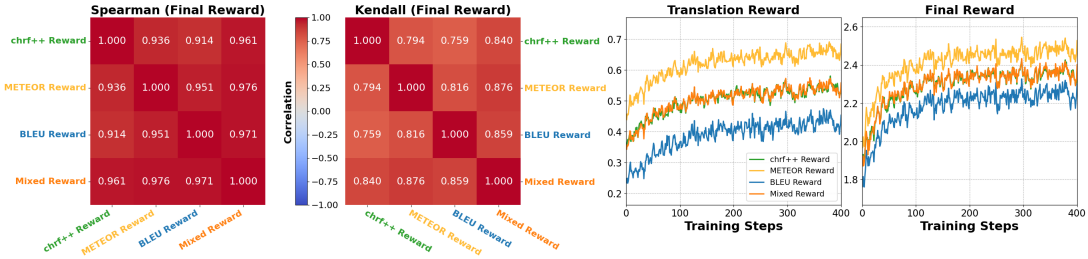

Figure 5: Analysis of translation rewards. **Left**: Kendall and Spearman correlation matrices between different individual metric rewards (chrF++, METEOR, BLEU) and the *Mixed Reward*, based on final reward. **Right**: Progression of *Translation Reward* and *Final Reward* when optimizing for individual metric reward versus the *Mixed Reward*.

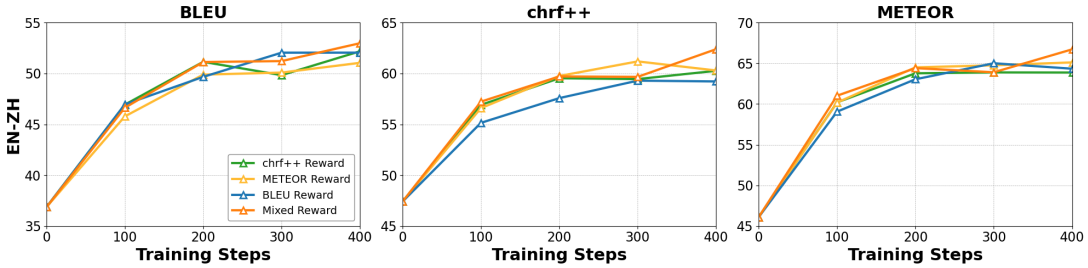

Figure 6: Effect of different translation reward metric choices ($R_{task-trans}$) on performance progression for EN-ZH translation on MIT-10M. Curves display BLEU, chrF++, and METEOR scores over training steps when optimizing with individual metrics versus the *Mixed Reward*.

embedded in images (Ma et al., 2022). Prevailing approaches include: (1) *Cascade systems* (Hinami et al., 2021; Sable et al., 2023; Zhang et al., 2023, 2025b), which sequentially combine Optical Character Recognition (OCR) and Neural Machine Translation (NMT), often facing issues like error propagation and latency. (2) *End-to-end (E2E) models* (Zhu et al., 2023; Lan et al., 2023; Ma et al., 2024; Niu et al., 2024; Liang et al., 2024), developed to unify training and improve efficiency. Early E2E methods integrated visual encoders and text decoders, with some bridging modality gaps using pre-trained components (Zhu et al., 2023) or dynamically assembling models (Liang et al., 2024). While advanced MLLMs (Chen et al., 2025; Bai et al., 2025) show promise for more effective E2E TIMT (Li et al., 2024), their specific application to this task remains underexplored.

MLLM Reasoning with Reinforcement Learning. RL has been shown to enhance the reasoning capabilities of LLMs (Team et al., 2025; Guo et al., 2025), leading to explorations of RL in MLLMs (Huang et al., 2025; Zhang et al., 2025a; Shen et al., 2025; Liu et al., 2025; Meng et al., 2025). Several studies employ R1-like training methodologies, often cultivating a "<think> then <answer>" paradigm. For instance, some focus on improving MLLM reasoning through synthesized data (Huang et al., 2025), while others design online RL frameworks for self-improvement via step-wise rewarding (Zhang et al., 2025a), or explore R1-style RL for general vision-language tasks (Shen et al., 2025). These efforts have primarily concentrated on tasks like multimodal mathematical reasoning or general visual understanding. However, there has been less focus on leveraging multi-task RL paradigms (not only <think>) or addressing application-oriented downstream tasks like TIMT, a gap our work aims to fill.

5 Conclusion

We introduce MT^3, the first framework to specialize MLLMs for end-to-end Text Image Machine Translation using systematic multi-task reinforcement learning. By synergistically optimizing recognition, reasoning, and translation with a novel multi-mixed reward mechanism, our MT^3-7B-Zero achieves leading performance on the MIT-10M benchmark and exhibits robust out-of-distribution generalization, significantly outperforming larger baselines. A key contribution is also the introduction of the XHSPost benchmark for realistic social media evaluation. Our comprehensive analyses confirm the benefits of multi-task synergy, offering a valuable blueprint for developing future RL-driven, specialized MLLMs.

Limitations

Although MT^3 demonstrates strong performance on the TIMT task across multiple benchmarks, there remains room for improvement in its large-scale multilingual capabilities. Including more metrics and larger MLLMs are also important improvement directions. In future work, we aim to leverage the robust text-grounding and visual understanding abilities of MLLMs to extend this approach to cross-lingual and multilingual VQA and reasoning tasks, while further enhancing MLLMs' alignment with user preferences.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *Preprint*, arXiv:2504.10160.

Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving llm-based machine translation with systematic self-correction. *arXiv preprint arXiv:2402.16379*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12998–13008.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Gant Laborde. Deep nn for nsfw detection.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3479–3491.

Bo Li, Shaolin Zhu, and Lijie Wen. 2024. Mit-10m: A large scale parallel corpus of multilingual image translation. *Preprint*, arXiv:2412.07147.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *Preprint*, arXiv:2503.01785.

Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. Improving end-to-end text image translation from the auxiliary text translation task. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1664–1670. IEEE.

Cong Ma, Yaping Zhang, Zhiyang Zhang, Yupu Liang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2024. Born a babynet with hierarchical parental supervision for end-to-end text image machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2468–2479.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2503.07365.

Liqiang Niu, Fandong Meng, and Jie Zhou. 2024. Umtit: Unifying recognition, translation, and generation for multimodal text image translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16953–16972.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André F. T. Martins. 2025. Fine-grained reward optimization for machine translation using error severity mappings. *Preprint*, arXiv:2411.05986.

Nilesh P Sable, Priya Shelke, Ninad Deogaonkar, Nachiket Joshi, Rudra Kabadi, and Tushar Joshi. 2023. Doc-handler: Document scanner, manipulator, and translator based on image and natural language processing. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6. IEEE.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Cong Ma, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025b. Understand layout and translate text: Unified feature-conductive end-to-end document image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhiyang Zhang, Yaping Zhang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. A novel dataset and benchmark analysis on document image translation. In *China Conference on Machine Translation*, pages 103–115. Springer.

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. Peit: bridging the modality gap with pretrained models for end-to-end image translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13433–13447.

A MT^3 Prompt

MT^3 Prompt

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. First, the assistant recognizes all the text in the image following the natural reading order. Then, the assistant carefully analyzes the recognized text and the visual elements in the image, considering the layout, objects, color schemes, spatial relationships, and other contextual clues that may influence meaning. This integrated understanding ensures the translation is accurate, coherent, and appropriate to the visual setting. After thorough reasoning based on both textual content and visual context, the assistant provides the user with the final translation in reading order. The recognized text, reasoning process, and final translation are enclosed within <recognize> </recognize>, <think> </think>, and <translate> </translate> tags, respectively. The format must be as follows: <recognize> recognized text here </recognize> <think> reasoning process here </think> <translate> final translation here </translate>
User: *{image}* Translate all the text in this image into *{target_language}* following the natural reading order.

Here, *{source_language}* and *{target_language}* indicate the source and target languages, and *{image}* denotes the input image requiring translation.

B Training Details

We employ a constant learning rate of 5e-7 and set the sampling temperature to 1.0. The maximum generation length for responses is capped at 4096 tokens. We set the KL penalty coefficient β to 0.01 and set the PPO clipping range ϵ to 0.2. The model is trained for 1 epoch on 32 NVIDIA H800 80G GPUs for about 15 hours.

C SFT Prompts

The prompts for SFT (TIMT) and SFT (OCR + TIMT) are structured as follows:

SFT (TIMT) Prompt

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. Then, the assistant provides the user with the final translation in reading order, separating text from different positions (boxes) with a line break. The final translation must be enclosed within <translate> </translate> tags. The format must be as follows: <translate>final translation here</translate>
User: *{image}* Translate all the text in this image into *{target_language}* following the natural reading order.
Assistant: <translate> *{target_text}* </translate>

SFT (OCR + TIMT) Prompt

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. First, the assistant recognizes all the text in the image following the natural reading order, separating text from different positions (boxes) with a line break. Then, the assistant provides the user with the final translation in reading order, separating text from different positions (boxes) with a line break. The recognized text and final translation must be enclosed within <recognize> </recognize> and <translate> </translate> tags, respectively. The format must be as follows: <recognize>recognized text here</recognize> <translate>final translation here</translate>
User: *{image}* Translate all the text in this image into *{target_language}* following the natural reading order.
Assistant: <recognize> *{source_text}* </recognize><translate> *{target_text}* </translate>

D Evaluation Details

When evaluating MLLM's performance on the test set, we deployed open-source models locally using

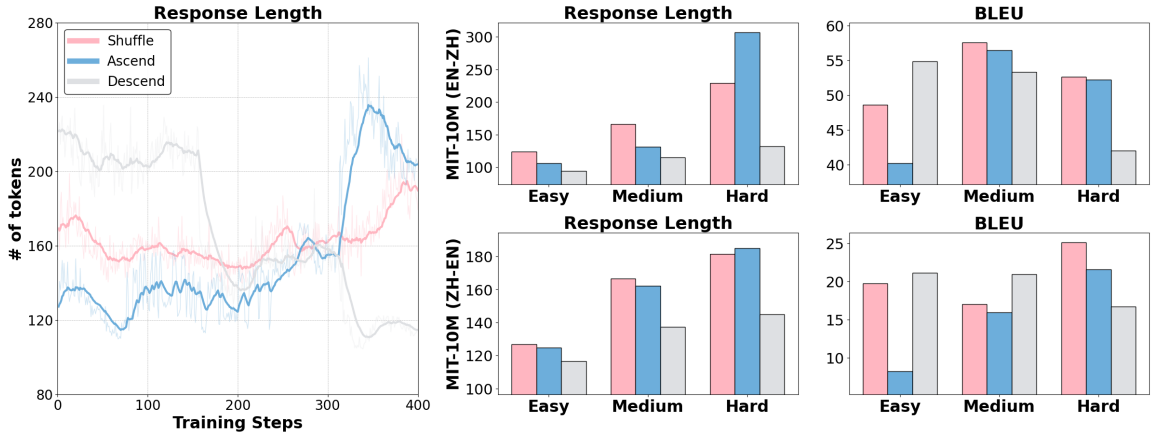

Figure 7: Influence of curriculum learning strategies on training dynamics and performance. **Left**: Average response length dynamics for *Shuffle*, *Ascending (easy-to-hard)*, and *Descending (hard-to-easy)* difficulty curricula during training. **Right**: Final response length and BLEU scores on MIT-10M difficulty splits (Easy, Medium, Hard) for EN-ZH and ZH-EN..

frameworks like vLLM[3] or lmdeploy[4] implementations. We use the greedy decoding strategy for all systems. The maximum generation length was capped at 4096 tokens. The prompt for evaluating MLLMs is structured as follows:

Evaluating MLLMs Prompt

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation.
User: *{image}* Translate the text in the image from *{source_language}* into *{target_language}* following the natural reading order without any explanation.

E Curriculum Learning

We explore curriculum learning by ordering training data by difficulty, categorized by Li et al. (2024) as Easy, Medium, or Hard using heuristics based on text length and scene complexity, with details in Appendix L. We compare three data presentation strategies: *Shuffle* (random), *Ascend* (easy-to-hard), and *Descend* (hard-to-easy). Figure 7 shows response length dynamics: *Ascend* leads to longer responses as harder samples appear, while *Descend* starts longer and shortens. Figure 7 also shows the final BLEU scores and response lengths on MIT-10M test splits. *Ascend* tends to yield competitive or superior BLEU scores on Hard instances, whereas *Descend* performs better on Easy instances. *Shuffle* demonstrates the most balanced performance across different difficulty splits. These results suggest that while specific curricula can tune for specific difficulty levels, a shuffled curriculum may be more broadly beneficial for robust TIMT.

Metric	Step	ZH-EN	EN-ZH	ZH-DE	DE-ZH	EN-DE	DE-EN
BLEU	0	11.79	35.86	10.72	35.05	26.81	25.08
	400	22.06	49.75	19.16	38.95	32.93	43.98
	800	22.39	53.92	19.84	41.68	35.95	46.84
METEOR	0	37.51	44.54	33.11	40.39	48.35	61.69
	400	44.46	64.53	38.50	55.46	52.57	69.93
	800	45.48	66.25	39.32	57.05	54.18	72.72
chrF++	0	33.55	46.79	29.87	44.44	53.90	55.53
	400	44.31	57.76	36.99	46.04	56.74	66.78
	800	45.80	60.02	39.54	47.45	58.64	69.95

Table 5: Performance progression of MT^3-7B-Zero trained simultaneously on three language pairs (ZH↔EN, ZH↔DE, EN↔DE).

F Multilingual Training Experiment

To further demonstrate the scalability and generalizability of our framework, we conducted additional experiments training a single model on three language pairs simultaneously: Chinese-English (ZH↔EN), Chinese-German (ZH↔DE), and English-German (EN↔DE). We used an equal number of MIT-10M training samples for each of the six translation directions and adopted the same training settings as described in the main paper.

The results, detailed in Table 5, show that the model learns effectively across all language pairs and consistently improves with training. For instance, the model achieves strong performance on unseen pairs like ZH ↔ DE and EN↔DE, providing robust evidence that the MT^3 framework is not

[3] https://github.com/vllm-project/vllm
[4] https://github.com/InternLM/lmdeploy

```
<recognize>
铁将军\n39 2.5\n36 2.5\nSTEEL MATE
</recognize>
<think>
The image shows a car dashboard with
a digital display device labeled
\"STEEL MATE.\" The text \"铁将军\"
in the top left corner is the brand
name \"Steel Mate\" in Chinese. The
display shows tire pressure readings
for four tires, with the numbers 39,
2.5, 36, and 2.5, which likely
represent the tire pressure in PSI
and the tire size in inches. The text
\"STEEL MATE\" is the brand name of
the device.
</think>
<translate>
Steel Mate\n39 2.5\n36 2.5\nSTEEL
MATE
</translate>
```


Zero-Start


```
<recognize>
铁将军\n39.25\n35.2.5\n2.5\n39\n2.6\n35
</recognize>
<think>
So I have this image to translate, and it seems to be an advertisement for a car product. The brand name is \"铁将
军,\" which I believe is \"Iron General\" in English. That makes sense because \"铁\" means iron, and \"将军\"
means general, so Iron General is a fitting translation.\n\nNow, looking at the product, it's placed on what
appears to be a car dashboard. The product has a digital display showing various numbers: 39.25, 35.2.5, 2.5, 39,
2.6, and 35. There are also some icons and symbols that suggest it's related to car maintenance or
safety.\n\nFirst, I need to understand what this device does. Given the numbers and the context, it looks like a
tire pressure monitoring system (TPMS). TPMS displays the pressure in tires to prevent blowouts and improve fuel
efficiency. The numbers likely represent the pressure in PSI or BAR for each tire.\n\nLet me confirm that. The
unit \"Bar\" is visible on the display, so it's probably measuring pressure in bars. For example, 2.5 Bar is
equivalent to about 36 PSI, which is a standard tire pressure for many vehicles.\n\nSo, the display shows
pressure readings for four tires: front left, front right, rear left, and rear right. The numbers are:\n\n- Front
left: 39 PSI (or BAR)\n\n- Front right: 25 PSI (or BAR)\n\n- Rear left: 35 PSI (or BAR)\n\n- Rear right: 2.5 BAR
(which is approximately 36 PSI)\n\nWait a minute, that seems inconsistent. If front left is 39 PSI and rear left
is 35 PSI, and front right is 25 PSI, then rear right should be 2.5 BAR, which is 36 PSI. But that doesn't add
up. Maybe there's a conversion error or misreading.\n\nLet me check the conversion. 1 BAR is approximately 14.5
PSI. So, 2.5 BAR would be about 36 PSI, which matches the reading. But the display shows 2.5 BAR and 35 PSI for
the rear left, which seems off.\n\nPerhaps the display is showing different units for different tires. For
example, front left and rear left might be in PSI, while front right and rear right are in BAR. But that seems
unlikely.\n\nAlternatively, maybe the display is showing the tire type or something else. The icons next to the
numbers suggest different tire types or statuses.\n\nLet me look closer at the icons:\n\n- Front left: A tire
with a pressure gauge\n\n- Front right: A tire with a pressure gauge\n\n- Rear left: A tire with a pressure
gauge\n\n- Rear right: A tire with a pressure gauge\n\nAnd there's a central icon that might indicate the overall
status or a specific function.\n\nGiven that, it's likely that the device is a TPMS that monitors tire pressure
in real-time and displays the readings on a digital screen.\n\nNow, considering the brand \"Iron General,\" which
is known for automotive security products, this TPMS fits into their product line. It's probably designed to be
easy to install and use, providing drivers with important safety information at a glance.\n\nThe background of
the image shows a scenic view with a sunset and mountains, which gives a sense of adventure and reliability,
aligning with the brand's image.\n\nIn summary, the text in the image is primarily the brand name \"铁将军\" and
the digital readings from the TPMS device, which are likely in PSI and BAR units, indicating tire pressure levels
for each tire.\n\nTo provide a clear translation, I should list the tire pressure readings as they appear on the
display:\n\n- Front left: 39 PSI (or BAR)\n\n- Front right: 25 PSI (or BAR)\n\n- Rear left: 35 PSI (or BAR)\n\n-
Rear right: 2.5 BAR (approximately 36 PSI)\n\nAdditionally, the brand name \"铁将军\" translates to \"Iron
General,\" which is already provided in the image.\n\nTherefore, the final translation would include the brand
name and the tire pressure readings in both PSI and BAR for clarity.
</think>
<translate>
Iron General\n39.25\n35.2.5\n2.5\n39\n2.6\n35
</translate
```


QVQ-Distill

Figure 8: Case Study. The model initialized with QVQ-Distill tends to generate significantly longer reasoning steps. In contrast, the Zero-start model, after RL training, often develops a more straightforward and less redundant reasoning path.

restricted to a specific language set and scales well to multilingual scenarios. We also observed that the different language pairs did not significantly interfere with each other; instead, all benefited from the multi-task RL training. This aligns with findings in other work regarding the language capacity and learning capabilities of MLLMs.

G Cold-Start Initialization

In this work, we utilize QVQ (Team, 2024) alongside the ground truth TIMT dataset to construct the cold-start dataset. During preliminary experiments, we observed that as a reasoning model, QVQ exhibits limited instruction-following capabilities, rendering it unsuitable for directly applying the MT^3 prompt to generate cold-start data. Instead, we provide QVQ with images and corresponding ground truth source texts from the TIMT dataset, and request only the reasoning process and final translation. We keep most parts of the prompt remain consistent with the MT^3 prompt:

Cold-Start Data Construction Prompt

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. First, the assistant carefully analyzes the recognized text together with the visual elements in the image, taking into account the layout, objects, color schemes, spatial relationships, and other contextual clues that may influence meaning. This integrated understanding ensures that the translation is accurate, coherent, and appropriate to the visual setting. After thorough reasoning based on both textual content and visual context, the assistant provides the user with the final translation in reading order. You should output your final *{target_language}* translation as: Final Translation: Your final translation. You should think step-by-step.
User: *{image}* Translate this text from *{source_language}* to *{target_language}*. *{source_language}*: *{source_text}*

For quality control, we extract the final translation of each sample using the regular expression, compute the BLEU score of each translation

against the ground truth target translation, and select the top 5K samples with the highest BLEU scores for both EN-ZH and ZH-EN directions.

Subsequently, we merge the ground truth source texts with the QVQ-generated reasoning processes and translations, following the MT^3 prompt format to construct the final cold-start dataset. This dataset is further used for the cold-start SFT initialization of MT^3.

H Case Study: Initialization Strategy Impact on Reasoning

We provide a qualitative case study in Figure 8 to illustrate the differences in reasoning patterns that emerge from RL training with different initialization strategies: MT^3-7B-Zero (*Zero-start*) versus MT^3-7B-QVQ-Distill (*QVQ-Distill*). As observed in the examples, the model initialized with QVQ-Distill tends to generate significantly longer reasoning steps within the `<think>` tags. These longer reasoning chains often exhibit self-reflection patterns that appear to be cloned or heavily influenced by the QVQ teacher model. In contrast, the Zero-start model, after RL training, often develops a more straightforward and less redundant reasoning path. This more concise reasoning appears to be directly focused on the core elements necessary for the TIMT task, potentially contributing to its superior performance and more efficient processing.

I Detailed Analysis of XHSPost Example

As illustrated in Figure 2, MT^3 correctly translate culturally nuanced terms (like the TV show *'Friends'* to its official Chinese title *'老友记'* by inferentially linking the visual cue on a phone screen to its broader context, unlike models that produce literal translations such as *'朋友'* (friendship)) and accurately interpret ambiguous numerical information in advertisements (discerning price/location from '800' instead of '800 square meters') stems from this integrated approach.

J Multi-Task Ablation Prompts

The prompts used for multi-task ablation are as follows:

w/o Reasoning

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. First, the assistant recognizes all the text in the image following the natural reading order, separating text from different positions (boxes) with a line break. Then, the assistant provides the user with the final translation in reading order, separating text from different positions (boxes) with a line break. The recognized text and final translation must be enclosed within <recognize> </recognize> and <translate> </translate> tags, respectively. The format must be as follows: <recognize>recognized text here</recognize> <translate>final translation here</translate>
User: *{image}* Translate all the text in this image into *{target_language}* following the natural reading order.

w/o Recognition

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. First, the assistant carefully analyzes the recognized text and the visual elements in the image, considering layout, objects, color schemes, spatial relationships, and other contextual clues that may influence meaning. This integrated understanding ensures the translation is accurate, coherent, and appropriate to the visual setting. After thorough reasoning based on both textual content and visual context, the assistant provides the user with the final translation in reading order. The reasoning process and final translation are enclosed within <think> </think> and <translate> </translate> tags, respectively. The format must be as follows: <think> reasoning process here </think><translate> final translation here </translate>
User: *{image}* Translate all the text in this image into *{target_language}* following the natural reading order.

Only TIMT

System: You are a helpful translation assistant. The user provides an image containing *{source_language}* text and asks for the corresponding *{target_language}* translation. Then, the assistant provides the user with the final translation in reading order, separating text from different positions (boxes) with a line break. The final translation must be enclosed within <translate> </translate> tags. The format must be as follows: <translate>final translation here</translate>
User: *{image}* Translate all the text in this image into *{target_language}* following the natural reading order.

K OOD BLEU Score

Due to the scope limits of the paper, we provide the BLEU score for these OOD evaluations in Table 6.

L Difficulty Levels for Curriculum Learning

The curriculum learning experiments (Appendix E) utilize difficulty categorizations for data instances from the MIT-10M dataset (Li et al., 2024). During the construction of the MIT-10M dataset, samples were classified into three difficulty levels based on heuristics considering the token length of the in-image text and the number of detected bounding boxes for text regions. These categories are defined as follows:

- **Easy:** Instances where the number of bounding boxes is less than or equal to 2, and the token length of the text is less than or equal to 16. These samples generally contain fewer textual elements and shorter texts, resulting in a relatively straightforward translation task.
- **Hard:** Instances where the number of bounding boxes is greater than 5, or the token length of the text is greater than 25. These samples typically feature more textual elements or longer, more complex texts, placing higher demands on the model's attention mechanism and processing capabilities.
- **Medium:** All other instances that do not fall into the Easy or Hard categories. These samples exhibit a wider spread in terms of the number of bounding boxes and text token lengths, representing more realistic and diverse image translation scenarios. This category is particularly useful for testing the generalization ability of multimodal translation models.

M XHSPost Benchmark Construction Details

Social media platforms like TikTok, Instagram, and Xiaohongshu (also known as Red Note) have significantly deepened intercultural communication worldwide. However, there is a notable absence of evaluation benchmarks specifically designed for TIMT in social media contexts.

The XHSPost benchmark was constructed to evaluate TIMT in real-world social media contexts, specifically using posts from Xiaohongshu, a globally-oriented platform. The construction process involved several key steps:

1. **Data Collection**: We targeted posts containing either entirely English or entirely Chinese text. Initial post candidates were gathered by querying keywords relevant to common social media usage patterns, such as 'PLOG' (Photo Log), '中文Plog' (Chinese Plog), and 'TikTok Refugee'.
2. **Content Appropriateness Filtering**: Following Li et al. (2024), all collected images underwent an NSFW detection process using an established tool (Laborde). Any images flagged as potentially inappropriate were discarded.
3. **Text Recognition and Viability Check**: We utilized GPT-4o (OpenAI's `gpt-4o-2024-11-20`) for detailed text recognition from the images. Posts where no precise textual information could be reliably extracted were excluded from further processing.
4. **Personal Data and Sensitive Content Removal**: To mitigate the risk of exposing personal data, images where the OCR-extracted text contained sensitive information (e.g., email addresses, phone numbers) were removed. Additionally, images containing NSFW characters in the recognized text were also excluded to further ensure the dataset's appropriateness.

Model	OOD (Language Pairs)			OOD (Datasets)			
	MIT-10M (EN-DE)	MIT-10M (ZH-FR)	MIT-10M (DE-FR)	OCRMT30K (ZH-EN)	DoTA (EN-ZH)	XHSPost (EN-ZH)	XHSPost (ZH-EN)
Strong Baseline							
Qwen2.5-VL-72B	20.49	9.79	18.78	15.14	45.43	52.97	29.52
InternVL2.5-78B	21.23	5.22	11.16	11.46	38.28	45.61	28.87
Same-size Baseline							
Qwen2.5-VL-7B	22.85	12.09	14.99	8.93	36.90	44.96	24.49
InternVL2.5-8B	16.76	6.68	8.49	7.10	27.25	35.74	19.21
MT^3-7B-Zero	29.50	15.97	26.92	14.21	40.46	49.41	26.94

Table 6: OOD BLEU score performance. Complement chrF++ and METEOR results in Table 2.

5. **Final Dataset Curation**: This multi-stage filtering process yielded a curated set of 106 English posts and 109 Chinese posts.

6. **Translation and Post-Editing**: For each selected post, the OCR-recognized text was translated by GPT-4o, using the corresponding image as contextual input to aid translation accuracy and appropriateness. Crucially, all machine-generated translations underwent manual post-editing by human reviewers to ensure high linguistic quality, fluency, and fidelity to the source text in its visual context. This step was vital for creating reliable ground-truth translations.

This process resulted in the creation of two parallel datasets: XHSPost (EN-ZH) and XHSPost (ZH-EN), designed to support research in social media TIMT. Illustrative examples can be found in Figure 2 and 9. With this work, we have taken an initial step in addressing TIMT for social media scenarios. We anticipate that this benchmark will encourage further research in this area, ultimately contributing to enhanced cross-cultural communication. The benchmark will be released upon acceptance.

Figure 9: Comparison of XHSPost with other TIMT datasets.