CONSERVE: CONTEXT-PRESERVING AND COMPUTE-CONSERVING TEXT IMAGE MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text Image Machine Translation (TIMT) requires models to recognize embedded text and use visual context to resolve translation ambiguities. However, existing approaches leave unclear which visual evidence supports translation decisions and how much context is necessary. We introduce Translation Context Lens (TCL), a causal diagnostic framework based on controlled visual interventions. TCL reveals that visual evidence for translation is distributed yet redundant, with different regions providing overlapping support. This finding motivates ConServe, a framework for preserving translation quality under visual context compression. Spatially balanced token sampling distributes the background budget across regions and selects representative tokens locally, preserving all visual tokens in target text regions. Privileged Context On-Policy Self-Distillation (PC-OPSD) addresses shifts in translation preferences caused by compression. The teacher uses full visual context, correct interpretations, and explanations of incorrect alternatives to supervise predictions on prefixes generated by the student. We also introduce SC-TIMT, a multilingual benchmark for screen translation that depends on visual context. Experiments show that our 9B model outperforms larger multimodal models on SC-TIMT and remains competitive on MCiT with substantially fewer visual tokens. These results suggest that efficient TIMT depends on preserving distributed contextual evidence rather than retaining the entire visual input.

1 INTRODUCTION

Text Image Machine Translation (TIMT) translates source-language text embedded in images into a target language, requiring text perception, visual understanding, and language generation in concert (Ma et al., 2023; Feng et al., 2026). Unlike text-only machine translation, translation decisions in TIMT depend not only on the textual content itself but also on visual cues such as page layout, neighboring text, objects, and their spatial relations (Feng et al., 2026; Li et al., 2026a). A TIMT model must therefore accurately recognize the source text while reasoning over its surrounding visual context to determine its intended meaning in the current scene.

Existing TIMT methods broadly follow cascaded and end-to-end paradigms (Ma et al., 2023). Cascaded systems combine optical character recognition (OCR) with machine translation but suffer from error propagation and limited access to visual context (Ma et al., 2023; Feng et al., 2026). End-to-end methods increasingly leverage Multimodal Large Language Models (MLLMs) to integrate visual and linguistic signals, with recent work exploring preference optimization (Zuo et al., 2025), multi-task reinforcement learning (Feng et al., 2026), structured reasoning (Li et al., 2026a), and global-context modeling with local text crops (Lu et al., 2026).

Despite these advances, existing methods primarily focus on enriching visual inputs or improving visual reasoning, leaving unclear which visual information in a full scene actually influences translation decisions. This is particularly consequential in TIMT, where translation-relevant cues coexist with substantial content only weakly related to the target text. Retaining the full scene incurs unnecessary representational and computational overhead (Zhang et al., 2025b; Yang et al., 2025), whereas aggressive compression may discard context critical to the intended translation. Efficient TIMT therefore requires not only reducing visual computation but also understanding how translation-relevant context is structured and how this structure can be preserved under compression.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To this end, we introduce Translation Context Lens (TCL), a post-hoc causal diagnostic method for MLLMs that applies controlled interventions to background regions outside the target text to assess their contribution to context-sensitive translation decisions. TCL reveals a consistent distributed yet redundant structure: translation-relevant evidence is spread across multiple background regions, while different regions can provide functionally overlapping support and no single region universally determines the translation decision. This finding suggests that efficient visual representations should preserve cross-region contextual coverage while reducing redundant background information.

Building on these findings, we propose ConServe: Context-Preserving and Compute-Conserving Text Image Machine Translation, which combines efficient visual-context modeling with translation-policy preservation for context-sensitive TIMT. For visual efficiency, we introduce Spatial-Quota Farthest Point Sampling (SQ-FPS), a training-free context-preserving token allocation strategy that preserves the target text region, maintains cross-region contextual coverage through spatial quotas, and samples representative visual tokens within each region to reduce redundancy. Beyond preserving visual evidence, efficient TIMT also requires maintaining translation preferences under the visual condition shift introduced by compression. We therefore introduce Privileged Context On-Policy Self-Distillation (PC-OPSD), which leverages privileged supervision from the full visual context and positive and negative context signals to preserve translation preferences under compressed visual conditions. Through on-policy alignment on student-generated states, PC-OPSD mitigates the mismatch between compressed visual representations and full-context translation decisions.

Beyond model design, existing TIMT benchmarks have progressively expanded from document-centric settings (Zhang et al., 2023; Liang et al., 2024) to multilingual and multi-scenario evaluation (Qian et al., 2024; Zhuang et al., 2025; Li et al., 2026a), but are not explicitly designed to isolate cases where visual context determines the intended translation. We therefore construct ScreenContext (SC-TIMT), a multilingual context-sensitive TIMT benchmark for real-world screens, covering English, Japanese, Korean, and Vietnamese to Chinese. SC-TIMT emphasizes examples whose correct translation depends on interface layout, neighboring content, or other visual cues, enabling targeted evaluation of context-sensitive translation. Inspired by prior work on MT error analysis (Freitag et al., 2021; Lu et al., 2024), we further introduce the Error-Aware Translation Agent (EATA), which diagnoses context misinterpretation, content omission, and linguistic errors, and evaluates faithfulness and fluency by accounting for both error type and severity.

Our main contributions are summarized as follows:

- We systematically study how MLLMs use visual context for TIMT and introduce TCL, a causal diagnostic framework for post hoc analysis that assesses the contribution of background regions through controlled visual interventions. TCL reveals a distributed yet redundant structure in the visual context relevant to translation, providing empirical guidance for visual compression that retains relevant evidence while reducing redundancy.
- We propose ConServe, which couples efficient visual representation with policy preservation for context-sensitive TIMT. SQ-FPS preserves the target text region while reducing background redundancy through spatial coverage and representative sampling, whereas PC-OPSD uses the full visual input and positive and negative context conditions as privileged supervision to guide a student using only the compressed representation, preserving context-sensitive translation preferences and mitigating compression-induced policy drift.
- We construct SC-TIMT, a multilingual context-sensitive TIMT benchmark for real-world screen scenarios, focusing on translation cases where visual context determines the intended translation. We further design EATA, which diagnoses fine-grained error types and severities to provide a more diagnostic evaluation of contextual translation quality.

2 Translation Context Lens

Although MLLMs can leverage visual information for translation disambiguation, it remains unclear how translation-relevant visual evidence is distributed spatially and whether different regions provide functionally overlapping or substitutable information. To investigate this, we use TCL for post-hoc causal analysis of MLLM translation decisions. Specifically, TCL keeps the target text region fixed, applies controlled interventions to the background at both global and regional levels, and measures the resulting shifts in the model's preference between context-consistent and context-

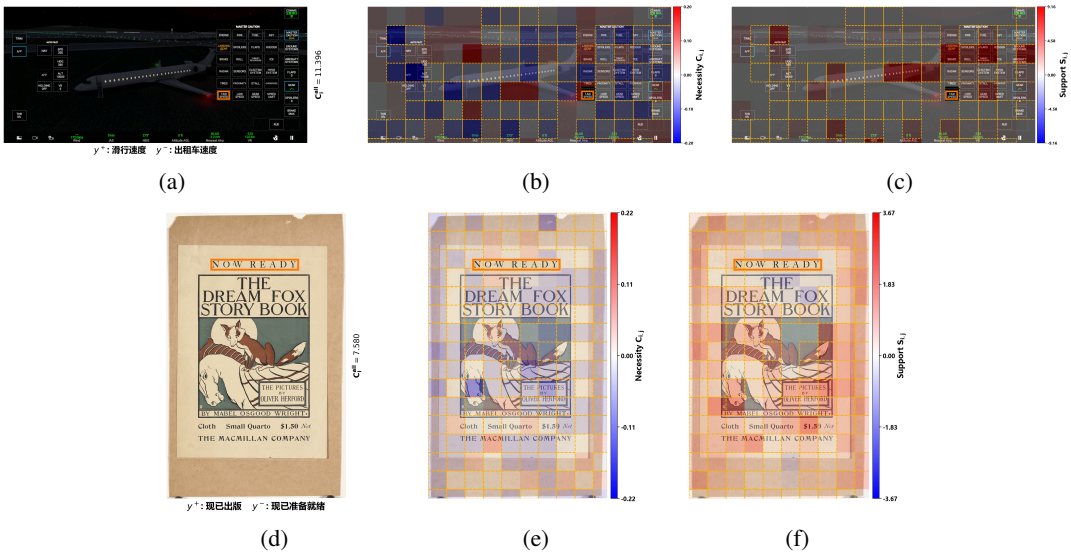


Figure 1: TCL regional intervention examples. Rows show a real-screen and an MCiT example, respectively; columns show the original image, regional necessity $C_{i,j}$, and regional support $S_{i,j}$. Solid orange boxes mark target text regions, while dashed orange boxes indicate regions satisfying $C_{i,j} \leq 0,\ S_{i,j} > 0$. Red/blue denote positive/negative intervention effects.

inconsistent translations, thereby characterizing the spatial organization of translation-relevant visual context. Fig. 1 shows two examples of TCL's regional intervention analysis.

Global Background Effect. We apply TCL to Qwen3.5-9B with all model parameters fixed. Candidate samples are drawn from 5,000 real-world mobile screenshots and 1,450 images from the MCiT benchmark (Zuo et al., 2025). For each paragraph-level target, we generate translations from the full image and from a variant in which the entire background is masked while the target text region is preserved. After manually removing pairs that differ only in wording, formatting, or semantically equivalent paraphrases, we obtain a diagnostic set $\mathcal{T}$ of 990 context-sensitive targets.

For each $\tau_i \in \mathcal{T}$, we construct a fixed pair of competing translations (y_i^+, y_i^-) that captures the key semantic ambiguity exposed by the full-image and background-masked outputs. Here, y_i^+ denotes the translation consistent with the current visual context, while y_i^- denotes the competing translation corresponding to a context-inconsistent interpretation. Since the two candidates may share a common prefix, we evaluate only the portion from their first differing token onward. Let K_i^+ and K_i^- denote the corresponding token positions; the shared prefix is excluded from scoring but retained as autoregressive context. Under visual condition x, we define

$$s_i^{\pm}(x) = \frac{1}{|K_i^{\pm}|} \sum_{t \in K_i^{\pm}} \log p_\theta \left(y_{i,t}^{\pm} \mid y_{i,<t}^{\pm}, x \right), \qquad D_i(x) \quad = s_i^+(x) - s_i^-(x). \tag{1}$$

A positive $D_i(x)$ indicates a preference for the context-consistent translation y_i^+. With the target text region held fixed, let x_{full} and x_{mask} denote the full-image and background-masked conditions, respectively, differing only in the visibility of the surrounding background. We then define

$$C_i^{\text{all}} = D_i(x_{\text{full}}) - D_i(x_{\text{mask}}), \qquad \hat{\mathcal{T}} \quad = \left\{ \tau_i \in \mathcal{T} \mid C_i^{\text{all}} > 0 \right\}. \tag{2}$$

Here, C_i^{all} measures the preference shift induced by the visual background, while $\hat{\mathcal{T}}$ contains targets for which this shift favors y_i^+. We perform the region-wise analysis on $\hat{\mathcal{T}}$.

Regional Necessity. For each $\tau_i \in \hat{\mathcal{T}}$, we partition the background into $\mathcal{G}_i = \{g_1, \ldots, g_j\}$. We remove each region g_j from the full image and measure the resulting change in translation preference. Based on these necessity contributions, we select the Top-k positively contributing regions:

$$C_{i,j} = D_i(x_{\text{full}}) - D_i(x_{\setminus g_j}), \qquad \mathcal{G}_i^{(k)} \quad = \text{Top}_k \left(\{ g_j \mid C_{i,j} > 0 \} \right). \tag{3}$$

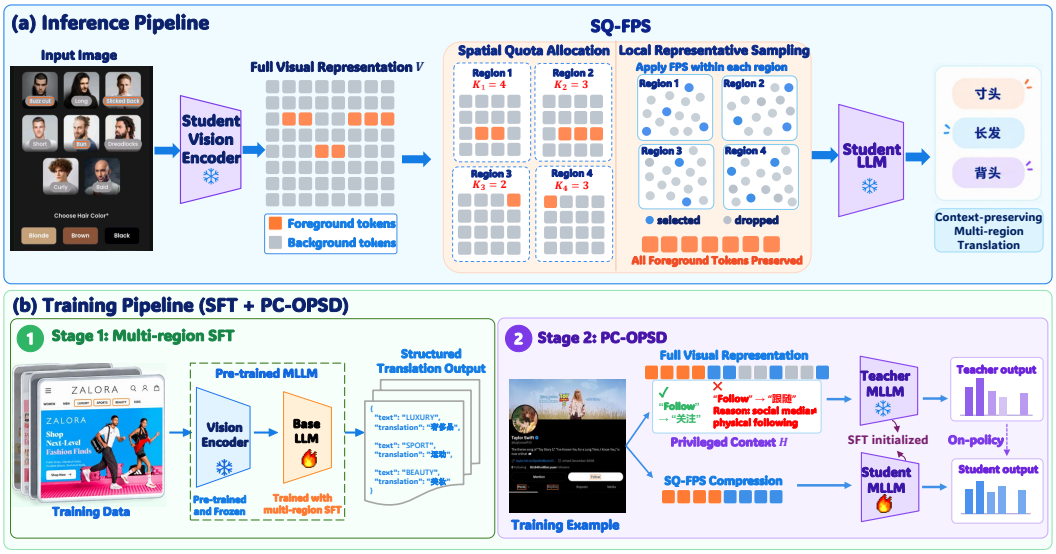


Figure 2: Overview of ConServe. (a) Multi-region inference with SQ-FPS background compression. (b) Two-stage training with Multi-region SFT followed by PC-OPSD.

Here, $C_{i,j} > 0$ indicates that removing g_j weakens the preference for y_i^+, and $\mathcal{G}_i^{(k)}$ collects the k regions with the largest positive $C_{i,j}$. We then jointly remove $\mathcal{G}_i^{(k)}$ and define the resulting effect $J_i(k)$ and normalized ratio $R_i^{(k)}$ as follows:

$$J_i(k) = D_i(x_{\text{full}}) - D_i\left(x_{\setminus \mathcal{G}_i^{(k)}}\right), \qquad R_i^{(k)} = \frac{J_i(k)}{C_i^{\text{all}}}. \tag{4}$$

A value of $R_i^{(k)} \approx 1$ indicates an effect comparable to full-background removal. The Top-1 region yields an average $R_i^{(1)}$ of only 0.1664; even for Top-20, the average $R_i^{(20)}$ is only 0.2996, with 80.85% of targets remaining below 0.5. As a control, randomly removing 20 background regions yields an average $R_i^{(20)}$ of 0.110, substantially below the 0.2996 obtained by the Top-20 regions, showing that the effect is not merely an artifact of region masking. Thus, even jointly removing the most necessary local regions produces an effect far smaller than removing the entire background, indicating that translation-relevant visual contributions are not concentrated in a few regions but exhibit a distributed structure. Fig. 1b and Fig. 1e illustrate this pattern.

Regional Support. Low regional necessity does not imply the absence of useful contextual information. We therefore start from the fully background-masked condition x_{mask} and restore each region g_j individually, denoting the resulting condition by x_{g_j}. We define its independent support as

$$S_{i,j} = D_i(x_{g_j}) - D_i(x_{\text{mask}}). \tag{5}$$

A positive $S_{i,j}$ indicates that region g_j alone increases the model's relative support for y_i^+. We focus on regions satisfying $C_{i,j} \leq 0$ and $S_{i,j} > 0$: removing such a region from the full image does not reduce the current translation preference, yet restoring it alone provides positive support. These regions therefore contain useful visual evidence that can be substituted by information elsewhere in the scene. The orange dashed boxes in Fig. 1c and Fig. 1f show such examples. We find that 98.97% of samples contain multiple such regions, revealing substantial functional overlap and substitutability across regions, indicating strong redundancy in translation-relevant visual context.

TCL Findings. TCL reveals a distributed yet redundant organization of translation-relevant visual context: multiple background regions contribute useful evidence, while many are functionally substitutable. This finding motivates SQ-FPS to preserve broad cross-region contextual coverage while reducing redundant local information through within-region representative sampling.

3 CONSERVE

Guided by the diagnostic findings of TCL, we propose ConServe, an efficient TIMT framework that preserves translation-relevant visual context under visual compression. As illustrated in Fig. 2, ConServe adopts a two-stage training scheme. Multi-region Translation Modeling first learns joint translation of multiple target regions under full visual context, establishing full-context translation capability and yielding an SFT model that initializes both the teacher and student. In the second stage, SQ-FPS preserves distributed contextual evidence by maintaining cross-region coverage while removing redundant visual representations, producing a compact input for the student. PC-OPSD then preserves translation preferences under compressed visual conditions by using a privileged teacher with access to full visual context and additional contextual signals to supervise student-generated states. During inference, the teacher and privileged context are entirely removed. SQ-FPS first constructs the compact visual representation from the input scene, after which the student jointly translates the target regions using only this compressed visual input, retaining contextual translation capability without introducing any teacher-side inference cost.

3.1 MULTI-REGION TRANSLATION MODELING

ConServe uses PaddleOCR (Li et al., 2022)[1] solely for text localization, retaining only the detected bounding boxes. We explicitly mark the current target regions on the original high-resolution image to distinguish them from other text regions in text-dense scenes. Translation is then performed directly from the marked image, preserving fine-grained visual information of the source text and its spatial relationship with the surrounding scene.

For an image I, we partition the detected text regions into target groups of at most K regions, with $K = 3$ by default. Let the q-th group contain $K_q \leq K$ targets with bounding boxes $\mathcal{B}^{(q)}$. We mark these targets on the original image using an annotation operator $\mathcal{A}(\cdot)$ and encode the resulting image with the visual encoder Φ_{vis}:

$$I^{(q)} = \mathcal{A}\Big(I; \mathcal{B}^{(q)}\Big), \qquad V^{(q)} = \Phi_{\text{vis}}\Big(I^{(q)}\Big) = \Big[v_1^{(q)}, \ldots, v_{M_q}^{(q)}\Big], \tag{6}$$

where M_q denotes the number of visual tokens. All target groups share the same full scene and differ only in their marked regions, allowing the model to jointly translate each group while retaining access to the surrounding visual context. As shown in Stage 1 of Fig. 2(b), we use Qwen3.5-397B-A17B (Qwen Team, 2026) to construct 33K multi-region translation examples and perform Multi-region SFT on the base MLLM. We freeze the visual encoder and optimize only the vision–language connector and language model. Given the structured supervision sequence $\mathbf{y}^{(q)} = (y_1^{(q)}, \ldots, y_{T_q}^{(q)})$ for group q, we minimize the standard autoregressive language modeling loss:

$$\mathcal{L}_{\text{SFT}} = -\sum_{q} \sum_{t=1}^{T_q} \log p_\theta \left(y_t^{(q)} \mid y_{<t}^{(q)}, V^{(q)}, P\right), \tag{7}$$

where P is the translation instruction and θ denotes the trainable parameters of the vision–language connector and language model. The resulting Multi-region SFT model initializes both the teacher and student in the subsequent PC-OPSD stage.

3.2 SPATIAL-QUOTA FARTHEST POINT SAMPLING

Motivated by the distributed yet redundant visual context revealed by TCL, we propose SQ-FPS to compress background visual tokens while preserving spatial coverage. For the q-th target group, we partition the full visual representation $V^{(q)}$ into foreground tokens $F^{(q)}$ and background tokens $G^{(q)}$ according to the target text regions. All foreground tokens are retained, while background tokens are sampled with a retention ratio r. SQ-FPS first allocates the background sampling budget across spatial regions according to their size and visual diversity, and then performs local FPS to select representative tokens within each region. The overall procedure is illustrated in Fig. 2(a).

Spatial Quota Allocation. To prevent global sampling from concentrating a limited budget in only a few regions, we partition the background tokens into spatial cells according to their two-dimensional

[1] https://github.com/PaddlePaddle/PaddleOCR

positions and assign each cell a local sampling quota. For a cell $G_c^{(q)}$, let $\hat{v}_i^{(q)}$ denote the normalized feature of token i and $\mu_c^{(q)}$ the mean normalized feature of the cell. We define its visual diversity $d_c^{(q)}$ and quota score $a_c^{(q)}$ as:

$$d_c^{(q)} = \frac{1}{|G_c^{(q)}|} \sum_{i \in G_c^{(q)}} \left\| \hat{v}_i^{(q)} - \mu_c^{(q)} \right\|_2^2, \qquad a_c^{(q)} = |G_c^{(q)}|^\gamma \left(1 + \lambda \widetilde{d}_c^{(q)}\right), \tag{8}$$

where $\widetilde{d}_c^{(q)}$ is the diversity normalized across spatial cells, and γ and λ control the contributions of cell size and visual diversity, respectively. The total background budget is distributed according to $a_c^{(q)}$, yielding a local budget $K_c^{(q)}$ for each cell. When the budget permits, every non-empty cell receives at least one token, while larger or more visually diverse cells receive larger quotas.

Local Representative Sampling. Given $K_c^{(q)}$, we perform FPS independently within each spatial cell to select representative background tokens. For tokens i and j in the same cell, we combine their visual and spatial distances as

$$D_{ij}^{(q)} = \left\| \hat{v}_i^{(q)} - \hat{v}_j^{(q)} \right\|_2^2 + 4\alpha^2 \left\| p_i^{(q)} - p_j^{(q)} \right\|_2^2, \qquad i^* = \arg \max_{i \in G_c^{(q)} \setminus S_c^{(q)}} \min_{j \in S_c^{(q)}} D_{ij}^{(q)}, \tag{9}$$

where $p_i^{(q)}$ is the normalized two-dimensional position of token i, α controls the contribution of spatial distance, and $S_c^{(q)}$ is the current representative set. Sampling continues until $|S_c^{(q)}| = K_c^{(q)}$, favoring tokens that are visually or spatially distinct from the current representatives, thereby reducing redundant selections within each cell. Finally, the selected background tokens are combined with all foreground tokens in their original spatial order to form the compressed visual representation $\widetilde{V}^{(q)}$, which is used as the student visual input in the subsequent PC-OPSD stage.

3.3 Privileged Context Online Policy Self-Distillation

Although SQ-FPS retains the major contextual cues required for translation, the change in visual conditions can still shift the model's translation preference away from that formed under the full visual representation. We therefore propose Privileged Context Online Policy Self-Distillation (PC-OPSD), where a privileged teacher with richer visual and semantic context supervises a student operating only on the compressed representation, with policy matching performed on student-generated states. The training procedure is illustrated in Stage 2 of Fig. 2(b).

Privileged-Context Teacher. Both teacher and student are initialized from the Multi-region SFT model but receive different inputs: the student receives only the compressed visual representation $\widetilde{V}^{(q)}$, whereas the teacher receives the full representation $V^{(q)}$ together with privileged context $H^{(q)}$ available only during training. The privileged context provides two complementary semantic signals: a context-consistent translation sense as positive guidance, and a competing but incorrect sense together with an explanation of why it is inconsistent with the current visual context. These signals are used only to strengthen the teacher's discrimination of ambiguous translations.

Online Policy Matching. At each decoding step, the student autoregressively generates a prefix $y_{<t}^{(q)}$ under the compressed visual condition, and the teacher predicts the next-token distribution conditioned on the same student-generated prefix. Policy matching therefore occurs directly on the states visited by the student. The teacher and student policies are defined as:

$$p_T^t(\cdot) = p_T\left(\cdot \mid y_{<t}^{(q)}, V^{(q)}, H^{(q)}, P_T\right), \qquad p_S^t(\cdot) = p_\theta\left(\cdot \mid y_{<t}^{(q)}, \widetilde{V}^{(q)}, P_S\right), \tag{10}$$

where P_T and P_S denote the teacher and student instructions, respectively. To reduce the cost of full-vocabulary policy matching, we retain the top-m tokens under the teacher distribution and renormalize both distributions over this restricted vocabulary, yielding $\widetilde{p}_T^t$ and $\widetilde{p}_S^t$. PC-OPSD then aligns their token-level policies by minimizing their KL divergence:

$$\mathcal{L}_{\text{PC-OPSD}} = \sum_t D_{\text{KL}}\left(\widetilde{p}_T^t \,\|\, \widetilde{p}_S^t\right). \tag{11}$$

After training, the teacher and privileged context are removed, and the student performs multi-region translation using only the visual representation compressed by SQ-FPS.

Table 1: Results on SC-TIMT. We report Agent Scores and overall SacreBLEU/METEOR scores. Best results are shown in blue boldface; second-best results are underlined.

Method	Size	Agent Score				Overall		
		en	ko	ja	vi	Agent ↑	BLEU ↑	METEOR ↑
GPT-4o	–	4.34	3.90	4.06	<u>4.09</u>	<u>4.16</u>	48.93	70.87
Qwen3.5-35B-A3B	35B-A3B	4.36	<u>3.98</u>	3.97	3.94	4.14	<u>53.23</u>	73.47
Gemma-4-31B	31B	<u>4.37</u>	3.78	<u>4.07</u>	4.02	4.14	52.73	**73.87**
InternVL3-38B	38B	4.07	3.56	3.56	3.78	3.81	35.84	51.75
InternVL3-9B	9B	4.05	3.32	3.52	3.01	3.67	22.65	41.35
Qwen3.5-9B	9B	4.31	3.80	3.93	3.71	4.05	46.68	68.38
Qwen3-VL-8B	8B	4.29	3.79	3.92	3.65	4.04	46.22	65.95
ConServe	9B	**4.56**	**4.03**	**4.21**	**4.09**	**4.32**	**53.83**	<u>73.49</u>

4 EXPERIMENTS

Our experiments investigate three questions: (1) whether ConServe preserves context-aware translation capability with fewer visual tokens; (2) how SQ-FPS and PC-OPSD contribute to effective learning from compressed visual representations; (3) how ConServe achieves a balance between translation quality and efficiency under different compression settings.

Baselines and Implementation. We evaluate ConServe from two perspectives: TIMT performance and visual compression. For translation performance comparison, we compare ConServe with uncompressed MLLMs of different scales, including GPT-4o (OpenAI, 2024), Qwen3.5-35B-A3B and Qwen3.5-9B (Qwen Team, 2026), Gemma-4-31B (Gemma Team, 2026), InternVL3-38B and InternVL3-9B (Zhu et al., 2025), and Qwen3-VL-8B (Bai et al., 2025). For visual compression evaluation, we compare SQ-FPS with baselines, including Random Sampling, global selection methods (K-Means (MacQueen, 1967) and FPS (Gonzalez, 1985)), spatial/local selection methods (MiniBatch K-Means (Sculley, 2010) and spatial/local K-Means variants), and MLLM compression methods (HoloV (Zou et al., 2025), VisPruner (Zhang et al., 2025a), and CRISP (Li et al., 2026b)).

We train ConServe based on Qwen3.5-9B with a two-stage pipeline. Multi-region SFT uses approximately 40K multi-region translation examples with at most $K = 3$ target regions, while PC-OPSD uses 1,365 positive-negative pairs from 1,069 images. During PC-OPSD training, the student receives SQ-FPS representations with $r = 0.5$, and the teacher provides privileged supervision through top-50 forward KL alignment of token-level translation distributions. Unless otherwise specified, inference uses $r = 0.3$. All experiments are conducted on NVIDIA H100 GPUs. Additional implementation details are provided in Appendix C.

Benchmarks and Evaluation. We evaluate ConServe on SC-TIMT and MCiT (Zuo et al., 2025). SC-TIMT contains 1,100 real-world screen images across English, Korean, Japanese, and Vietnamese, covering diverse scenarios with Chinese as the target language. Many samples require visual context for accurate translation. MCiT contains 1,450 English-to-Chinese image translation samples across Document, Scene, and Poster scenarios.

We report SacreBLEU (Post, 2018) and METEOR (Banerjee & Lavie, 2005) for reference-based evaluation. Since these metrics overlook semantic equivalence and error severity (Freitag et al., 2021), we further introduce Error-Aware Translation Agent (EATA), which evaluates Faithfulness and Fluency errors and maps them into a 0–5 Agent Score. On SC-TIMT, we additionally report Contextual Consistency Rate (CCR), defined as the ratio of samples whose translations are judged by an LLM evaluator (Zheng et al., 2023) to be consistent with the visual context. Details of EATA scoring criteria, CCR evaluation protocols, and prompts are provided in Appendix D.

Overall Performance. As shown in Table 1 and Table 2, ConServe achieves strong TIMT performance while using compressed visual representations. It consistently outperforms the base Qwen3.5-9B and remains competitive with much larger MLLMs, achieving an overall Agent Score of 4.32 on SC-TIMT and 4.71 on MCiT. Notably, ConServe surpasses GPT-4o on SC-TIMT by 0.16 Agent Score and achieves the highest SacreBLEU score (53.83), demonstrating that it preserves context-aware TIMT capability under compressed visual inputs.

Table 2: Results on MCiT. We report Agent Scores and overall SacreBLEU/METEOR scores. Best results are shown in blue boldface; second-best results are underlined.

		Document			Scene			Poster		Overall		
Method	Size	Paper	News	Novel	Title	Sign	Introduction	Cover	Leaflet	Agent ↑	BLEU ↑	METEOR ↑
GPT-4o	–	4.89	4.63	4.05	4.96	4.90	**4.81**	**4.60**	4.74	4.70	**46.05**	**65.65**
Qwen3.5-35B-A3B	35B-A3B	4.88	4.52	3.92	4.95	4.92	4.73	4.56	4.73	4.65	43.34	63.20
Gemma-4-31B	31B	4.83	**4.63**	**4.13**	4.96	**4.93**	4.80	4.56	4.76	**4.71**	39.53	59.84
InternVL3-38B	38B	4.84	4.36	3.18	4.93	4.84	4.68	4.42	4.64	4.49	37.99	57.18
InternVL3-9B	9B	4.80	4.33	3.40	4.93	4.78	4.72	4.20	4.65	4.49	25.43	44.37
Qwen3.5-9B	9B	4.86	4.51	3.52	4.93	4.88	4.72	4.53	4.74	4.59	42.76	57.71
Qwen3-VL-8B	8B	4.86	4.50	3.37	4.93	4.88	4.67	4.53	4.72	4.55	40.18	57.99
ConServe	9B	**4.90**	4.59	4.06	**4.97**	4.92	4.80	4.59	**4.79**	4.71	44.72	62.62

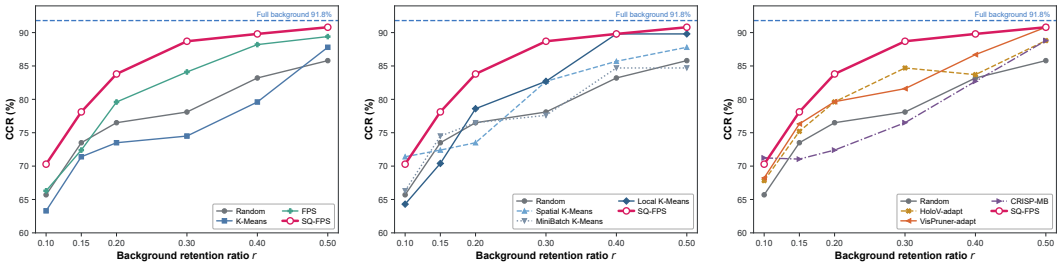
Ablation Studies. We analyze the contribution of SQ-FPS and PC-OPSD through ablation studies. All variants are evaluated on the same test sets as the main results under the same token retention ratio. Results are shown in Table 3. The SQ-FPS ablations show that translation-relevant context preservation requires more than similarity-based selection. Removing Spatial Quota or Diversity Allocation degrades performance on both benchmarks, highlighting the importance of spatial coverage and intra-region diversity. Replacing visual tokens with region centroids leads to the largest degradation, likely due to the loss of fine-grained visual evidence during aggregation. The PC-OPSD ablations verify the importance of preserving translation preferences under compressed visual conditions. Removing PC-OPSD or either supervision source degrades performance, demonstrating the necessity of privileged contextual supervision for mitigating compression-induced preference shifts. Privileged Context has a larger impact, highlighting the value of additional visual and semantic cues beyond the compressed representation. Negative Only outperforms Positive Only, while their combination provides complementary supervision. Finally, removing on-policy rollout further degrades performance, indicating that supervising the student's generated states reduces the mismatch between compressed-visual student policies and full-context teacher decisions.

Table 3: Ablation results of ConServe. Agent Scores on SC-TIMT and MCiT are reported.

Group	Variant	SC-TIMT	MCiT
SQ-FPS	w/o Spatial Quota	4.15	4.57
	w/o Diversity Allocation	4.17	4.50
	w/ Region Centroid	3.99	4.38
PC-OPSD	w/o PC-OPSD	4.01	4.48
	w/o Full Context	4.26	4.63
	w/o Privileged Context	4.08	4.53
	Positive Only	4.10	4.52
	Negative Only	4.22	4.67
	w/o On-policy Rollout	4.14	4.58
	Full Model	**4.32**	**4.71**

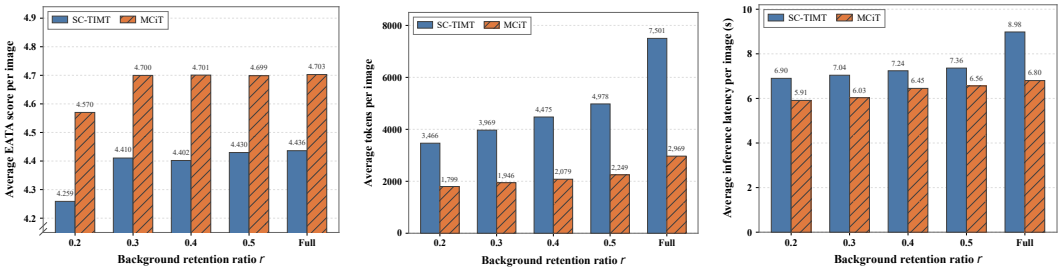
Evaluating Visual Compression. We further compare different background visual token compression strategies under the same token retention ratio, where all methods preserve target text region tokens and are evaluated by CCR. As shown in Fig. 3, SQ-FPS consistently achieves the best performance across different retention ratios. At a low retention ratio ($r = 0.10$), all methods experience noticeable degradation, suggesting that aggressive compression may remove necessary contextual evidence. As the retention ratio increases, SQ-FPS maintains a clear advantage over competing methods. Compared with FPS, which retains the same number of background tokens, SQ-FPS improves CCR by 4.2 and 4.6 points at $r = 0.20$ and $r = 0.30$, respectively, demonstrating that preserving spatially distributed contextual evidence is more important than selecting locally representative features. Other compression methods mainly optimize generic visual importance or token similarity, which may not fully capture the contextual cues required for TIMT disambiguation. At $r = 0.50$, SQ-FPS achieves 90.8% CCR, only 1.0 point below full background input. Together with TCL analysis, these results indicate that translation-relevant visual evidence is distributed yet redundant: effective compression should preserve spatial coverage rather than rely on salient regions.

Compression Robustness. We further analyze the trade-off between translation quality and efficiency under different visual retention ratios. During PC-OPSD training, we fix the SQ-FPS retention ratio to $r = 0.5$, while inference-time retention ratios can be adjusted without additional training. As shown in Fig. 4, ConServe maintains stable translation performance under stronger compression. When reducing r from 0.5 to 0.3, the Agent Scores on both SC-TIMT and MCiT remain close to those with full visual inputs, with notable degradation only under more aggressive



(a) Global Selection **(b) Spatial/Local Selection** **(c) MLLM Compression**

Figure 3: Comparison of visual compression strategies. We report CCR under different token retention ratios across three categories: (a) global selection, (b) spatial/local selection, and (c) MLLM compression. SQ-FPS is used as a reference, with the dashed line indicating full background input.



(a) Translation Quality **(b) Sequence Length** **(c) Inference Latency**

Figure 4: Compression robustness and efficiency trade-off. PC-OPSD is trained with $r = 0.5$, while other retention ratios require no additional training. We report (a) Agent Score, (b) Total Tokens, and (c) latency per image. Full denotes retaining all background visual tokens during inference.

compression. These results indicate that PC-OPSD mitigates translation preference shifts induced by compressed visual representations, enabling robust adaptation across different retention ratios rather than overfitting to a specific compression level. Meanwhile, $r = 0.3$ substantially reduces visual token overhead while preserving translation quality. Compared with full visual inputs, it reduces average total tokens by 47.1% and 34.5% on SC-TIMT and MCiT, respectively, while reducing end-to-end latency to 7.04 s and 6.03 s. Based on this trade-off, we adopt $r = 0.3$ as the inference setting, reducing visual computation while preserving context-sensitive TIMT capability.

5 Discussion and Conclusion

Through TCL, we analyze how different visual regions influence translation decisions and characterize the causal role of visual context under controlled interventions. TCL does not identify universally critical regions, but instead reveals how translation decisions depend on context-specific visual evidence. Our analysis shows that translation-relevant evidence is instance-dependent: a single translation decision may rely on multiple regions, while different regions can provide overlapping support. Therefore, rather than directly converting post-hoc causal signals from individual samples into fixed region selection rules, we adopt a learning-free spatial-aware compression strategy that preserves distributed contextual coverage. Developing dynamic visual evidence extraction models that can further leverage the fine-grained causal signals revealed by TCL remains an open direction.

We revisit how MLLMs utilize visual context in TIMT from the perspective of translation decisions. Through controlled visual interventions with TCL, we reveal that translation-relevant visual context is distributed yet redundant: useful evidence is spread across spatial regions, while different regions may provide functionally overlapping support. Based on this finding, we propose ConServe, which uses SQ-FPS to preserve cross-region contextual coverage while compressing redundant visual representations, and PC-OPSD to mitigate translation policy shifts caused by visual condition changes. Experiments demonstrate that ConServe substantially reduces visual tokens and inference cost while maintaining translation performance. These results suggest that efficient TIMT does not necessarily rely on identifying a fixed set of universally critical regions, but instead requires preserving sufficient distributed contextual evidence while removing redundant visual representations. This perspective may provide insights for other vision-dependent multimodal tasks.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, literature search, and selected steps in constructing our training data. During manuscript preparation, the tools helped polish the language, improve clarity, and adjust some LaTeX formatting. The authors verified all technical claims, experimental results, numerical values, equations, and citations. AI tools also helped identify potentially relevant papers; the authors read each cited paper in its original form and checked its relevance and content before including it.

We also used generative AI to produce translation labels for a subset of training examples. Only labels that passed screening and quality control were used for supervised fine-tuning of the 9B model. The authors reviewed and validated the data processing procedure, criteria for sample selection, and final training data. AI-generated content was not used as an experimental conclusion or a factual statement in the paper without verification.

The authors take full responsibility for the final manuscript, citations, training data, experimental results, and conclusions.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72. Association for Computational Linguistics, 2005.

Zhaopeng Feng, Yupu Liang, Shaosheng Cao, Jiayuan Su, Jiahan Ren, Zhijie Zhou, Wenxuan Huang, Jian Wu, and Zuozhu Liu. MT^3: A synergistic multi-task RL framework for specializing MLLMs in text image machine translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10140–10157. Association for Computational Linguistics, 2026.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. doi: 10.1016/0304-3975(85)90224-5.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 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)*, pp. 3479–3491. Association for Computational Linguistics, 2023.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022.

Gengluo Li, Chengquan Zhang, Yupu Liang, Huawen Shen, Yaping Zhang, Pengyuan Lyu, Weinong Wang, Xingyu Wan, Gangyan Zeng, Han Hu, Can Ma, and Yu Zhou. MMTIT-Bench: A multilingual and multi-scenario benchmark with cognition-perception-reasoning guided text-image machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16593–16602, 2026a.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133:6794–6812, 2025. doi: 10.1007/s11263-025-02491-7.

Xu Li, Yi Zheng, Mengyang Zhao, Yuxuan Liang, Zhe Liu, Rui Zhu, Xiaolei Chen, Wei Zhou, Baoquan Zhao, and Juncen Guo. CRISP: Pre-LLM yet text-driven visual token pruning for efficient LVLM inference. *arXiv preprint arXiv:2607.16326*, 2026b.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 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)*, pp. 7084–7095. Association for Computational Linguistics, 2024.

Junxin Lu, Tengfei Song, Zhanglin Wu, Pengfei Li, Xiaowei Liang, Hui Yang, Kun Chen, Ning Xie, Yunfei Lu, Jing Zhao, Shiliang Sun, and Daimeng Wei. Global-local dual perception for MLLMs in high-resolution text-rich image translation. *arXiv preprint arXiv:2602.21956*, 2026.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8801–8816. Association for Computational Linguistics, 2024.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. E2TIMT: Efficient and effective modal adapter for text image machine translation. In *Document Analysis and Recognition – ICDAR 2023*, pp. 70–88. Springer Nature Switzerland, 2023.

Cong Ma, Yaping Zhang, Zhiyang Zhang, Yupu Liang, Yang Zhao, Yu Zhou, and Chengqing Zong. 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*, pp. 2468–2479, 2024.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297, 1967.

OpenAI. Gpt-4o system card. OpenAI Technical Report, 2024.

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

Zhipeng Qian, Pei Zhang, Baosong Yang, Kai Fan, Yiwei Ma, Derek F. Wong, Xiaoshuai Sun, and Rongrong Ji. AnyTrans: Translate AnyText in the image with large scale models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2432–2444. Association for Computational Linguistics, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, 2026.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 1177–1178, 2010.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025a.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 74840–74857. PMLR, 2025b.

Zhiyang Zhang, Yaping Zhang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. A novel dataset and benchmark analysis on document image translation. In *Machine Translation: 19th China Conference, CCMT 2023*, pp. 103–115. Springer Nature Singapore, 2023.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Wanru Zhuang, Wenbo Li, Zhibin Lan, Xu Han, Peng Li, and Jinsong Su. PATIMT-bench: A multi-scenario benchmark for position-aware text image machine translation in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16572–16588. Association for Computational Linguistics, 2025.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in MLLMs: Revisiting visual holistic context retention. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1327.

Fei Zuo, Kehai Chen, Yu Zhang, Zhengshan Xue, and Min Zhang. InImageTrans: Multimodal LLM-based text image machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20256–20277. Association for Computational Linguistics, 2025.

A Related Work

Text Image Machine Translation. Text image machine translation (TIMT) translates source-language text embedded in an image into a target language. A conventional pipeline first detects and recognizes text before translation. Although effective, this decomposition suffers from error propagation, and visual cues such as layout, neighboring content, and spatial relations become difficult to exploit once images are reduced to plain text. Early end-to-end approaches therefore explored joint modeling of visual and linguistic representations. E2TIMT (Ma et al., 2023) connects a pre-trained recognition encoder with a translation decoder and adapts representations across modalities. Multimodal Codebook (Lan et al., 2023) learns a shared visual–textual representation that leverages textual semantics to compensate for uncertain recognition. BabyNet (Ma et al., 2024) introduces hierarchical supervision from recognition and translation models to transfer knowledge between visual encoding and translation generation. These methods mainly focus on bridging the perception-to-translation gap, while how visual context contributes to individual translation decisions remains less explored.

Multimodal large language models (MLLMs) provide a unified framework for visual perception and language generation. InImageTrans (Zuo et al., 2025) targets repetition and omission errors through multi-condition preference learning. MT^3 (Feng et al., 2026) jointly optimizes recognition, contextual reasoning, and translation through multitask reinforcement learning. CPR-Trans (Li et al., 2026a) organizes scene understanding, text perception, and translation as an explicit reasoning process. GLoTran (Lu et al., 2026) combines a low-resolution global view with multiscale local text regions for high-resolution, text-rich images. These methods improve text perception, completeness, and contextual reasoning, but leave open a more fundamental question: within a full visual scene, how is translation-relevant evidence distributed, and how much contextual information is necessary

for reliable translation decisions? We study this question through causal analysis and leverage the resulting insights to design context-preserving visual compression.

Efficient Visual Compression for MLLMs. High-resolution images retain fine text and rich scene context but produce long visual-token sequences. Recent work reduces the sequence passed to an MLLM through selection, pruning, aggregation, or resampling. TokenPacker (Li et al., 2025) injects high-resolution local features into a compact coarse-to-fine representation. SparseVLM (Zhang et al., 2025b) progressively prunes tokens according to their interaction with the text query. VisionZip (Yang et al., 2025) selects representative tokens from redundant visual-encoder features. Together, these studies show that dense visual representations contain substantial redundancy and can often be compressed while maintaining general multimodal performance. Recent methods also consider where the retained information is located. HoloV (Zou et al., 2025) allocates a compression budget across spatial regions to avoid repeatedly selecting similar salient areas. VisPruner (Zhang et al., 2025a) combines visual importance and feature similarity to balance useful content against redundancy. CRISP (Li et al., 2026b) first selects text-conditioned evidence, then supplements it with semantically diverse scene information. Thus, effective visual compression depends on coverage, representativeness, and task relevance rather than token reduction alone.

TIMT imposes a more specific constraint. In our setting, all visual tokens associated with the target text region are retained; compressed background tokens mainly provide context for interpreting that text. The relevant evidence may lie in neighboring labels, layout, object relations, or visually inconspicuous regions, and may not align with generic visual importance criteria. Consequently, visual salience, global feature representativeness, and text–image similarity need not coincide with the context required for a particular translation. TCL characterizes how such context contributes to translation decisions. Its finding of distributed yet redundant evidence motivates SQ-FPS as a *context-preserving token allocation strategy*: spatial quotas maintain cross-region coverage, while within-region representative sampling removes local redundancy. The objective is translation-specific context preservation rather than generic image reconstruction.

On-Policy Distillation for Vision-Language Models. Knowledge distillation transfers behavior from a teacher to a student. In autoregressive generation, however, inference-time states depend on the student's own generated tokens. Training only on fixed teacher trajectories can therefore supervise a state distribution different from the one the student actually visits. Generalized Knowledge Distillation (GKD) (Agarwal et al., 2024) addresses this mismatch by sampling student trajectories and matching teacher token distributions on those trajectories. On-policy self-distillation (Zhao et al., 2026) further considers asymmetric contexts, where teacher and student share an initialization while the teacher accesses privileged information unavailable to the deployed student.

This principle has also been extended to vision-language models. Vision-OPD (Yuan et al., 2026) transfers fine-grained perception knowledge from a privileged teacher to a student using the student's own rollouts. PC-OPSD adopts a different form of information asymmetry. Its teacher accesses the full visual representation and privileged translation context, whereas its student operates only on the compressed representation available during inference. On-policy supervision therefore targets the translation-preference shift induced by visual compression. It complements SQ-FPS, which reduces redundant visual representations while preserving contextual coverage, by maintaining context-sensitive translation behavior under compressed visual conditions.

B Datasets and Data Construction

Evaluation Sets and Gold Translations. SC-TIMT contains 1,100 screen images with source text in English, Japanese, Korean, or Vietnamese and Chinese as the target language. The images are collected from real-world screen scenarios commonly encountered in everyday translation use, covering a diverse range of applications, including social media and messaging interfaces, e-commerce and shopping pages, travel and transportation applications, food ordering and restaurant menus, news and information pages, entertainment and video platforms, system settings and utility applications, as well as other general mobile and web interfaces. We specifically select examples in which the intended translation cannot be reliably determined from the isolated text alone, but instead depends on visual context such as neighboring interface labels, page structure, icons, objects, or broader scene information. All images are manually screened to exclude user-private or personally identifiable

information. A team of five translation experts provides the Chinese gold translations for SC-TIMT. For MCiT, we use the dataset-provided gold translations as references for evaluation.

Neither SC-TIMT nor MCiT is used in Multi-region SFT or PC-OPSD training. All evaluation images and associated annotations are excluded from training data construction to ensure that the reported results measure generalization to unseen visual-context-sensitive translation examples.

Multi-region SFT Data. We construct approximately 40K candidate multi-region translation examples for supervised fine-tuning. For each training image, up to three target-text regions are grouped into a single example, requiring the model to jointly translate multiple regions while observing the full visual context. Translation supervision is generated by Qwen3.5-397B-A17B, deployed locally on eight H100 GPUs. After data preparation and filtering, the SFT training run uses 33,325 examples. The resulting SFT model establishes full-context translation capability and initializes both the teacher and student models for PC-OPSD training.

PC-OPSD Data Construction. We construct the privileged-context supervision used by PC-OPSD from translation instances whose intended meanings depend on visual context. For each ambiguous translation instance, human annotators identify (i) a contextually consistent interpretation, (ii) a linguistically plausible but contextually inconsistent alternative, and (iii) the visual evidence that distinguishes the two interpretations. Each annotation therefore forms a positive–negative pair that explicitly captures the context-sensitive translation preference. The privileged teacher is provided with the disambiguating contextual evidence, whereas the student must recover the same preference from the compressed visual representation available at both training and inference time. The resulting PC-OPSD dataset contains 1,365 positive–negative pairs, corresponding to 2,730 instances, collected from 1,069 images. We use 1,310 pairs from 1,019 images for training, comprising 1,310 positive and 1,310 negative instances. The remaining 55 pairs, corresponding to 110 instances from 50 images, form the validation split. The PC-OPSD data are collected independently of SC-TIMT and MCiT and are used exclusively for PC-OPSD training and validation.

C Implementation and Inference Details

SFT configuration. ConServe is initialized from Qwen3.5-9B and fine-tuned on 33,325 training examples, with 631 additional examples used for validation. Each example contains at most three target regions. We fine-tune the language model and multimodal projector while keeping the vision encoder frozen. Training is performed on 16 GPUs across two nodes using bfloat16 precision, DeepSpeed ZeRO-2, PyTorch scaled-dot-product attention (SDPA), and gradient checkpointing. We use AdamW with a learning rate of 3×10^{-6}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay of 0.1, a cosine learning-rate schedule, and a warmup ratio of 0.03. The per-GPU batch size is 1 without gradient accumulation, resulting in a global batch size of 16. Training runs for 1.5 epochs (3,125 optimization steps), with gradient clipping at 1.0 and random seed 42. The maximum context length is 8,192 tokens, and the maximum image resolution is $1{,}048{,}576$ pixels. Checkpoints and validation evaluations are performed every 200 steps. We use the `qwen3_5_nothink` template during both training and inference.

PC-OPSD configuration. The teacher and student are initialized from the same SFT checkpoint, with checkpoint 2000 used in our experiments. The student takes the framed image with the compressed visual representation as input, while the frozen teacher additionally receives a low-resolution global image and the privileged context constructed from the positive or negative translation branch described in Sec. 3. The framed image is capped at $1{,}048{,}576$ pixels and the auxiliary global image at $200{,}704$ pixels. For each training example, the student samples one on-policy response using vLLM with temperature 1.0. The teacher evaluates the student-generated response and provides the top-50 next-token log probabilities used for distillation. We optimize the student using the KL distillation objective described in Sec. 3, while keeping all teacher parameters frozen. Training uses 12 GPUs for the FSDP student actor and four GPUs for independent vLLM teacher replicas. We use bfloat16 precision, SDPA, gradient checkpointing, and parameter and optimizer offloading. AdamW is used with a learning rate of 1×10^{-6} and a warmup ratio of 0.05. Training lasts five epochs with a global batch size of 48 and a mini-batch size of 24. The maximum student prompt length is 2,048 tokens, the maximum response length is 768 tokens, and the teacher may use up to 2,560 additional

privileged-context tokens. Unless otherwise specified, the background retention ratio used for the student during PC-OPSD training is $r = 0.5$.

SQ-FPS configuration. SQ-FPS is applied independently to the visual tokens associated with each target group. All foreground tokens are retained, while background tokens are partitioned into a 4×4 spatial grid. We set the cell-size exponent to $\gamma = 0.5$, the diversity weight to $\lambda = 1.0$, and the spatial weighting coefficient to $\alpha = 0.25$. Visual features are ℓ_2-normalized before diversity computation, and token coordinates are normalized to $[0, 1]^2$. Diversity scores are min–max normalized across non-empty cells; when all non-empty cells have identical diversity values, the normalized diversity term is set to zero. The retention ratio r is applied only to background visual tokens, while all foreground tokens are preserved. When the sampling budget is sufficient, each non-empty spatial cell is guaranteed at least one retained token, and the remaining budget is allocated according to the spatial quota defined in Sec. 3. Within each cell, representative tokens are selected using deterministic farthest point sampling with seed 0 and a cell-specific seed offset. We use the `medoid` setting throughout, such that the selected original visual tokens are retained directly rather than replaced by averaged representations. Unless otherwise specified, we use $r = 0.3$ for the main inference results and $r = 0.5$ during PC-OPSD training.

Code availability. To facilitate reproducibility, we provide an anonymous repository at `https://anonymous.4open.science/r/ConServe-ICLR2027-anonymous`. The repository contains the implementation of our proposed SQ-FPS compression method, M3 inference framework, PC-OPSD training pipeline, and the evaluation tools used in our experiments. We provide complete instructions for reproducing the training, inference, and evaluation procedures. The SC-TIMT benchmark will be released upon acceptance.

D Evaluation Details

D.1 Reference-Based and Error-Based Metrics

SacreBLEU and METEOR use the expert-produced Chinese references for SC-TIMT and the dataset-provided references for MCiT. We additionally report the EATA Agent Score to assess meaning-transfer and fluency errors that reference-overlap metrics may miss. The Agent Score is computed from EATA's text branch; its separate visual-rendering, layout, and background assessments are not included.

EATA protocol. The evaluator takes an original image and its translated image as input. An image-text preprocessor transcribes visible text in each image and aligns source and translated spans into evaluation units. Uncertain alignments are retained as uncertain rather than scored as translation errors. Faithfulness and Fluency agents independently propose errors on aligned units. Triggered detail reviews refine these proposals, after which a Text Critic removes duplicate root causes and assigns a final error type and severity. An additional coverage audit checks source text without a translated counterpart. The image-text preprocessor and text agents use `gemini-3.1-pro-preview` with temperature 0.

EATA agent instructions. We report the decision rules from the text-agent prompts that determine the Agent Score. Candidate errors are returned as structured JSON with the affected text pair, error type, severity, evidence spans, and confidence.

Faithfulness. This agent checks whether the translation preserves source meaning, including entities, actions, quantities, conditions, relations, and context-dependent word senses. When text alignment is uncertain, it flags the pair for review instead of guessing a translation error. It does not judge target-language style.

Fluency. This agent checks whether the target-language text is grammatical, natural, and readable. It does not penalize changes in source meaning: a fluent mistranslation is assessed under Faithfulness, while an awkward but semantically correct translation can receive a Fluency error.

Text Critic. This agent adjudicates the proposed errors without retranslating the image. It requires exact source or target evidence, merges reports of the same root cause, and suppresses fluency penalties that are only secondary effects of faithfulness errors.

Detail skills. The `Text Pairing Detail` skill re-examines both images when extraction or alignment is uncertain, without scoring quality. The `Accuracy Detail` and `Fluency Style` skills review their respective candidate errors when severity is Major/Critical, confidence is below 0.65, the region is high-risk, or at least three Minor candidates require the same skill. The former checks semantic transfer and consequential facts; the latter checks independent readability defects. Both may confirm, reject, or revise a candidate before final adjudication. A deterministic coverage check handles unmatched source text separately.

Score computation. For each dimension, let m, M, and C be the counts of adjudicated Minor, Major, and Critical errors. The dimension score is

$$s(m, M, C) = \begin{cases} 0, & C > 0 \\ \max(0, 3 - 0.5(M-1) - 0.3m), & M > 0 \\ \max(0, 5 - 0.3m), & M = 0 \end{cases}. \tag{12}$$

The Agent Score is the mean of the Faithfulness and Fluency scores on a 0–5 scale. Following the implementation, any adjudicated Critical error sets the overall Agent Score to 0.

Human validation. To assess the reliability of EATA as an automatic evaluator, we compared its error judgments with expert annotations on 600 images containing 2,916 text pairs. Against the initial expert annotations, EATA achieved 81.3% precision, 91.9% recall, and 86.2% F1 in translation-error identification, indicating high coverage of expert-identified errors and strong overall agreement at the error-detection level. We further asked experts to re-examine cases in which EATA disagreed with the initial annotations. A substantial portion of these discrepancies was attributable to differences in annotation granularity: EATA often identified local semantic shifts, omissions, and expression defects as separate errors, whereas such issues had been merged or left implicit in the original expert annotations. Consequently, some apparent false positives were confirmed as valid errors upon re-evaluation. In addition, translation experts manually reviewed all errors flagged by EATA on SCTIMT and MCiT and verified the corresponding error explanations. Together, these results support the use of EATA for large-scale diagnostic evaluation. We note that the reported precision, recall, and F1 quantify agreement in error identification rather than correlation with continuous Agent Scores.

D.2 Contextual Consistency Rate

CCR measures the fraction of translations whose intended meaning is compatible with the full image. The judge receives the uncropped image, source text, an expert gold translation, and the candidate translation. The gold specifies the intended reading but is not treated as an exact-string target. We use `chatgpt-5-sol` and obtain ten binary judgments per item. The item-level decision is determined by vote; a 5–5 split is resolved by human review. CCR is the fraction of items assigned a consistent decision after this procedure.

CCR judge prompt. The judge returns a single binary label: 1 for a translation consistent with the intended image-grounded meaning and 0 otherwise. When the image does not resolve an ambiguity, the expert gold anchors the intended meaning; the candidate is not marked inconsistent solely because the image is inconclusive.

System message.

> You are a multimodal translation evaluator. Examine the entire supplied image, including surrounding text, objects, people, actions, layout, and scene context. Determine whether CANDIDATE preserves the intended meaning indicated by GOLD in the image context. Accept natural paraphrases, synonyms, and grammatical differences when meaning is preserved.
>
> Return 0 if CANDIDATE chooses a different sense of an ambiguous word or phrase, contradicts the visual context or semantic relations, omits information needed to preserve the

intended meaning, or adds materially incorrect information. Do not require exact lexical overlap with GOLD or invent visual context. If the image alone does not resolve an ambiguity, use GOLD to determine the intended meaning; do not return 0 solely because the image is inconclusive. Return 1 if CANDIDATE preserves that meaning. Output exactly one digit, 0 or 1, with no explanation, JSON, punctuation, or other text.

User message.

Evaluate the translation using the entire supplied image.
`SOURCE: [SOURCE]`
`GOLD: [GOLD]`
`CANDIDATE: [CANDIDATE]`
Decide whether CANDIDATE preserves the intended meaning of GOLD in the image context. Return only 1 for consistent or 0 for inconsistent.

E ADDITIONAL TCL VISUALIZATIONS

These examples complement Fig. 1. Each triptych shows the original image, regional necessity, and regional support. The labels y^+ and y^- mark the intended and alternative translations; dashed orange boxes mark blocks with nonpositive necessity and positive support. Color scales are specific to each example, so the maps describe local intervention responses rather than a global ranking of regions.

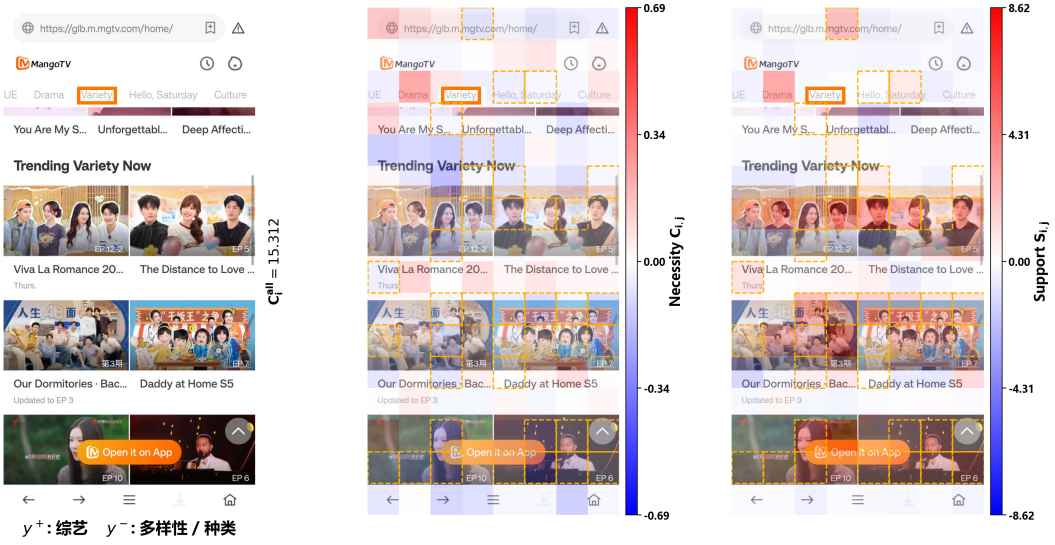


Figure 5: The "Variety" tab in a streaming interface. Program thumbnails and the "Trending Variety Now" heading support the television-genre sense rather than the general sense of diversity or kinds.

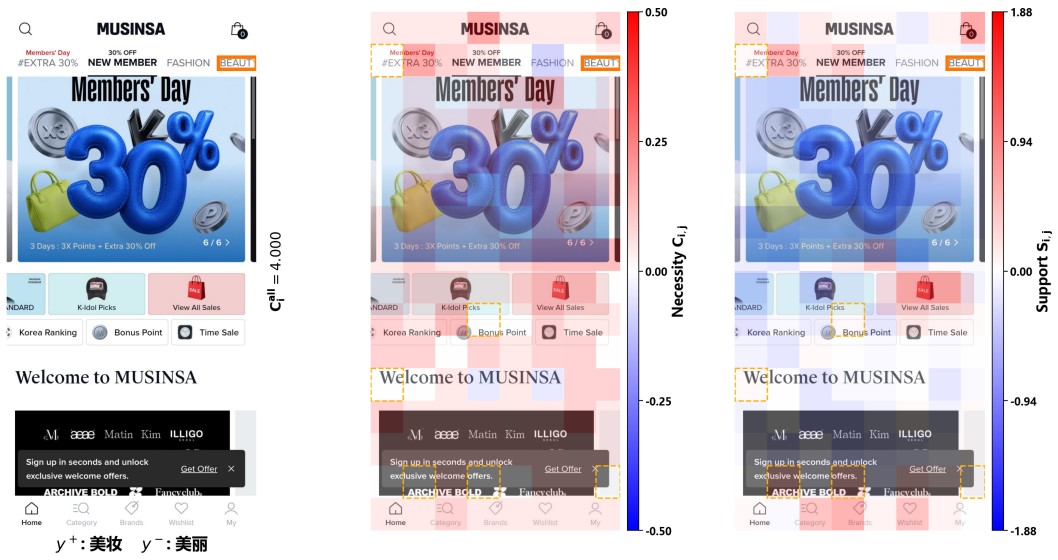


Figure 6: The "Beauty" category in a shopping interface. Navigation labels and merchandise cues favor the cosmetics-category sense over the abstract sense of beauty.

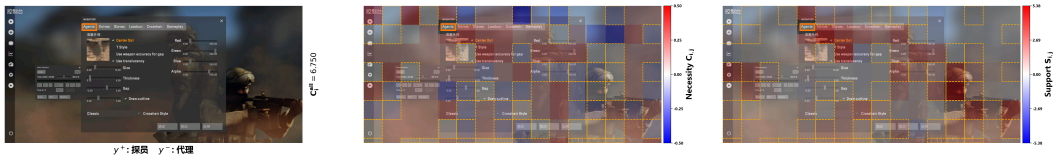


Figure 7: The "Agents" tab in a game interface. The surrounding inventory and character imagery support the game-character sense over a generic agent or proxy sense.

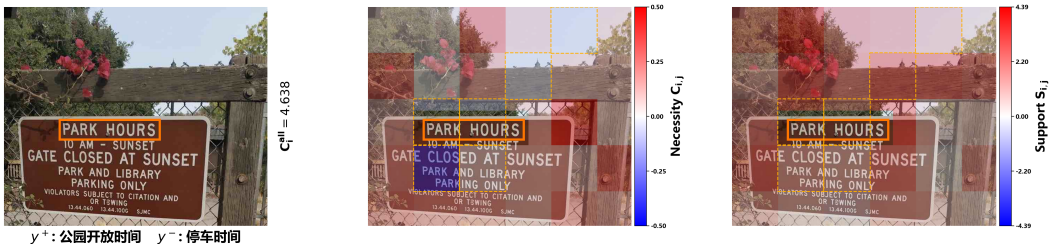


Figure 8: The "PARK HOURS" sign refers to opening hours despite a separate "PARKING ONLY" line.

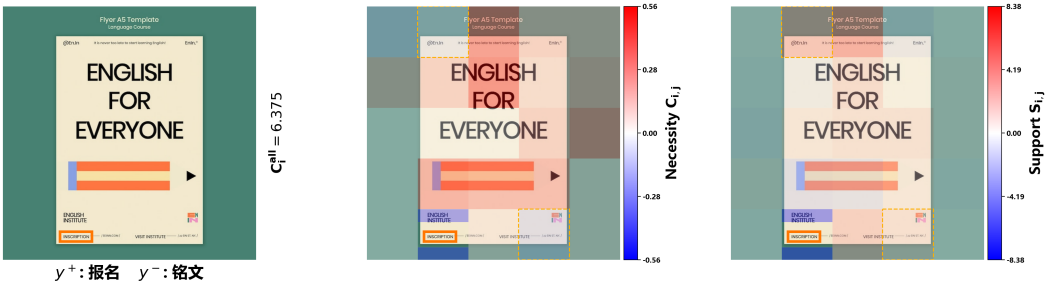


Figure 9: A course leaflet: "INSCRIPTION" denotes registration rather than an engraving.

F Qualitative Examples of Context-Aware Translation



Figure 10: Qualitative comparison on SC-TIMT. Each example presents the complete translated text and EATA evaluation of ConServe, GPT-4o, and Qwen3.5-35B-A3B.