

# MMTIT-Bench: A Multilingual and Multi-Scenario Benchmark with Cognition–Perception–Reasoning Guided Text-Image Machine Translation

## 1. Implementation Details of the CPR-Trans Paradigm

This section details the VLLM-assisted multi-stage generation pipeline used to construct CPR-Trans supervision. The pipeline is designed to elicit structured reasoning traces from a vision–language large model (VLLM) and align them with human-provided OCR and translation labels. It produces the three reasoning stages—cognition, perception, and translation reasoning—that form the core of the CPR-Trans paradigm.

### 1.1. Step 1: Scene Cognition

The first stage of CPR-Trans focuses on establishing a high-level understanding of the visual scene. Given an input image, the VLLM is instructed to summarize the overall layout, key objects, and contextual elements while strictly avoiding any transcription or inference of textual content. This design ensures that visual cognition remains fully decoupled from subsequent text perception, preventing premature leakage of textual information.

To operationalize this constraint, we employ a carefully designed prompt that restricts the model to describing only non-textual visual attributes—such as scene composition, document structure, object arrangement, and the surrounding environment—using a concise, neutral summary. The prompt specification, along with representative scene-cognition outputs, is provided in Figure 1 for reference.

This stage provides a clean, text-agnostic representation of the image, forming a stable foundation for downstream text perception and translation reasoning within CPR-Trans.

### 1.2. Step 2: Text Perception

The second stage of CPR-Trans focuses on reconstructing how a human would progressively perceive and read the textual content within an image. Given the input image, the human OCR annotations, and the *cognition* summary from Step 1, the VLLM is instructed to simulate a detailed, incremental reading process instead of directly outputting the final transcription.

To achieve this, we employ a structured prompt that guides the model to articulate its reading pathway—

including layout navigation, ordering of text regions, merging of broken words, handling of hyphenation and casing, and resolving visually ambiguous characters. The model then concludes with a “final read text” that must match the human OCR annotation in all non-whitespace characters, ensuring that the perception reasoning remains tightly aligned with the target textual content.

This stage outputs a richly interpretable perception trace enclosed in the `<perception>` tag, capturing both *how* the model reads the text and *what* it ultimately reads. Representative prompt specifications and outputs are visualized in Figure 2, illustrating how the model determines reading order, handles accented characters, resolves typographic ambiguities, and reconstructs the stabilized OCR sequence.

By explicitly modeling the visual reading pathway, this stage forms a crucial bridge between scene cognition and downstream translation reasoning, ensuring stable grounding of textual content before linguistic transformation.

### 1.3. Translation Reasoning

The third stage of CPR-Trans focuses on modeling the *translation reasoning process* that bridges visual perception and final multilingual generation. Given the input image, its human-verified OCR transcription, and the scene-level `<cognition>` and text-level `<perception>` outputs from the previous stages, the VLLM is instructed to articulate an explicit chain of translation decisions before producing the final translated text enclosed in `<trans>` tags.

In this stage, a structured prompt enforces a clear separation between *reasoning* and *final translation*. The model must explain how it segments the source content, selects appropriate tone and register, normalizes terminology, resolves ambiguities, handles numerals and proper nouns, and makes punctuation or format adjustments. To avoid memorization or verbatim copying, the prompt further restricts the model from reproducing long spans of source text, instead requiring the use of short semantic labels (e.g., “[Title]”, “[Key Point]”, “[Description]”) when referencing content during the reasoning process. After completing this detailed decision-making trace, the model outputs an English translation that converges to the human-provided target while re-

## Scene Recognition

Input: Image

Prompt=\*\*\*

You are an image content analyst. Your goal is to briefly summarize the visible content of the image without transcribing any text, so that the result can be used for subsequent text extraction.

Task Requirements (strictly follow):

- Describe only the main subjects, scene, layout, and key visual elements (e.g., people, objects, UI components, icons, color scheme, language environment).
- Do not copy, transcribe, or guess any specific text, numbers, full brand names, or dialogue. If text appears, only describe its presence, language, approximate location, or purpose.
- If the image is a document, briefly state the type and structure of the document.
- Use terms such as "may," "likely," or "unclear" when uncertain—do not hallucinate hidden content.
- Length should be 1-3 sentences, objective, neutral, and suitable for guiding later text extraction.

Output Format (must follow exactly):

\*\*\*  
< cognition > Image content summary: provide a 1-3 sentence overview of the main subjects, scene, and layout without transcribing any specific text. < / cognition >  
\*\*\*

Output: "< cognition > xxxxxx < / cognition >"

## Visualize



Figure 1. **Scene Cognition Stage of CPR-Trans.** The figure illustrates the prompt specification and representative outputs generated by the VLLM during the scene-cognition stage. The model provides a high-level visual summary while strictly avoiding transcription or inference of any textual content.

maintaining semantically faithful and stylistically coherent.

Figure 3 presents the complete prompt specification and representative examples generated by the VLLM under this stage. These cases illustrate how the model performs fine-grained translation reasoning—resolving linguistic ambiguities, preserving stylistic cues, and reconciling layout-derived semantics—before arriving at the final translation. This stage completes the CPR-Trans pipeline by aligning visual interpretation, textual perception, and linguistic rendering into a unified and coherent reasoning chain.

### 1.4. Translation Reasoning

The third stage of CPR-Trans focuses on modeling the *translation reasoning process* that bridges text perception and final multilingual generation. In this stage, the model receives three inputs: the image, the human-verified OCR transcription, and the corresponding human translation. The goal is to simulate how a translator makes step-by-step decisions when converting the fixed source text into the target language.

To ensure full decoupling from the earlier < cognition > and < perception > stages, the model is explicitly instructed to treat the OCR transcription as already-resolved textual content, rather than re-performing any form of text recognition. The emphasis is placed entirely on translation-oriented reasoning—including segmentation, terminology selection, ambiguity resolution, stylistic adjustments, and formatting choices—before converging to the final translated output enclosed in < trans >

tags.

Representative prompt specifications and visualization examples demonstrating this process are provided in Figure 3.

For completeness, we also provide a collection of randomly sampled examples along with their full annotations in the supplementary material, enabling reviewers to inspect the detailed behaviors of CPR-Trans across diverse scenarios.

## 2. VLLM-based Judging Setup

This section provides additional details on the VLLM-based evaluation framework used in our experiments, complementing the brief description in the main paper.

### 2.1. Judge Models

We adopt two vision–language large models as automatic judges:

- **Gemini 2.5 Flash** (closed-source), which demonstrates strong multimodal reasoning capability and stable evaluation behavior.
- **Qwen3-VL-235B-A22B-Instruct** (open-source), chosen for its transparent access and robust multilingual understanding.

These two models are selected to reflect both commercial and open-source evaluation settings, ensuring that our results are not biased toward a single family of judge models.

## Text Perception

Input: Image, OCR\_anno, < cognition >

Prompt={

You are an expert in visual text perception. Your task is to simulate how a human would gradually perceive and read the text in an image, based on an existing scene summary, and then produce the final list of transcribed text.

# Given

[Scene summary (from previous step)]

{< cognition >}

[Target text (ground-truth content, may contain line breaks / spacing / case / punctuation differences)]

{OCR\_anno}

# Objective

Using a first-person style such as "I observe... / I read... / I infer...", describe in step-by-step detail how you perceive the text from layout to content (long-chain style), including: 1) browsing order of regions, 2) merging split words, 3) handling line breaks / hyphenation / casing / punctuation, 4) distinguishing similar characters (e.g., 0 vs O, l vs I), 5) and how you resolve uncertain cases.

At the end, provide the "final read text", whose non-whitespace character sequence is line-by-line consistent with {OCR\_anno}. (You may only normalize spaces and line breaks; do not add, delete, or change semantics).

# Output format (only this block, no extra text)

< perception >

[Perception process]

- Browsing order: ...  
- Shape and ambiguity handling: ...  
- Inclusion/exclusion decisions: ...

[Final read text]

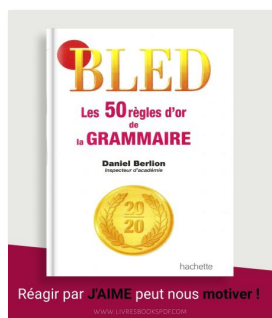
(write line by line following the order of {OCR\_anno}; non-whitespace characters must match the target, only spaces/line breaks may be normalized)

</ perception >

}}

Output: "< perception >xxxxxx</ perception >"

## Visualize



```
< perception >
[Perception process]
- Browsing order: I begin at the top of the cover with the most prominent golden title 'BLED', with a red circular decoration on its left. I then move downward to the pink subtitle lines, reading them in order: 'Les 50 règles d'or', 'de', 'la GRAMMAIRE'. Next, I read the author name 'Daniel Berlion' and the title 'Inspecteur d'académie'. In the middle, inside the golden badge, the stacked numbers '20' over '20' are clearly visible. At the bottom right is the publisher 'hachette'. Finally, my sight moves to the dark red banner across the bottom of the image, where I read the white slogan 'Réagir par JAIME peut nous motiver!' and the website 'www.LIVRESBOOKSPDF.COM'.
- Shape and ambiguity handling: I confirm that 'BLED' is fully uppercase with no ligatures; the 'è' in 'règles' is an accented vowel, not a plain 'e'; the apostrophe in 'd'or' is a French contraction, not a broken word; 'GRAMMAIRE' is all uppercase and remains semantically connected to 'la', even though they appear on separate lines; within the badge '20/20' appears visually as two independent lines '20' and '20', not as a fraction; the apostrophe in 'JAIME' reflects standard French usage; 'WWW' in the URL is visually uppercase, but I convert it to lowercase 'www' to match the target format; all punctuation (e.g. exclamation mark, apostrophes) is preserved; no ambiguous characters such as 0/O or l/I are detected.
- Inclusion/exclusion decisions: I strictly follow what is visually displayed, without adding or removing any semantic characters. Line breaks (such as 'de' appearing alone, or the two '20' lines) are retained. Spacing is normalized only for readability without altering characters. I confirm that 'JAIME' is uppercase, consistent with the target. Although the URL appears visually with 'WWW' in uppercase, I match the target format by writing 'www.LIVRESBOOKSPDF.COM', which preserves the intended character sequence.
[Final read text]
# BLED
## Les 50 règles d'or
de
### la GRAMMAIRE
Daniel Berlion
Inspecteur d'académie
20
20
hachette
Réagir par JAIME peut nous motiver !
www.LIVRESBOOKSPDF.COM
</ perception >
```

Figure 2. **Text Perception Stage of CPR-Trans.** The figure presents the prompt specification used to guide the VLLM in simulating a human-like text-reading process, followed by a representative perception example. The model integrates the scene-level summary with OCR annotations to infer reading order, merge fragmented words, resolve typographic ambiguities, and produce a structured perception trace that aligns line-by-line with human OCR annotations.

## 2.2. Judging Inputs

Each judge receives the following three inputs:

1. the *original image*;
2. the *human reference translation*;
3. the *system-generated translation*.

For models trained with intermediate reasoning (e.g., Simple CoT, Distillation, CPR-Trans), we provide only the final translated text to the judge, explicitly excluding any internal reasoning traces. This ensures that the evaluation focuses purely on translation quality and prevents judge models from being influenced by the presence or style of chain-of-thought outputs.

## 2.3. Evaluation Dimensions

Following a unified scoring rubric, each judge assigns scores along four complementary dimensions:

- **Faithfulness** — whether the translation accurately preserves the meaning and key information of the source content.
- **Fluency** — grammatical correctness and naturalness of the translation in the target language.

- **Readability** — clarity, coherence, and stylistic suitability of the translation for human readers.
- **Terminology Consistency** — correctness and coherence of domain-specific or named-entity translations.

These dimensions jointly reflect both semantic adequacy and linguistic quality, offering a more holistic assessment than conventional word-overlap metrics.

## 2.4. Score Aggregation

For each translation instance, the judge produces four numerical sub-scores. The final score is computed as the **simple average** of the four dimensions:

$$\text{Score} = \frac{1}{4} (\text{Faithfulness} + \text{Fluency} + \text{Readability} + \text{Terminology})$$

This averaging scheme provides a balanced view of translation quality without overemphasizing any single aspect.

## 2.5. Evaluation Prompt

For reproducibility, we provide here the exact VLLM evaluation prompt used in all experiments. The judge model

### Translation Reasoning

Input: Image, OCR\_anno, Trans\_anno

Prompt=#

You are a professional English translator. You already have the "text perception" result from the previous step (the source text has been fixed). In this step, your only task is to perform translation reasoning and convergence from the source text into English.

# Given

[Source text (final text list from Step 2; provided only as the basis for translation, do NOT restate it)]

{OCR\_anno}

[Target English (reference / training target)]

{Trans\_anno}

# Objective

- Without restating or copying the source text line by line, simulate the translation decision-making process: segmentation into chunks, tone and register, style choices, terminology consistency, disambiguation, treatment of numbers/units/proper nouns, and punctuation/format decisions.

- Finally, produce an English result that converges to the 'Target English'

# Strict Constraints

- You MUST NOT copy or re-list large portions of Source text in your output. If you need to refer to the source, use SHORT LABELS (e.g., [Title], [Button]), [Description]) and avoid quoting more than 3 consecutive original words.

- You MUST NOT use these words or their synonyms: OCR, annotation, original text, line number, alignment, source, reference, ground truth, verbatim, provided data.

- The translation is ONLY for the English result. Do not embed explanations/notes into the final English paragraph; explanations should appear ONLY in the "Translation Reasoning" section.

- Proper nouns and models: explain your handling strategy in the reasoning; the final result should follow the 'Target English'.

- Output exactly ONE block, which MUST be wrapped with '<trans> ... </trans>'. Do not output any other tags or extra text.

- Make the decision process as detailed as possible.

# Output Structure (STRICT)

<trans>

[Translation Reasoning]

- Segmentation Strategy: Using short labels (such as [Title] / [Key Point] / [Button]), explain how you segment the source and what English style you choose for each (formal / informal / product copy, etc.), without copying the original text.

- Terminology and Format: Explain your unified strategy and examples (using short labels) for numbers/units (% , \$ , kg , GB , etc.), time, currency, and proper nouns (brands / person names / organizations).

- Ambiguity and Choices: List potential ambiguities and candidate translations, and explain why you chose the final wording (based on context, common sense, and layout hierarchy).

- Punctuation and Readability: Explain how you handle differences between Chinese and English punctuation, long sentence splitting, bulleting, etc.

- Quality Self-check: Check for accuracy (no omissions or mistranslations), consistency (terminology and tone), and naturalness (fluent and easy to read).

[Final English Result]

(Give your final English translation line by line; it must be semantically equivalent to the Target English. For strict character-level consistency, see the optional constraint below.)

</trans>

#

Output: "<trans>xxxxxx</trans>"

### Visualize



```
<trans>
- Chunking strategy: For the [Title] "Stalingrad -", I keep a direct translation of the historical place name without added modifiers; for the [Subtitle] "diesmal ohne u", I translate it as "this time without u", preserving the colloquial shorthand "u" to match the original informal tone; for the [Explanation section], I enclose it in parentheses and translate it as "and also without our tanks and tax money", maintaining the parallel structure and the colloquial "our" to convey the intended sense of ownership.
- Terminology and formatting: The proper noun "Stalingrad" is kept unchanged; "Panzer" is translated as "tanks", which is the standard rendering in military contexts; "Steuergelder" becomes "tax money", which is more colloquial than "tax funds" and fits the tone; there are no numbers or units requiring additional processing.
- Ambiguity and choices: In "ohne u", the "u" could mean "you" or "us" (us). Considering the later use of "unsere" (our) and the overall context, translating it as "without u" (without you) better matches the ironic tone that implies shifting responsibility. Translating it as "without us" would be redundant with the later phrase "our tanks", so that reading is rejected.
- Punctuation and readability: I preserve the original dash and parenthetical structure; in English, a space is kept after the dash and line breaks within the parentheses follow the source formatting: "and" links coordinated elements to avoid a long run-on sentence and improve readability.
- Quality self-check: No omissions or mistranslations; terminology such as "tanks" and "tax money" is consistent and context-appropriate; the tone remains colloquial and aligns with the original ironic style; punctuation and formatting match the target output.
# Stalingrad
## this time without u
(and also without our tanks
and tax money)
</trans>
```



```
<trans>
- Chunking strategy: The [Title] is rendered as a product name plus a descriptive phrase in a commercial signage style; the [Key points] (火・木・土) are converted to English weekday names, keeping them concise; the [Time] retains the 24-hour format, joined by a hyphen; the [Information note] is phrased as an instructive sentence with a polite, formal tone; the [URL] is kept unchanged; the [Asterisked clauses] are translated as disclaimer-style notes introduced with an asterisk, maintaining a neutral and precise tone.
- Terminology and formatting: The weekdays 火・木・土 are translated as 'Tuesday, Thursday, Saturday', consistent with international conventions; the times '15:00 17:00' are written as '15:00 - 17:00' to fit typical English formatting for time ranges; 'HP・SNS' becomes 'HP, SNS, etc.', retaining the abbreviations and separating them with commas; '特設' is translated as 'designated area', which is semantically accurate and common in parking instructions; '一切責任を負いません' is translated as 'not responsible for any accidents', which is clear and concise.
- Ambiguity and choices: The brand name 'ふくぐる' is transliterated as 'Fukuguru', consistent with proper noun handling; 'その他臨時休業あり' could be interpreted as either 'temporary closure' or 'temporary break', but given the surrounding context (latest information) and the business scenario, 'temporary closures' is the more appropriate translation; 'ご覧下さい' being a polite request, is translated as 'please check' for a natural and respectful tone.
- Punctuation and readability: The Japanese middle dots '・' are converted into English commas; Japanese full stops '。' become English periods; a hyphen is added between time ranges for clarity; asterisks introducing clauses are preserved to match standard English cautionary formatting; the long sentence '最新情報はHP・SNS等をご覧下さい' is split for readability and to avoid overlong structures.
- Quality self-check: No omissions or mistranslations; terminology is consistent (e.g., 'HP, SNS', 'designated area'); tone is coherent throughout, maintaining a commercial notice style; the final rendering reads naturally to English speakers without awkward literalism.
# Wood-fired Oven Bread Fukuguru
## Tuesday, Thursday, Saturday
## 15:00 - 17:00
Other temporary closures may occur
For the latest information, please check our HP, SNS, etc.
http://fuku-pan.com
*Please park your car within the designated area.
*We are not responsible for any accidents within the parking lot.
</trans>
```

Figure 3. Translation Reasoning Stage of CPR-Trans. The figure illustrates the structured prompt and representative outputs from the VLLM during the translation reasoning stage. The model first provides a detailed chain of translation decisions—including segmentation, terminology normalization, ambiguity resolution, and stylistic choices—before generating the final translated output within <trans> tags.

receives three inputs: (1) the original image, (2) the reference translation, and (3) the system-generated translation.

The judge is instructed to score the model translation along four dimensions and to output a structured evaluation consisting of a "scoring process" section and a final averaged score. The complete English evaluation prompt is provided below.

## 2.6. Evaluation Prompt

To ensure consistent behavior across different judge models, we design a unified and carefully structured evaluation prompt. This prompt explicitly defines the judge's role, required evaluation steps, scoring criteria, and output format. The full prompt used in all experiments is provided below for completeness.

## Evaluation Prompt.

You will be given the **original image**, the **reference translation**, the **evaluation criteria**, and the **scoring rules** (0–5 scale). Your task is to evaluate the translation provided in `<Model Translation>` and produce the final result inside `<Evaluation Output>`.

Your output **must strictly follow** the structure shown in the scoring example, consisting of two sections: (1) **Scoring Process** and (2) **Final Score**.

In the “Scoring Process” section, briefly identify the weaknesses of the model translation and explain them concisely. Then evaluate the translation along each dimension independently and assign a numerical score for each criterion. In the “Final Score” section, compute the average score over all four dimensions.

You must not modify, add, or invent any evaluation criteria or scoring rules.

### Evaluation Criteria

1. **Faithfulness:** Whether the translation fully and accurately conveys the meaning of the source content.
2. **Fluency:** The grammatical correctness and naturalness of the translation in the target language.
3. **Readability:** Whether the chosen expressions follow the cultural and stylistic norms of the target language.
4. **Terminology Consistency:** Correct and coherent translation of domain-specific terms, named entities, and institutional references.

**Scoring Rules (0–5)** Scores must be assigned strictly. Faithfulness serves as the foundational criterion—if it receives a low score, all other scores must be *less than or equal to* it.

- **0 — Very Poor:** Severe mistranslations; large portions missing; translation is incoherent or unrelated to the source.
- **1 — Poor:** No catastrophic errors but significantly flawed; key terms mistranslated; meaning coverage insufficient; multiple awkward or incorrect expressions.
- **2 — Fair:** Generally understandable; meaning mostly preserved; contains minor issues such as slight misuse of non-technical words, minor named-entity errors, or small omissions.
- **3 — Good:** Accurate, fluent, and clear overall; conveys the source meaning well; only minor issues remain.

- **4.0–5.0 — Excellent:** Highly accurate, natural, culturally aligned, and terminologically consistent; close to human-level translation quality.

### Important Notes

- The model translation may include layout descriptions or repeated source content—**ignore these completely**. Evaluate only the actual translated text.
- Numbers and LaTeX formulas do not require translation and should not affect the score.

### Scoring Example

```
<Model Translation>:
[[model-generated translation here]]
```

```
<Evaluation Output>:
Scoring Process:
---- Begin Scoring ----
Feedback on issues and suggestions:
[[brief explanation of weaknesses
and how to improve]]
```

```
1. Faithfulness:
[[12 sentence justification]] score
: X
```

```
2. Fluency:
[[12 sentence justification]] score
: X
```

```
3. Readability:
[[12 sentence justification]] score
: X
```

```
4. Terminology Consistency:
[[12 sentence justification]] score
: X
```

```
Overall final score (average of the
four scores):
```

```
[[ (X1 + X2 + X3 + X4) / 4 = Final
Score ]]
```

```
---- End Scoring ----
```

```
Final Score: [[Final Score]]
```

### Evaluation Input:

```
Original Image: {image}
Reference Translation: {reference}
```

```
<Model Translation>:
{system_output}
```

```
<Evaluation Output>:
```

## 2.7. Reproducibility and Stability

To minimize judge variance, we:

- fix the temperature to 0 for deterministic scoring;
- standardize prompt ordering and formatting across all instances;
- random-check sampled judgments to ensure alignment between judge explanations and assigned scores.

Across all experiments, we observe stable behavior from both judges, with consistent score tendencies across systems and languages.

## 3. Multilingual Text Parsing Evaluation

Beyond translation, we also construct *human-verified structured OCR annotations* for the entire MMTIT dataset. These annotations enable **MMTIT-Bench** to serve not only as a translation benchmark but also as a rigorous evaluation suite for *multilingual, multi-scenario text parsing*.

We adopt a standard edit-distance evaluation protocol and perform a systematic comparison of the VLLMs and OCR expert models referenced in the main paper. The results are summarized in Table 1.

Model	Param	Thinking	Score
<i>General VLLMs</i>			
Gemini 2.5 Flash [2]	-	✓	<b>82.64</b>
Qwen3-VL-235B-Instruct [8]	235B-A22B	-	81.40
Qwen3-VL-235B-Think [8]	235B-A22B	✓	81.98
Seed1.6 [6]	-	-	81.78
Seed1.6-Thinking [6]	-	✓	82.01
MiMo-VL [7]	7B	-	78.45
dots.vlm [5]	671B-A37B	✓	81.89
<i>OCR-specialized Models</i>			
MinerU 2.5 [3]	1.2B	-	52.05
dots.ocr [4]	3B	-	77.50

Table 1. **Multilingual text parsing accuracy on MMTIT-Bench.** Scores are computed using edit distance against human-verified OCR annotations. General-purpose VLLMs demonstrate strong multilingual parsing ability, often outperforming dedicated OCR models in complex multilingual and natural-scene conditions.

We observe a clear *positive correlation between parsing ability and translation quality*: models achieving higher translation scores on MMTIT-Bench also tend to demonstrate stronger multilingual text parsing.

Moreover, general-purpose VLLMs consistently outperform specialized OCR systems. This is largely because existing OCR models are primarily optimized for *digital-born or scanned document* scenarios, whereas MMTIT-Bench spans a significantly broader set of conditions—including natural scenes, posters, signage, and diverse multilingual

layouts—which introduces substantial challenges for conventional OCR pipelines.

## 4. Translation Visualization

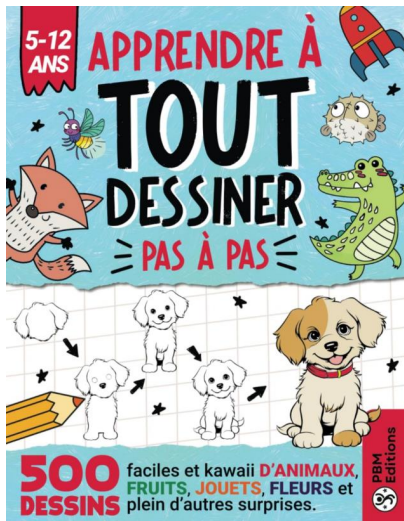
To provide an intuitive understanding of the qualitative improvements introduced by CPR-Trans, we include a set of representative examples from the MMTIT-Bench in this section. For each selected test instance, we visualize the original image alongside three system outputs obtained under the **Qwen2.5-VL-7B [1]** setting:

1. **Direct Translation** — translating the image content end-to-end without any intermediate processing;
2. **Simple CoT(Parsing-then-Translation)** — first parsing visible text elements, then translating the parsed result without any reasoning;
3. **CPR-Trans** — our proposed multi-stage pipeline aligned with scene cognition, text perception and translation reasoning.

These visualizations allow readers to directly observe the differences in translation behavior across methods. Notably, CPR-Trans produces translations with substantially higher semantic fidelity, improved handling of multi-script or noisy text, and better cultural and stylistic alignment—particularly on examples involving dense layout, multilingual fragments, or visually degraded text. Representative examples are provided in Figures 4, 5, and 6.

## References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [2] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasapat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6
- [3] Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025. 6
- [4] rednote hilab. dots.ocr: Multilingual document layout parsing in a single vision-language model, 2025. 6
- [5] rednote hilab. dots.vlm1: The official repository of the dots.vlm1 instruct models, 2025. GitHub repository. 6
- [6] ByteDance Seed Team. Seed 1.6: A universal multimodal deep-thinking foundation model, 2025. Foundation model released by ByteDance Seed Team. 6
- [7] LLM-Core-Team Xiaomi. MIMO-VL technical report, 2025. 6
- [8] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6



### Direct Translation

5-12 岁  
 ## 学习  
 # 画遍所有  
 ### 一步一步  
 ## 500  
 简单又可爱的动物、水果、玩具、鲜花  
 以及  
 许多其他惊喜。

### Simple CoT

```
<ocr>
### 5-12 ANS
## APPRENDRE À
# TOUT DESSINER
## PAS À PAS
## 500 dessins
faciles et kawaii D'ANIMAUX,FRUITS, JOUETS, FLEURS et
plein d'autres surprises.
</ocr>
<trans>
### 5-12 岁
## 学习
# 画遍所有
## 一步一步
## 500 幅画
简单又可爱的动物、水果、玩具、鲜花和
许多其他惊喜。
</trans>
```

### CPR-Trans

<cognition>

图片内容概括：封面以浅蓝色为背景，主体为大号红黑字体标题，配有多只卡通动物（狐狸、青蛙、狮子等）及铅笔、星星等装饰元素；底部展示小狗的绘画步骤图解，右下角有出版社标识，整体为儿童绘画教程风格。

</cognition>

<perception>

[感知过程]

- 版面浏览顺序：从左上方红色标签“5-12 ANS”开始，向右读取主标题“APPRENDRE À”，再向下逐行扫描“TOUT DESSINER”“PAS À PAS”，接着观察底部大号数字“500 DESSINS”与描述性文字“faciles et kawaii D'ANIMAUX, FRUITS, JOUETS, FLEURS et plein d'autres surprises.”，最后定位右下角竖排“PBM Editions”。

- 形态与歧义处理：确认“5-12”为连字符连接的数字范围，“ANS”为独立词；“APPRENDRE À”中“A”带重音符，非“A”；“TOUT”与“DESSINER”为独立词，无连字；“PAS À PAS”中两个“A”均带重音；“500”为纯数字，无字母混淆；“D'ANIMAUX”中撇号为缩写标记；“FRUITS, JOUETS, FLEURS”为逗号分隔列表；“plein d'autres surprises.”末尾句点清晰；“PBM Editions”中“Editions”为竖排，字母无变形。

- 取舍说明：忽略装饰性星星、箭头、动物轮廓等非文字元素；不合并或拆分任何单词；保留所有标点与重音符；“500”与“DESSINS”虽在视觉上被图形部分遮挡，但根据上下文与字体结构判断为完整词；“kawaii”为日语借词，按原拼写保留；“et”重复出现，均保留；出版社名“PBM Editions”按竖排顺序读取，不调整为横排。

[最终读到的文字]

5-12 ANS

## APPRENDRE À

# TOUT DESSINER

## PAS À PAS

# 500 DESSINS

faciles et kawaii D'ANIMAUX, FRUITS, JOUETS, FLEURS et plein d'autres surprises.

PBM Editions

</perception>

<trans>

- 分块策略：[年龄标识]采用“5-12岁”口语化数字表达；[主标题]“APPRENDRE À TOUT DESSINER”拆为“学习+画万物”，强化动作与范围；[副标题]“PAS À PAS”译为“循序渐进”，贴合教学语境；[内容亮点]“faciles et kawaii...”处理为“简单又可爱的...”，保留数量与形容词并列结构；[出版社]“PBM Editions”统一译为“PBM 出版社”，符合中文出版物惯例。

- 术语与格式：数字“5-12”保留连字符，单位“ANS”译为“岁”；“kawaii”不直译，采用“可爱”传达情感色彩；“D'ANIMAUX, FRUITS...”等类别词首字母大写转为中文无大小写处理，用顿号分隔；“plein d'autres surprises”译为“许多其他惊喜”，避免直译“充满其他惊喜”的生硬感。

- 歧义与取舍：“TOUT DESSINER”可译“画所有东西”或“画万物”，选后者更简洁有力；“faciles et kawaii”若译“容易且可爱”略显生硬，调整为“简单又可爱”更口语自然；“plein d'autres surprises”若译“充满其他惊喜”语义正确但冗长，取“许多其他惊喜”平衡准确与流畅。

- 标点与可读性：法文逗号、句号转为中文全角标点；

- 质量自检：无漏译（覆盖所有文本块）；术语统一（如“kawaii”始终译“可爱”，“Editions”始终译“出版社”）；语气一致（全篇保持儿童向、轻松活泼口吻）；自然度高（符合中文儿童读物标题习惯，无欧化句式）。

5-12 岁

## 学习

# 画万物

## 循序渐进

# 500幅画

简单又可爱的动物、水果、玩具、花卉和其他惊喜内容。

PBM 出版社

</trans>

Figure 4. Qualitative example of CPR-Trans on an MMTIT-Bench sample(other2zh)-Case1.



Direct Translation
# 餐馆
LO BIANCO 有限责任公司
E. AMARI 街 104 号
90139 巴勒莫 (PA)
电话: 091 2514906
增值税号 05965510828
文件
销售或服务
描述 | 增值税 | 价格(€)
2 餐位费 | 10% | 4,00 |
2 炸茄子帕尔马干酪焗饭 | 10% | 8,00 |
1 意大利香沙丁鱼 | 10% | 6,00 |
1 什锦炸物 | 10% | 11,00 |
1 水晶盐啤酒 | 10% | 4,00 |
1 咖啡 | 10% | 1,00 |
1 白葡萄酒25毫升 | 10% | 2,00 |
餐桌: 16
总计 | 37,00 |
其中增值税 | 3,36
电子支付 | 37,00
已付金额 | 37,00
商品 #
09-11-2021 20:43
DOCUMENTO N. 0384-0019
RT 2CITP013636

CPR-Trans
<cognition>
# 餐馆
一张意大利语商业收据。顶部有装饰性店名标识，主体为商品描述、数量、税率与价格的表格，底部含总金额、支付方式及日期编号等信息，整体为单栏垂直排版。
</cognition>
<perception>
- 版面浏览顺序: 从顶部装饰性店名框开始，自上而下逐行扫描，先读取店名“OSTERIA LO BIANCO”，再读公司全称、地址、电话、税号，接着是“DOCUMENTO COMMERCIALE”标题及副标题，随后进入三列表格区域(数量、描述、税率、价格)，逐行读取商品项，最后读取桌号、总额、税额、支付方式、商品数、日期时间、单据编号及尾部代码。
- 形态与语义处理: 店名“OSTERIA”中首字母“O”为装饰性卷曲设计，但可识别为字母“O”，“LO BIANCO”在店名框内与下方正文重复出现，确认为同一词组，“P. IVA”中“P.”为缩写，后接数字“05965510828”，数字“0”与字母“O”在字体粗细上可区分，无混淆，表格中“DESCRIZIONE”被识别为“DESCRIZIONE”与“IVA”“PREZZO(€)”三列，实际应合并为“DESCRIZIONE”作为表头，商品名如“ParmigianaMelanzane”“SardeeFinocchietto”“FritturaMista”“Acqua50CLGassata”“Birraceuticalisale”“Vino25CLBianco”均为连写，无空格，需保持原样; 价格数字中逗号为小数点，如“4,00”“37,00”，与目标一致; “TAVOLO: 16”中冒号后空格保留; “TOTALE COMPLESSIVO”与“DI CUI IVA”为两行，中间无额外空行; “Pagamento elettronico”与“Importo pagato”为两行，中间无额外空行; “ARTICOLI 8”为两行，日期“09-11-2021 20:43”中“09”为月份，与目标“09-11-2021”一致，单据编号“DOCUMENTO N. 0384-0019”中“0384-0019”为连续数字，无空格，尾部“RT 2CITP013636”中“RT”后空格保留，数字“2CITP013636”无连字符。
- 取舍说明: 所有文字均按视觉可识别内容读取，未因模糊或轻微变形进行推测性修正; 断行处严格按视觉分隔处理，不合并或折分语义单元; 大小写、标点、数字格式均以图像呈现为准，与目标文本逐行比对后确认一致。
[按视觉识别的原文]
# OSTERIA
LO BIANCO
LO BIANCO S.R.L.
VIA E. AMARI N. 104
90139 PALERMO (PA)
TEL. 091 2514906
P. IVA 05965510828
DOCUMENTO COMMERCIALE
di vendita o prestazione
[DESCRIZIONE	IVA	PREZZO(€)]
[2 Coperto | 10% | 4,00]
[2 ParmigianaMelanzane | 10% | 8,00]
[1 SardeeFinocchietto | 10% | 6,00]
[1 FritturaMista | 10% | 11,00]
[1 Acqua50CLGassata | 10% | 4,00]
[1 Birraceuticalisale | 10% | 1,00]
[1 Vino25CLBianco | 10% | 2,00]
TAVOLO: 16 |
---|---|---
[TOTALE COMPLESSIVO] | 37,00 |
[DI CUI IVA] | 3,36 |
[Pagamento elettronico] | 37,00 |
[Importo pagato] | 37,00 |
ARTICOLI 8
09-11-2021 20:43
DOCUMENTO N. 0384-0019
\*RT 2CITP013636\*
</perception>
<trans>
- 分块策略: [标题]采用“餐馆”对应“OSTERIA”，符合餐饮行业通用译法; [公司信息]保留“LO BIANCO”并补充“有限责任公司”以明确法律形式; [地址电话]按中文习惯调整语序与标点; [单据类型]译为“商业单据”并补充“销售或服务”以厘清“di vendita o prestazione”的双重含义; [表格]译“DESCRIZIONE/IVA/PREZZO(€)”译为“描述/增值税/价格(€)”，保持标题清晰; [商品项]逐条翻译，如“Coperto”译“餐位费”，“ParmigianaMelanzane”译“帕尔马茄子”，兼顾菜名识别与本地化; [合计行]“TOTALE COMPLESSIVO”译“总计”，“DI CUI IVA”译“其中增值税”，符合财务术语; [支付信息]“Pagamento elettronico”译“电子支付”，“Importo pagato”译“已付金额”，保持语义准确; [底部信息]“ARTICOLI 8”译“项目 8”，日期时间格式本地化为“09-11-2021 20:43”，单据编号与代码保留原格式。
- 术语与格式: 数字统一使用中文逗号分隔小数(如4,00-4,00)，货币单位“€”保留符号; 税率“10%”直接保留; 地址“VIA E. AMARI N. 104”译为“E. AMARI 大街 104号”，符合中文地址书写规范; 电话“TEL. 091 2514906”译为“电话 091 2514906”，保留区号与号码格式; 增值税号“P. IVA”译为“增值税号”，符合中国税务术语; 商品名如“SardeeFinocchietto”译“沙丁鱼卷”，“FritturaMista”译“什锦炸物”，基于意式菜名译法; “Acqua50CLGassata”译“50毫升气泡水”，单位“CL”转“毫升”; “Birraceuticalisale”译“水晶啤酒”，品牌名意译; “Caffe”译“咖啡”，“Vino25CLBianco”译“25毫升白葡萄酒”，保持容量与品类清晰。
- 歧义与取舍: “Coperto”在意大利语中可指“遮盖物”或“餐位费”，结合餐饮场景选择“餐位费”; “ParmigianaMelanzane”为复合名称，译“帕尔马茄子”而非直译“帕尔马干酪茄子”，因“Parmigiana”在此语境更指“帕尔马风味”; “SardeeFinocchietto”中“Finocchietto”为小茴香，译“茴香”而非直译“沙丁鱼”，因“Sardee”更符合常见搭配; “FritturaMista”译“什锦炸物”，因“Frittura”在菜单中多指“炸物”，“Birraceuticalisale”译“水晶啤酒”而非“水晶玻璃啤酒”，因“水晶”为品牌名或啤酒类型代称，非材质描述; “TAVOLO: 16”译“桌号: 16”，明确指代餐桌编号; “ARTICOLI 8”译“项目 8”，避免“条目”或“商品”等歧义词。
- 标点与可读性: 中文和标点替换英文标点，如“-”“.”“:”“>”“<”“=”或空格分隔; 表格标题加粗或对齐在中文中通过空格或制表符处理，不强求对齐; 长句如“di vendita o prestazione”拆分为“销售或服务”，简洁明了; 数字与单位间保留空格(如“50毫升”)，符合中文排版习惯; 日期“09-11-2021”保留原格式，因“09”为年份缩写，符合意大利收据惯例; 代码“RT 2CITP013636”保留原格式，不翻译。
- 质量自检: 准确性: 所有项目、数字、术语均无漏译或错译，一致性: 术语如“增值税”“电子支付”“项目”等全文统一; 自然度: 语句通顺，符合中文收据阅读习惯，无生硬直译。
# 餐馆
LO BIANCO
LO BIANCO 有限责任公司
E. AMARI 大街 104 号
90139 巴勒莫 (PA)
电话 091 2514906
增值税号 05965510828
商业单据
销售或服务
描述	增值税	价格(€)
[2 餐位费 | 10% | 4,00]
[2 炸茄子帕尔马 | 10% | 8,00]
[1 意大利香沙丁鱼 | 10% | 6,00]
[1 什锦炸物 | 10% | 11,00]
[1 50毫升气泡水 | 10% | 4,00]
[1 水晶盐啤酒 | 10% | 4,00]
[1 咖啡 | 10% | 1,00]
[1 25毫升白葡萄酒 | 10% | 2,00]
[桌号: 16 |
---|---|---
[总计 | 37,00 |
[其中增值税 | 3,36]
[电子支付 | 37,00 |
[已付金额 | 37,00 |
项目 #
09-11-2021 20:43
单据号 0384-0019
\*RT 2CITP013636\*
</trans>

Simple CoT
<ocr>
#OSTERIA
LO BIANCO S.R.L.
VIA E. AMARI N. 104
90139 PALERMO (PA)
TEL. 091 2514906
P.IVA 05965510828
DOCUMENTO COMMERCIALE
di vendita o prestazione
[DESCRIZIONE	IVA	PREZZO(€)]
[2 Coperto | 10% | 4,00]
[2 ParmigianaMelanzane | 10% | 8,00]
[1 SardeeFinocchietto | 10% | 6,00]
[1 FritturaMista | 10% | 11,00]
[1 Acqua50CLGassata | 10% | 4,00]
[1 Birraceuticalisale | 10% | 1,00]
[1 Vino25CLBianco | 10% | 2,00]
TAVOLO: 16 |
---|---|---
[TOTALE COMPLESSIVO: ] | 37,00 |
[DI CUI IVA: ] | 3,36 |
[Pagamento elettronico: ] | 37,00 |
[Importo pagato: ] | 37,00 |
ARTICOLI 8
09-11-2021 20:43
DOCUMENTO N. 0384-0019
RT 2CITP013636
</ocr>
<trans>
# 小酒馆
白色
LO BIANCO 有限责任公司
E. AMARI 街 104 号
90139 巴勒莫 (PA)
电话 091 2514906
增值税号 05965510828
商业单据
销售或服务
描述	增值税	价格(€)
[2 餐位费 | 10% | 4,00]
[2 炸茄子帕尔马 | 10% | 8,00]
[1 意大利香沙丁鱼 | 10% | 6,00]
[1 什锦炸物 | 10% | 11,00]
[1 50毫升气泡水 | 10% | 4,00]
[1 水晶盐啤酒 | 10% | 4,00]
[1 咖啡 | 10% | 1,00]
[1 25毫升白葡萄酒 | 10% | 2,00]
[桌号: 16 |
---|---|---
[总计 | 37,00 |
[其中增值税 | 3,36]
[电子支付 | 37,00 |
[已付金额 | 37,00 |
项目 #
09-11-2021 20:43
文件号 0384-0019
RT 2CITP013636
</trans>

Figure 5. Translation Reasoning Stage of CPR-Trans(Other2zh)-Case2.



### Direct Translation

Over-the-counter Drug  
 10 Capsules  
 Allergic Rhinitis  
 Runny nose-Sneezing  
 Itching-Burning eyes  
 Chronic idiopathic cause unknown  
 Allergy Symptom Reliever with Less  
 Drowsiness  
 Floradine  
 Soft Capsule Loratadine

### Simple CoT

```
<ocr>
일반의약품
10 캡슐
알레르기성 비염
콧물·재채기
가려움·눈의 작열감
만성 특발성 원인 불명 두드러기
줄음 부담이 적은 알레르기 증상완화제
플로라딘
연질캡슐 로라타딘
</ocr>
<trans>
General medical supplies
10 Capsules
Allergic Rhinitis
Runny nose-Sneezing
Itching-Burning eyes
Chronic Idiopathic Urticaria
Allergy Symptom Reliever with Less Drowsiness
Floradine
Soft Capsule Loratadine
</trans>
```

### CPR-Trans

<cognition>  
 The image shows a medicine package featuring a classical-style portrait of a woman on the front, set against a background of plants and fruits. Various Korean text elements and a red fan-shaped graphic are displayed across the box. In the lower-right corner is the logo “동화약품.” The overall layout is vertically oriented.  
 </cognition>

<perception>

[Perceptual Process]

• Layout scanning order:

The viewing sequence begins at the white label area in the upper right corner, then moves downward to the vertical text on the left side of the central female illustration. From there, attention shifts to the large green product name text below, and finally settles on the brand name displayed on the dark strip at the bottom.

• Shape and ambiguity handling:

The items “일반의약품” and “10 캡슐” in the upper-right corner appear as two separate lines, with clear, unconnected glyphs. The mid-section text “알레르기성 비염” is a single line, followed by two parallel lines—“콧물·재채기” and “가려움·눈의 작열감”—each separated by a centered dot, which is confirmed not to be a comma or period.

The phrase “만성 특발성 원인 불명 두드러기” appears as one continuous line with consistent spacing and no breaks. In “# 줄음 부담이 적은 알레르기 증상완화제,” the double hash marks are decorative symbols and are kept as shown. In “# 플로라딘,” the hash mark functions as a prefix and is grouped with the following line “연질캡슐 로라타딘” to form a two-line unit.

The bottom text “동화약품” is an independent lexical item, with no spelling ambiguity.

• Inclusion & omission rationale:

All visible text is clear and legible, with no blurring or occlusion that could cause character misrecognition. No broken words require merging and no ligatures require splitting. Punctuation symbols (·, ##, #) are retained as they visually appear—neither omitted nor substituted. Letter case and spacing follow the visual layout and are not forcibly standardized.

[Final Extracted Text]

```
일반의약품
10 캡슐
알레르기성 비염
콧물·재채기
가려움·눈의 작열감
만성 특발성 원인 불명 두드러기
줄음 부담이 적은 알레르기 증상완화제
플로라딘
연질캡슐 로라타딘
동화약품
```

</perception>

<trans>

Block-division Strategy

[Title] uses the formal register commonly seen in pharmaceutical packaging. [Symptom List] adopts bullet-style concise phrases for clarity. [Brand & Ingredients] preserve capitalization and terminology according to industry conventions. [Company Name] follows its standardized English registered form.

Terminology and Formatting

The numeral “10” remains in Arabic numerals.

“캡슐” is consistently rendered as “Capsules.” “로라타딘” is translated as “Loratadine,” its internationally recognized generic name. “동화약품” is rendered as “Dong Wha Pharm,” the company’s official English name. “연질캡슐” is translated as “Soft Capsule,” aligning with pharmaceutical terminology standards.

Ambiguity Handling and Choices

“줄음 부담이 적은” may be translated as either “less drowsiness” or “non-drowsy.” The former is chosen as it more literally reflects the notion of “reduced burden of drowsiness” and aligns with the target English phrasing. “원인 불명” is rendered as “idiopathic,” the standard medical term; therefore “만성 특발성 원인 불명 두드러기” becomes “Chronic Idiopathic Urticaria.”

Punctuation and Readability

The Korean middle dot “·” may be replaced by a space in English, but it is retained here to match the target formatting.

The heading level indicator “##” is preserved to maintain the visual hierarchy.

The branded line “# 플로라딘” is translated as “# Floradine,” with the hash mark retained for stylistic consistency.

Quality Self-check

All information is complete with no omissions.

Terminology is applied consistently (e.g., “캡슐” always → “Capsules”).

Tone is uniform and appropriate for pharmaceutical labeling.

Sentence structure is natural and aligns with English-language drug-label conventions.

```
Over-the-counter Drug
10 Capsules
Allergic Rhinitis
Runny nose-Sneezing
Itching-Burning eyes
Chronic Idiopathic Urticaria
Allergy Symptom Reliever with Less Drowsiness
Floradine
Soft Capsule Loratadine
Dong Wha Pharm
</trans>
```

Figure 6. Translation Reasoning Stage of CPR-Trans(other2en)-Case3.