MAPLE: *M*ULTILINGUAL *A*LIGNMENT OF *P*REDICTIVE *L*ATENT *E*MBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Enhancing Visual-Language Models (VLMs) with external tools improves perception but incurs substantial overhead. Common solutions rely on specialised supervision; however, these are predominantly English-centric and fail when scaling to multilingual spaces. We propose **MAPLE** (***M***ultilingual ***A***lignment of ***P***redictive ***L***atent ***E***mbeddings), a representation-learning framework based on predictive alignment that learns from expert tool-use trajectories in the latent space, eliminating the need for explicit tool invocation at inference time. Specifically, MAPLE learns multilingual predictive embeddings, bypassing reconstruction-based methods that generate latent tokens and suffer from a training-inference mismatch, as well as limited support for multi-step tool use. By preserving the standard vision-language generation pipeline, our framework is model-agnostic and straightforward to train. Crucially, it naturally supports trajectories with multiple tool calls across languages. We evaluate **MAPLE** on a comprehensive suite of perception benchmarks and introduce their corresponding multilingual versions, demonstrating that it matches or outperforms standard tuning and reconstruction-based latent-reasoning baselines across multiple languages. Finally, we expose reconstruction methods as learning static embeddings, showing that predictive alignment is a principled and globally scalable paradigm[1].

1 INTRODUCTION

Enhancing Visual-Language Models (VLMs) with external tools has emerged as a critical mechanism to improve fine-grained perception (Gao et al., 2025; Su et al., 2025; Wu et al., 2026; Wang et al., 2025a). By enabling models to iteratively manipulate visual inputs, these approaches allow VLMs to think with images, pushing them beyond standard textual reasoning constraints Davis et al. (2025); Zhang et al. (2025); Yan et al. (2026).

Indeed, explicit tool invocation introduces practical bottlenecks across different stages of the process. During inference, external calls incur latency Wang et al. (2025c) and, even with strict supervision, often result in misleading invocations that introduce computational costs and, at the same time, inject bias into the context, influencing the reasoning process Xu et al. (2026). Moreover, managing different real-world toolsets requires dynamic planning that current frameworks struggle to provide Chandhok et al. (2025). These brittle mechanisms collapse when scaled to multilingual settings, as rigid tool execution fails to generalise across languages, aggravating error propagation and limiting global applicability Wang et al. (2025b).

As an alternative to external tools, recent works simulate visual manipulations internally within continuous latent spaces Hao et al. (2025); Adhikari & Lapata (2026). Yang et al. (2025); Li et al. (2025); Wang et al. (2026) utilise autoregressive reconstruction constraints to map latent tokens to specific visual tool outputs. Although effective, they induce a distributional shift between training and inference, constraining supervised models to decode a sparse token subset at test time, systematically degrading performance Xiong et al. (2025). This formulation is restricted to single-step transformations and precludes step-wise reasoning. Finally, the reliance on monolingual resources renders these paradigms inherently English-centric. This raises the question of the extent to which tool-augmented visual reasoning can be internalised within a continuous multilingual representation space, without requiring either explicit tool execution or latent-token reconstruction at inference time.

[1] **code & data:** Link anonymous repository

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.

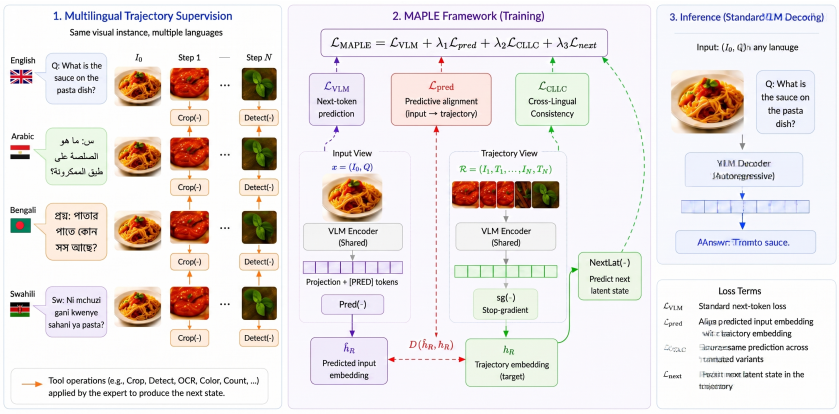


Figure 1: Overview of MAPLE. Multilingual tool-use trajectories supervise predictive latent alignment during training, enabling standard tool-free decoding at inference.

To make multimodal reasoning globally scalable and efficient, we propose **MAPLE** (***M***ultilingual ***A***lignment of ***P***redictive ***L***atent ***E***mbeddings), a framework that internalises visual tool operations directly within a language-agnostic latent space. Specifically, MAPLE aligns the predicted representation of an image-question pair with the encoded embedding of the reasoning trajectory. By implementing this predictive alignment across multiple languages, our method avoids the latent-space limitations intrinsic to autoregressive token generation. This alignment enables the model to anticipate the outcome of sequential visual manipulations without requiring explicit tool execution or test-time latent decoding. By going through the complexity of training tool interaction, MAPLE retains the standard vision-to-text inference pipeline, ensuring maximum efficiency and broad architectural scalability.

We construct MAPLE by optimising a unified set of objectives. To ensure robust multilingual proficiency, the model retains a standard autoregressive loss. At the same time, a predictive embedding loss minimises the distance between the projected input representation and a stop-gradient target derived from an expert multimodal trajectory. Completing this formulation, a next-latent prediction regulariser forces the hidden states to function as informative, language-invariant belief states, thus modelling the sequential dynamics of the visual editing process. Through this design, our framework internalises both isolated tool operations and their sequential dependencies. Furthermore, training on multilingual trajectories yields a language-agnostic latent space where structural visual transformations map directly to semantic objectives, facilitating robust multimodal reasoning across different landscapes.

To validate our framework, we evaluate MAPLE on V*, MMVP, BLINK, MARVEL, and M5, and introduce multilingual variants of the English perception benchmarks. Empirical results demonstrate that our method consistently matches or outperforms both standard fine-tuning and state-of-the-art reconstruction baselines across all tested languages.

Our contributions are as follows:

- We introduce MAPLE, a parameter-efficient framework for multilingual multimodal latent reasoning that internalises visual tool-use trajectories without requiring explicit tool invocation at inference time.
- We propose a latent-state predictive objective that replaces autoregressive token reconstruction with trajectory-level embedding alignment, enabling models to internalise step-wise visual trajectories.
- We present novel multilingual extensions of standard benchmarks, introducing new test splits in 11 diverse languages, thereby establishing a rigorous and comprehensive evaluation suite for cross-lingual visual reasoning.
- We validate the global scalability and empirical robustness of MAPLE across diverse architectures. Crucially, our analysis demonstrates that predictive alignment is a more effective and principled

paradigm than traditional reconstruction baselines, establishing a robust foundation for globally accessible multimodal perception.

2 MAPLE: MULTILINGUAL ALIGNMENT OF PREDICTIVE LATENT EMBEDDINGS

Explicit tool-use in Vision-Language Models (VLMs) induces inference overhead and struggles to generalise across different contexts. We introduce Multilingual Alignment of Predictive Latent Embeddings (MAPLE), which trains a VLM to internalise visual tool-use trajectories as predictive latent states, while preserving the standard vision-to-text decoding path at inference time. Given an image and a question, the model learns to predict the latent representation induced by an expert multimodal trajectory. We first formalise multilingual trajectory learning and its independent encoding strategy (§2.1), then introduce trajectory-level latent prediction (§2.2), define the optimisation objective and a cross-lingual consistency formulation (§2.3), and describe how to use MAPLE (§2.4).

2.1 MULTILINGUAL TRAJECTORY LEARNING

Problem formulation. We define a multilingual trajectory-learning setting over a set of languages $\mathcal{L}$. Each instance consists of an image-question pair $\mathbf{x}^{(l)} = \langle \mathcal{I}_0, \mathcal{Q}^{(l)} \rangle$, where $l \in \mathcal{L}$ denotes the language, together with a vision trajectory:

$$\mathcal{R}^{(l)} = (\mathcal{I}_1, \mathcal{T}_1^{(l)}, \mathcal{I}_2, \mathcal{T}_2^{(l)}, \ldots, \mathcal{I}_N, \mathcal{T}_N^{(l)}),$$

where each $\mathcal{I}_i$ is a tool-produced visual state, such as a crop, highlight, or spatial transformation, and each $\mathcal{T}_i^{(l)}$ is the associated reasoning text in language l. The final step $\mathcal{T}_N^{(l)}$ contains the answer to $\mathcal{Q}^{(l)}$. The visual states are shared across languages, while the rationale is language-specific. MAPLE learns from multilingual realisations of the same underlying visual trajectory. The objective is to learn from these trajectories without using tools. Unlike reconstruction-based latent reasoning methods, which generate latent tokens for visual-edit reconstruction (Li et al., 2025; Yang et al., 2025), MAPLE predicts a trajectory-level latent representation directly from the original image-question pair. This internalises tool-mediated reasoning while preserving the standard generation pipeline.

Independent trajectory encoding. To disentangle the multilingual input from the trajectory target, we encode the original query and the expert trajectory through two independent forward passes. We serialise the input view $\langle \mathcal{I}_0, \mathcal{Q}^{(l)} \rangle$ and the interleaved trajectory view $(\mathcal{I}_1, \mathcal{T}_1^{(l)}, \ldots, \mathcal{I}_N, \mathcal{T}_N^{(l)})$. For each view, we extract the hidden state of the last token to get the corresponding representation:

$$h_{\mathbf{x}}^{(l)} = \mathrm{Enc}(\mathbf{x}^{(l)}), \qquad h_{\mathcal{R}}^{(l)} = \mathrm{Enc}(\mathcal{R}^{(l)}).$$

Separate encoding prevents cross-view information leakage while keeping the method architecture-agnostic. A lightweight predictor is then trained to anticipate the trajectory embedding $h_{\mathcal{R}}$ from the input representation $h_{\mathbf{x}}$. This allows MAPLE to internalise step-wise multilingual visual reasoning while bypassing the training–inference mismatch of token-level reconstruction.

2.2 TRAJECTORY-LEVEL LATENT PREDICTION

Prediction tokens. To map the input representation $h_{\mathbf{x}}$ into the latent space, we reuse the VLM's self-attention stack as a tied-weight predictor. Following (Huang et al., 2025), in training, we append K learnable tokens (`[PRED]`) to the serialised input view and use the hidden state of the final prediction token as the predicted trajectory representation:

$$\hat{h}_{\mathcal{R}} = \mathrm{Pred}_\theta(\mathbf{x}^{(l)}, \texttt{[PRED]}_{1:K}).$$

This strategy enables additional non-linear processing of the input before latent approximation, while reducing to an identity-style mapping when $K = 0$. Since the predictor reuses the existing VLM stack, MAPLE avoids the need for separate MLPs or auxiliary transformers, thereby remaining parameter-efficient.

Trajectory alignment loss. To govern the mapping from multilingual inputs to trajectory-level latent states, MAPLE aligns the predicted trajectory representation $\hat{h}_{\mathcal{R}}^{(l)}$ with the independently encoded expert trajectory representation $h_{\mathcal{R}}$:

$$\mathcal{L}_{\text{pred}} = D\Big(\hat{h}_{\mathcal{R}}^{(l)}, \text{sg}[h_{\mathcal{R}}^{(l)}]\Big)$$

.

where $\text{sg}[\cdot]$ is the stop-gradient operation and D is a distance function. The stop-gradient ensures that the trajectory encoder provides a stable target, while the predictor is optimised to approximate this target from the original image-question pair. By avoiding token-level reconstruction, we encourage the model to encode the semantic effect of visual manipulations in a shared language-agnostic latent space.

2.3 Optimisation Objective

Cross-lingual latent consistency. To explicitly enforce a language-agnostic reasoning space, we introduce a Cross-Lingual Latent Consistency (*CLLC*) objective. Given the same image $\mathcal{I}_0$ and query $\mathcal{Q}$ formulated in two distinct languages, $l_a, l_b \in \mathcal{L}$, we minimise the distance between their respective predicted trajectory embeddings:

$$\mathcal{L}_{\text{CLLC}} = D\Big(\hat{h}_{\mathcal{R}}^{(l_a)}, \hat{h}_{\mathcal{R}}^{(l_b)}\Big)$$

This constraint discourages language-specific shortcuts and encourages semantically equivalent queries to converge towards similar visual reasoning states.

Autoregressive supervision. To support reasoning capabilities across $\mathcal{L}$, we retain the standard VLM objective over the textual components of the expert trajectory. Given the interleaved multimodal context, the model is trained to predict each token in the segments $\mathcal{T}_1^{(l)}, \ldots, \mathcal{T}_N^{(l)}$:

$$\mathcal{L}_{\text{VLM}} = -\sum_{n=1}^{N} \sum_{t=1}^{|\mathcal{T}_n^{(l)}|} \log p_\theta\big(\mathcal{T}_{n,t}^{(l)} \mid \mathbf{C}_{n,t}^{(l)}\big)$$

where $\mathbf{C}_{n,t}^{(l)}$ includes the input $\langle \mathcal{I}_0, \mathcal{Q}^{(l)} \rangle$, all preceding image-text pairs, and the prefix of the current reasoning step.

Latent-dynamics regularisation. To empower the summary capacity of multilingual latent representations, we incorporate a next-latent prediction objective inspired by transformer latent dynamics. Given a hidden state h_t within the serialised trajectory, we optimise a predictor to predict future states $\hat{h}_{t+i}$ over a horizon d:

$$\mathcal{L}_{\text{next}} = \mathbb{E}_t\Big[\tfrac{1}{d}\textstyle\sum_{i=1}^{d} \text{L1}_{\text{sm}}(\text{sg}[h_{t+i}], \hat{h}_{t+i})\Big]$$

This formulation encourages hidden states to encode the trajectory's temporal evolution, transforming them into belief states that summarise the full context history (Teoh et al., 2025).

Overall objective. The final objective combines generative, predictive, and consistency supervision. MAPLE is trained by optimising these losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VLM}} + \lambda\big[\mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{CLLC}} + \mathcal{L}_{\text{next}}\big]$$

where λ controls the contribution of the latent learning signals relative to the autoregressive generation loss. Together, they connect discrete multilingual text generation with continuous, language-invariant latent reasoning.

2.4 INFERENCE

MAPLE uses the standard decoding pipeline of the underlying VLM. It receives only the original multilingual image-question pair $\langle \mathcal{I}_0, \mathcal{Q}^{(l)} \rangle$ and directly generates the answer. The framework does not invoke external APIs, generate visual transformations, or decode latent reasoning tokens. Moreover, the `[PRED]` tokens used during training are discarded, keeping inference as lightweight as vanilla decoding.

3 EXPERIMENTAL SETUP

To evaluate the extent to which MAPLE internalises expert tool-use trajectories, we conduct experiments across different levels of visual complexity and multilingual perception. We describe the training regimes and trajectory datasets used for latent alignment (§3.1). We then present the evaluation benchmarks, including both English perception-intensive suites and our multicultural extensions (§3.2). Finally, we specify the models, baselines, and implementation (§3.3).

3.1 TRAINING REGIMES AND MULTILINGUAL DATA

We evaluate MAPLE under three training regimes to test different forms of latent tool-use internalisation: *(i)* single-tool trajectories with one visual operation; *(ii)* multi-tool trajectories with one visual operation per instance; and *(iii)* single-tool trajectories with multiple sequential operations.

For the first setting, we use the data from Li et al. (2025), which provides regions of interest (RoIs) for crop-based visual operations. For the second setting, we use ThinkMorph (Gu et al., 2025), which contains four categories of visual transformations: bounding boxes, chart highlights, jigsaw reconstructions, and spatial maze navigation. Finally, we use PixelReasoner (Wang et al., 2025a), which includes trajectories with up to three sequential visual crops.

To construct multilingual training trajectories, we translate the associated text $\mathcal{T}$ into a set of target languages $\mathcal{L}$. The visual inputs, tool-use structures, and target operations are kept fixed, while the linguistic realisation of each trajectory is different. This design enables the model to map multilingual instructions and reasoning traces onto the same trajectory, rather than learning language-specific tool-use shortcuts. Dataset statistics and examples are provided in Appendix H and T.

3.2 MULTILINGUAL PERCEPTION BENCHMARKS

We evaluate MAPLE on perception-intensive Visual Question Answering (VQA) benchmarks. Following Li et al. (2025), our base evaluation includes V* (Wu & Xie, 2023), which targets guided visual search through object–attribute queries (V^*_{DA}) and relative-position queries (V^*_{RP}); BLINK (Fu et al., 2025); and MMVP (Tong et al., 2024), which probes robustness to CLIP-blind visual patterns.

To assess cross-lingual robustness, we introduce multilingual versions of these benchmarks across ten target languages: Italian, Spanish, French, German, Russian, Chinese, Arabic, Bengali, Swahili, and Telugu. Both training trajectories and evaluation prompts are translated only at the textual level, preserving the original images, visual states; details are in Appendix F. We measure the performance drop relative to English, Δ_{en}, and introduce an **Exact Match Consistency (EMC)** score, which quantifies the proportion of instances where the model predicts the exact same answer across all translated variants of a given query, serving as a direct metric for language-agnostic latent stability. Further details on translation validation, output parsing, and metric computation are reported in Appendices E and G.

Culturally Grounded Evaluation. Evaluating on translated benchmarks risks inheriting Western-centric visual biases. Therefore, we incorporate **MARVEL** (Liu et al., 2021), a multicultural benchmark naturally designed for cross-lingual visual reasoning over diverse cultural concepts. This allows us to assess the true global scalability of our framework. We evaluate on M5 (Schneider & Sitaram, 2024) as an external reference point for general multilingual vision-language competence.

Regime	Method	En ↑	HR ↑	LR ↑	Gap ↓	m-Avg ↑	Δ_{en} ↓	EMC ↑	m-V*	m-MMVP	m-BLINK	MARVEL
Zero-shot, no fine-tuning												
–	Qwen2.5-VL-7B-Instruct	73.6	65.7	60.2	5.5	63.5	10.1	41.3	71.2	60.5	64.8	65.1
Single-type, single tool call (m-LVR trajectories)												
m-LVR	CoVT	72.0	63.4	58.9	4.5	61.6	10.4	35.8	70.4	52.3	60.1	65.6
	LVR (4 steps)	78.7	69.5	64.1	5.4	67.3	11.4	48.6	74.8	65.2	68.4	68.8
	SFT	77.6	67.6	62.2	5.4	65.4	12.2	45.2	73.1	62.4	66.7	67.8
	MAPLE	**78.9**	**76.8**	**74.6**	**2.2**	**75.9**	**3.0**	**74.5**	**79.5**	**71.8**	**76.2**	**79.7**
Multiple-type, single tool call (m-ThinkMorph trajectories)												
m-ThinkMorph	SFT	63.8	48.8	43.7	5.1	46.7	17.1	22.4	51.5	40.1	49.8	51.4
	MAPLE	**75.9**	**72.3**	**69.6**	**2.7**	**71.2**	**4.7**	**68.2**	**72.1**	**74.5**	**71.0**	**72.4**
Single-type, multiple tool calls (m-PixelReasoner trajectories)												
m-PixelReasoner	PixelReasoner	**81.2**	70.5	66.0	4.5	68.7	12.5	42.1	74.5	63.8	**74.1**	69.2
	MAPLE	78.0	**75.5**	**73.5**	**2.0**	**74.7**	**3.3**	**71.8**	**76.8**	**69.4**	73.8	**80.8**

Table 1: Main multilingual visual reasoning results (Qwen2.5-VL-7B-Instruct). HR/LR denote high- and lower-resource averages. Gap is the HR–LR difference, m-Avg excludes English, and Δ_{en} is the drop from English. Detailed per-language results, statistics, and baseline configuration are in Appendices U, H, and S.

3.3 Models and Baselines

Our primary experiments use Qwen2.5-VL-7B-Instruct (Wang et al., 2024) as the backbone, enabling direct head-to-head comparison with existing reconstruction-based baselines. However, to show that MAPLE is actually architecture-agnostic and does not overfit to a single model family's representation space, we extend our evaluation to include fundamentally different VLM architectures. Specifically, we incorporate **PaliGemma-3B** (Beyer et al., 2024)—which integrates a SigLIP vision encoder with a Gemma language backbone—and **Llama-3.2-11B-Vision-Instruct**. We additionally report results for smaller Qwen variants (Qwen2.5-VL-3B-Instruct and Qwen3-VL-4B (Bai et al., 2025)) to assess the framework's scalability across varying parameter counts.

For single-step settings, we compare against LVR (Li et al., 2025), using its 4-latent-token configuration, and CoVT (Qin et al., 2025). For multi-tool settings, we compare MAPLE with a LoRA fine-tuned variant of the base VLMs trained on the same data. For sequential tool-use, we benchmark against PixelReasoner (Wang et al., 2025a), which relies on explicit tool invocation.

MAPLE and all comparison systems are trained and evaluated under the same protocol. We include the original instruction-tuned backbone in a zero-shot setting and an SFT baseline trained on the same multilingual trajectories, but without MAPLE's latent alignment and consistency objectives. Implementation details, models, dataset construction, validation procedures, evaluation metrics are reported in Appendices A, B, D, E, dataset statistics are reported in Appendices G, and H.

4 Results

MAPLE improves cross-lingual visual reasoning performance across diverse operational regimes, demonstrating that predictive latent alignment establishes a robust, language-agnostic representation space. We first evaluate the extent to which MAPLE improves over standard fine-tuning and reconstruction-based latent reasoning, focusing on the gap between high-resource and lower-resource languages (§4.1). We then test whether latent tool-use knowledge transfers to languages unseen during trajectory training (§4.2), and isolate the contribution of each MAPLE objective (§4.4). Finally, we assess whether the approach generalises across VLM architectures (§4.3), analyse the learned latent space (§4.5), and the limits of autoregressive reconstruction in multilingual settings (§4.6).

4.1 Multilingual Robustness

Table 1 reports the multilingual comparison, combining benchmark-level averages with a resource-level breakdown. We report English, high- (HR) and low-resource (LR) average accuracy, the HR-LR gap, multilingual (m-Avg), Δ_{en}, Exact Match Consistency (EMC), and task-specific averages.

Across all regimes, MAPLE improves both accuracy and cross-lingual stability, with robust gains across different seeds (Appendix N). In the m-LVR setting, LR accuracy increases from 64.1% with LVR to 74.6%, while the HR–LR gap decreases ($5.4 \rightarrow 2.2$ points). The same pattern appears in the

m-ThinkMorph regime: SFT falls to 43.7% on LR languages, and MAPLE reaches 69.6%. These results show that predictive alignment reduces cross-lingual degradation of tool-mediated visual reasoning.

Isolated Visual Perception. On m-LVR, MAPLE reaches 75.9% m-Avg accuracy and 74.5% EMC, outperforming both SFT (65.4% m-Avg, 45.2% EMC) and LVR (67.3% m-Avg, 48.6% EMC). The gain is clear on LR languages, where MAPLE improves accuracy from 64.1% under LVR to 74.6%. This suggests that $\mathcal{L}_{\mathrm{CLLC}}$ helps map translated queries onto a shared trajectory representation, improving consistency beyond reconstruction-based latent decoding.

Heterogeneous Tool Supervision. Under m-ThinkMorph's diverse operations, SFT drops to 46.7% m-Avg accuracy and 22.4% EMC. MAPLE reaches 71.2% m-Avg accuracy and 68.2% EMC, with LR accuracy increasing from 43.7% to 69.6%. This indicates that a trajectory-level target is more effective than standard next-token fine-tuning when the supervision mixes different visual operations.

Sequential Trajectory Internalisation. On m-PixelReasoner, MAPLE improves over PixelReasoner in m-Avg (74.7% vs. 68.7%) and EMC (71.8% vs. 42.1%). PixelReasoner remains stronger on English accuracy and slightly stronger on m-BLINK, but relies on explicit tool invocation at inference time. MAPLE internalises sequential tool-use supervision while retaining standard VLM decoding and zero tool latency (Appendix K).

Correct Consistency and Cultural Grounding. Appendix L shows that higher EMC mainly reflects stable correct predictions, not repeated errors across languages. The gains also extend to MARVEL and M5, two naturally multicultural or multilingual benchmarks, where MAPLE outperforms the corresponding baselines (Appendix I).

Training Setting	En ↑	HR Avg. ↑	LR Avg. ↑	Non-En Avg. ↑	EMC ↑
Zero-Shot	73.6	65.7	60.2	63.5	41.3
SFT En-only	78.2	63.6	57.8	61.3	33.8
MAPLE En-only	78.4	72.1	68.6	70.7	62.7
SFT multilingual	77.6	67.6	62.2	65.4	45.2
MAPLE multil	**78.9**	**76.8**	**74.6**	**75.9**	**74.5**

Table 2: Zero-shot cross-lingual transfer on m-LVR.

4.2 Cross-Lingual Transfer of Latent Tool-Use

A strict test of language-agnostic latent reasoning is whether tool-use knowledge transfers to languages unseen during trajectory training. Table 2 assesses whether tool-use supervision learned from English trajectories transfers to other languages. English-only SFT increases English accuracy to 78.2%, but reduces the non-English to 61.3% and EMC to 33.8%, suggesting that tuning on English traces can specialise the model toward English patterns.

MAPLE En-only achieves a similar English accuracy (78.4%) while improving non-English accuracy to 70.7% and EMC to 62.7%. Full multilingual MAPLE remains strongest, with 75.9% non-English accuracy and 74.5% EMC. Thus, predictive alignment supports transfer from English trajectories, while multilingual trajectory supervision further consolidates the shared latent space. We report additional training-language ablations and Correct Consistency test in Appendices J and L.

4.3 Cross-Architectural Generalisation

To verify that our findings are not isolated to the Qwen representation space, Table 3 evaluates MAPLE across PaliGemma-3B, Qwen3-VL-4B, and Llama-3.2-11B-Vision. MAPLE improves over zero-shot inference and SFT on all tested backbones, with consistent gains in both m-Avg accuracy and EMC. On PaliGemma-3B, EMC increases from 31.5% under SFT to 59.3%, and m-Avg accuracy rises ($54.1\%\rightarrow64.8\%$). For stronger backbones, the gains remain substantial: Qwen3-VL-4B reaches 78.5% m-Avg and 76.4% EMC, while Llama-3.2-11B-Vision obtains the highest absolute performance, with 83.4% m-Avg and 81.5% EMC.

Backbone	Method	En Acc. ↑	m-Avg Acc. ↑	Δ_{en} ↓	EMC ↑	MARVEL ↑
PaliGemma-3B	Zero-Shot	56.8	52.4	4.4	28.1	50.3
	SFT (m-LVR)	58.7	54.1	4.6	31.5	51.8
	MAPLE (m-LVR)	**66.1**	**64.8**	**1.3**	**59.3**	**63.2**
Qwen3-VL-4B	Zero-Shot	74.6	68.7	5.9	45.2	70.1
	SFT (m-LVR)	75.8	69.2	6.6	47.1	71.3
	MAPLE (m-LVR)	**80.4**	**78.5**	**1.9**	**76.4**	**81.2**
Llama-3.2-11B-Vision	Zero-Shot	78.9	72.1	6.8	49.8	73.5
	SFT (m-LVR)	80.6	74.3	6.3	51.2	75.6
	MAPLE (m-LVR)	**84.7**	**83.4**	**1.3**	**81.5**	**86.7**

Table 3: Cross-architectural generalisation of MAPLE. Lower Δ_{en} indicates stronger cross-lingual stability.

The results suggest that MAPLE is architecture-agnostic and yields the largest improvement on the smaller PaliGemma and the strongest results on the larger Llama backbone, indicating that predictive alignment remains effective across different VLM.

4.4 Component Ablation

Table 4 isolates the contribution of each MAPLE objective. Removing $\mathcal{L}_{\text{pred}}$ produces the largest drop, reducing HR and LR accuracy to SFT levels and lowering EMC from 74.5% to 45.2%. This shows that the main gains come from aligning the input representation with the trajectory-level target.

Variant	En ↑	HR Avg. ↑	LR Avg. ↑	Δ_{en} ↓	EMC ↑
MAPLE	**78.9**	**76.8**	**74.6**	**3.0**	**74.5**
w/o $\mathcal{L}_{\text{pred}}$	77.6	67.6	62.2	12.2	45.2
w/o $\mathcal{L}_{\text{CLLC}}$	79.0	73.2	69.1	7.4	58.9
w/o $\mathcal{L}_{\text{next}}$	78.1	74.7	71.8	4.6	68.1
w/o `[PRED]` tokens	78.3	74.1	70.6	5.6	64.8

Table 4: MAPLE component ablations on m-LVR.

Removing $\mathcal{L}_{\text{CLLC}}$ keeps English accuracy high, but reduces LR accuracy to 69.1% and EMC to 58.9%. This confirms that the cross-lingual constraint is needed to stabilise predictions across translated inputs. Ablating $\mathcal{L}_{\text{next}}$ or the `[PRED]` tokens also reduces performance, indicating that both components refine the trajectory representation. Sensitivity to K and λ is analysed in Appendix M.

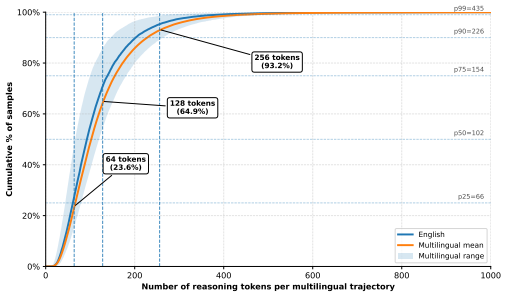


Figure 2: CDF of reasoning tokens per multilingual trajectory. The length variation across multilingual traces does not match the fixed 4–16 token budget used by autoregressive latent reconstruction.

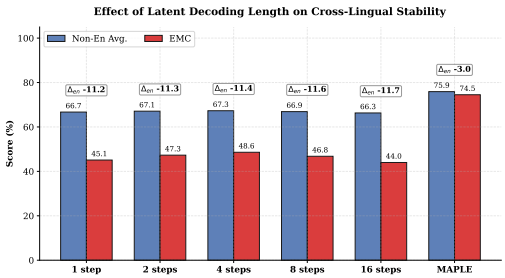


Figure 3: Effect of latent decoding length on cross-lingual stability.

4.5 LANGUAGE-AGNOSTIC LATENT SPACE

To verify the geometric organisation of the representation space, Figure 4 visualises input-view and trajectory-view embeddings for translated variants of the same visual question, comparing multilingual SFT against MAPLE.

Under standard SFT, representations remain clustered by language, indicating fragmented cross-lingual alignment. MAPLE collapses these linguistic boundaries, tightly binding input-view embeddings to their corresponding trajectory-view targets. The resulting clusters organise strictly by shared visual reasoning structure, fully abstracting away surface language variations.

This geometric restructuring explains the observed behavioural stability. MAPLE achieves high EMC by anchoring diverse multilingual inputs into a unified, language-agnostic latent trajectory space. We provide quantitative diagnostics supporting this shift in Appendix Q.

4.6 THE VULNERABILITY OF AUTOREGRESSIVE RECONSTRUCTION IN MULTILINGUAL SETTINGS

Reconstruction-based latent reasoning (LVR, CoVT) assumes autoregressive latent token generation simulates intermediate image edits. Yet models often entangle this spatial reconstruction with the dominant language of the training set, leading to instability in distant languages.

This instability is amplified by the length variation of reasoning traces (Figure 2). These traces often span hundreds of tokens reconstruction methods compress the reasoning process into a fixed 4–16 token latent budget. Figure 3 shows that increasing the number of LVR decoding steps does not resolve the issue; indeed, EMC remains below 50%, suggesting that the bottleneck is connected to the autoregressive latent-decoding interface.

MAPLE avoids this bottleneck by replacing token-level reconstruction with trajectory-level predictive alignment, projecting the input into a shared trajectory embedding and applies cross-lingual consistency directly in that space, while preserving the VLM inference path. Latent-step analysis details are in Appendix P. Appendix O provides trajectory-length statistics across languages, and Appendix K summarises inference-time characteristics. Qualitative trajectory and examples are in Appendix T.

5 CONCLUSION

We introduced MAPLE, which replaces autoregressive latent-token reconstruction with predictive embedding alignment to internalise visual tool-use trajectories in VLMs. Across languages and architectures, MAPLE improves accuracy and cross-lingual consistency without inference-time tools, showing that predictive alignment is a stable alternative for multilingual multimodal reasoning.

LIMITATIONS

While MAPLE improves cross-lingual visual reasoning, it inherits certain limitations from its underlying architecture and operational paradigm. First, because MAPLE projects the entire reasoning trajectory into a single forward-pass embedding, it cannot dynamically halt or correct its reasoning path in response to intermediate visual feedback. In deep sequential tasks (e.g., navigating complex mazes or evaluating highly ambiguous spatial overlaps), an early misalignment in the latent space propagates through to the final answer without the iterative self-correction typical of active external tools. Second, by internalising visual operations, MAPLE abstracts away from explicit symbolic processors. Consequently, the framework struggles with tasks that require explicit Optical Character Recognition (OCR), especially when the text within the image is written in low-resource or non-Latin scripts (e.g., Bengali, Telugu). In these cases, MAPLE is strictly bottlenecked by the base VLM's native visual-textual tokenisation limits. Finally, while translation-based trajectory alignment forces a globally shared semantic space, it does not fully resolve culturally grounded visual biases. Objects or cultural practices underrepresented in the backbone's pretraining data may still suffer from performance degradation, suggesting that while the *reasoning space* is language-agnostic, the *visual priors* remain tied to the base model's original training distribution. A qualitative breakdown of the remaining failure modes is provided in Appendix R.

Ethics Statement

All experiments and data collection in this work were conducted in strict accordance with standard ethical guidelines for academic research. During the preparation of the paper, we use LLMs assistant for text editing, grammatical correction, and structural refinement. We thoroughly reviewed and verified all AI-assisted modifications and take full responsibility for the scientific content, quality, and accuracy of the final paper.

AI Use Statement

We used LLM-based assistants to polish the text of the paper. All AI-assisted content was produced under human supervision: the authors reviewed and verified every modification and take full responsibility for the content of the paper.

References

Ashutosh Adhikari and Mirella Lapata. Multimodal latent reasoning via predictive embeddings, 2026. URL https://arxiv.org/abs/2604.08065.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

Shivam Chandhok, Wan-Cyuan Fan, Vered Shwartz, Vineeth N. Balasubramanian, and Leonid Sigal. Response wide shut? surprising observations in basic vision language model capabilities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25530–25545, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1241. URL https://aclanthology.org/2025.acl-long.1241/.

Anthony C Davis, Burhan Sadiq, Tianmin Shu, and Chien-Ming Huang. Augmented vision-language models: A systematic review, 2025. URL https://arxiv.org/abs/2507.22933.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 148–166, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73337-6.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage, 2025. URL https://arxiv.org/abs/2412.15606.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2510.27492.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2025. URL https://arxiv.org/abs/2412.06769.

Hai Huang, Yann LeCun, and Randall Balestriero. Llm-jepa: Large language models meet joint embedding predictive architectures, 2025. URL https://arxiv.org/abs/2509.14252.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning, 2025. URL https://arxiv.org/abs/2509.24251.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.818. URL https://aclanthology.org/2021.emnlp-main.818/.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens, 2025. URL https://arxiv.org/abs/2511.19418.

Florian Schneider and Sunayana Sitaram. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks, 2024. URL https://arxiv.org/abs/2407.03791.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. Openthinkimg: Learning to think with images via visual tool reinforcement learning, 2025. URL https://arxiv.org/abs/2505.08617.

Jayden Teoh, Manan Tomar, Kwangjun Ahn, Edward S. Hu, Pratyusha Sharma, Riashat Islam, Alex Lamb, and John Langford. Next-latent prediction transformers learn compact world models, 2025. URL https://arxiv.org/abs/2511.05963.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, June 2024.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning, 2025a. URL https://arxiv.org/abs/2505.15966.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Peng Wang, Ruihan Tao, Qiguang Chen, Mengkang Hu, and Libo Qin. X-WebAgentBench: A multilingual interactive web benchmark for evaluating global agentic system. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19320–19335, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.988. URL https://aclanthology.org/2025.findings-acl.988/.

Xiasi Wang, Tianliang Yao, Simin Chen, Runqi Wang, Lei Ye, Kuofeng Gao, Yi Huang, and Yuan Yao. VLMInferSlow: Evaluating the efficiency robustness of large vision-language models as a service. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16035–16050, Vienna, Austria, July 2025c. Association

for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.781. URL https://aclanthology.org/2025.acl-long.781/.

Zhangyun Wang, Ni Ding, and Aniket Mahanti. Learning robust vision-language models from natural latent spaces. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=7G9YKty2UZ.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use, 2026. URL https://arxiv.org/abs/2505.19255.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, Chaofan Tao, Shen Yan, Huaxiu Yao, Lingpeng Kong, Hongxia Yang, Mi Zhang, Guillermo Sapiro, Jiebo Luo, Ping Luo, and Ngai Wong. Autoregressive models in vision: A survey, 2025. URL https://arxiv.org/abs/2411.05902.

Yige Xu, Yongjie Wang, Zizhuo Wu, Kaisong Song, Jun Lin, and Zhiqi Shen. Do vision-language models truly perform vision reasoning? a rigorous study of the modality gap, 2026. URL https://arxiv.org/abs/2604.16256.

Muyang Yan, Miras Mengdibayev, Ardon Floros, Weihang Guo, Lydia E. Kavraki, and Zachary Kingston. Using vlm reasoning to constrain task and motion planning, 2026. URL https://arxiv.org/abs/2510.25548.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens, 2025. URL https://arxiv.org/abs/2506.17218.

Congzhi Zhang, Jiawei Peng, Zhenglin Wang, Yilong Lai, Haowen Sun, Heng Chang, Fei Ma, and Weijiang Yu. Vrest: Enhancing reasoning in large vision-language models through tree search and self-reward mechanism, 2025. URL https://arxiv.org/abs/2506.08691.

A Implementation and Hyper-parameters

MAPLE is trained with LoRA adapters using rank r=64 and scaling factor α=128. We use the same multiple-choice parsing protocol for the base VLM, SFT, reconstruction baselines, and MAPLE. Prediction tokens are used only during training and are discarded at inference time. All experiments are conducted using four NVIDIA A100 GPUs.

Backbone	**LoRA** r/α	**LR**	K	λ
Qwen2.5-VL-7B	64 / 128	1.0×10^{-5}	4	0.20
PaliGemma-3B	64 / 128	1.5×10^{-5}	4	0.20
Qwen3-VL-4B	64 / 128	1.0×10^{-5}	4	0.20
Llama-3.2-11B-Vision	32 / 64	8.0×10^{-6}	4	0.15

Table 5: Backbone-specific hyper-parameters. K is the number of prediction tokens; λ weights the latent objectives.

Parameter	**Value**	**Parameter**	**Value**	**Notes**
Optimiser	AdamW	Weight decay	0.01	All regimes
Warmup ratio	0.03	Epochs	2–3	Regime-dependent
Batch size	128	Grad. accum.	8	Effective batch
Distance D	SmoothL1	Horizon d	4	$\mathcal{L}_{\text{next}}$
Max input len	2048	Max output len	512	Multilingual traces

Table 6: General optimisation settings shared across experiments unless otherwise specified.

B Model Versions

Table 7 reports the exact model checkpoints used in our experiments.

Model	**Checkpoint / Version**
Qwen2.5-VL-7B	`Qwen/Qwen2.5-VL-7B-Instruct`
Qwen2.5-VL-3B	`Qwen/Qwen2.5-VL-3B-Instruct`
Qwen3-VL-4B	`Qwen/Qwen3-VL-4B-Instruct`
PaliGemma-3B	`google/paligemma-3b-mix-448`
Llama-3.2-11B-Vision	`meta-llama/Llama-3.2-11B-Vision-Instruct`

Table 7: Model checkpoints used in the experiments. Identifiers correspond to Hugging Face model names. Access verified on 05/2026.

C Dataset and Benchmark Sources

Table 8 reports the training-trajectory sources and evaluation benchmarks used in this work.

Resource	Source	Use
m-LVR	LVR (Li et al., 2025)	Single-type, single-tool trajectory supervision with RoI-based visual operations.
m-ThinkMorph	ThinkMorph (Gu et al., 2025)	Multiple-type, single-tool supervision: bounding boxes, chart highlights, jigsaw reconstruction, spatial maze navigation.
m-PixelReasoner	PixelReasoner (Wang et al., 2025a)	Sequential trajectory supervision with multiple visual crop operations.
m-V*	V* (Wu & Xie, 2023)	Multilingual object–attribute and relative-position visual search.
m-MMVP	MMVP (Tong et al., 2024)	Multilingual evaluation of CLIP-blind visual pattern robustness.
m-BLINK	BLINK (Fu et al., 2025)	Multilingual fine-grained visual perception and comparison.
MARVEL	MARVEL (**?**)	Naturally multicultural visual reasoning evaluation.
M5	M5 (Schneider & Sitaram, 2024)	General multilingual vision-language evaluation.

Table 8: Training-trajectory sources and evaluation benchmarks. The prefix "m-" denotes our multilingual extension of an English source.

D Multilingual Data Construction

We construct multilingual trajectory data by translating the textual components of each expert trajectory while preserving the visual states and tool-use structure. The initial image $\mathcal{I}_0$, the intermediate visual states $\mathcal{I}_1, \ldots, \mathcal{I}_N$, the tool order, and the final visual target are kept fixed across languages. Only the question $\mathcal{Q}$ and the reasoning texts $\mathcal{T}_1, \ldots, \mathcal{T}_N$ are translated. This ensures that all variants correspond to the same visual reasoning path, so the model is not trained to learn language-specific tool-use policies, but to align different linguistic realisations with a shared trajectory-level representation, a property central to MAPLE's cross-lingual consistency objective. For multiple-choice benchmarks, questions, and answer options are translated, preserving the correct answer index. Instances are discarded when translation changes the intended visual attribute, alters the spatial relation, and collapses two options into near-synonyms, or introduces ambiguity absent from the English source.

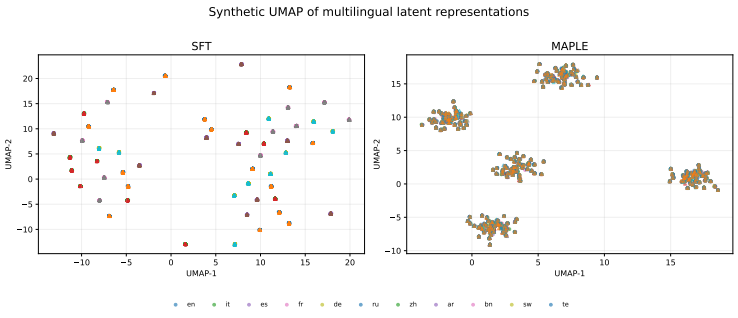


Figure 4: Projection of input and trajectory embeddings across languages in SFT and MAPLE.

E TRANSLATION VALIDATION

To reduce the risk that translation artefacts drive differences in multilingual performance, we apply a lightweight validation pipeline. We verify that each translated question preserves the same visual target as the English source; for multiple-choice items, we check that the correct answer remains semantically equivalent and that options stay mutually distinct. We back-translate a random subset into English and flag cases in which the queried object, attribute, spatial relation, or answer semantics change.

Table 9 reports the statistics. Lower-resource languages show a higher proportion of flagged cases, mainly due to answer-option collapse, spatial-relation ambiguity, and culturally specific lexical choices. Most flagged instances are manually revised; a small proportion is discarded.

Language Group	Checked	Flagged	Revised	Removed
High-resource	1,250	47 (3.8%)	35 (2.8%)	12 (1.0%)
Lower-resource	1,000	71 (7.1%)	53 (5.3%)	18 (1.8%)
All languages	2,250	118 (5.2%)	88 (3.9%)	30 (1.3%)

Table 9: Translation validation statistics.

F LANGUAGE GROUPING AND TRANSLATION MODEL

For the resource-level analysis, we group languages according to their relative availability in common pretraining sources and multilingual web corpora. English is treated as the reference language and is not included in the HR/LR averages. We define Italian, Spanish, French, German, Russian, and Chinese as high-resource (HR) languages, and Arabic, Bengali, Swahili, and Telugu as lower-resource (LR) languages. This grouping is not intended as an absolute typological classification, but as an operational distinction for analysing cross-lingual degradation in multilingual VLM evaluation.

All multilingual versions of the training trajectories and evaluation benchmarks were produced using `GPT-5.4 mini` through the OpenAI API. We translated only the textual components of each instance, while preserving the original images, visual states, tool-use structure, answer options, and correct answer indices. The translated data were then validated following the procedure described in Appendix E.

G EVALUATION PROTOCOL

Models are prompted to return a single choice; we constrain the maximum generation length to 16 tokens to eliminate variance in open-ended verbal outputs across languages, and we normalise by extracting the option marker. If the model outputs answer text, we map it to the corresponding choice when the match is exact and unambiguous. Outputs with no valid option, multiple incompatible options, or explanations without a recoverable answer are counted as incorrect. We report English accuracy, HR and LR averages, the HR–LR gap, the non-English average (m-Avg), and the English performance drop: $\Delta_{\text{en}} = \text{Acc}_{\text{en}} - \text{mAvg}$.

Exact Match Consistency (EMC). Let $\mathcal{Q}_i^{(l)}$ denote a visual query translated into language $l \in \mathcal{L}$, and let $f(\cdot)$ be the model prediction function. For a dataset of size M:

$$\text{EMC} = \frac{1}{M} \sum_{i=1}^{M} \prod_{l \in \mathcal{L}} \mathbb{I}\Big[f\Big(\mathcal{Q}_i^{(l)}\Big) = f\Big(\mathcal{Q}_i^{(\text{en})}\Big) \Big].$$

A score of 1 is assigned if and only if the model selects the same option across *all* translated variants of a visual instance. EMC, therefore, measures cross-lingual decision stability beyond correctness.

H Dataset and Benchmark Statistics

Table 10 reports the number of base expert trajectories and their multilingual realisations. Visual states are shared across languages; reasoning text is translated into the ten target languages.

Regime	Base Traj.	Multilingual Traj.	Avg. Visual States	Avg. Reasoning Tokens
m-LVR	18,920	208,120	1.0	86.4
m-ThinkMorph	12,480	137,280	1.0	112.7
m-PixelReasoner	9,750	107,250	2.6	138.5

Table 10: Training-trajectory statistics.

I Naturally Multicultural and Multilingual Benchmarks

Table 11 reports results on MARVEL and M5, which complement the translated evaluation by testing multicultural visual-language competence.

Method	MARVEL ↑	M5 ↑	Avg. ↑
Zero-Shot	65.1	61.8	63.5
SFT (m-LVR)	67.8	62.9	65.4
LVR	68.8	64.2	66.5
MAPLE (m-LVR)	**79.7**	**75.8**	**77.8**
SFT (m-ThinkMorph)	51.4	47.3	49.4
MAPLE (m-ThinkMorph)	**72.4**	**70.1**	**71.3**
PixelReasoner	69.2	65.9	67.6
MAPLE (m-PixelReasoner)	**80.8**	**74.9**	**77.9**

Table 11: Evaluation on naturally multicultural and multilingual benchmarks.

J Training-Language Ablation

Table 12 isolates the effect of multilingual trajectory data from MAPLE's predictive alignment objective. MAPLE outperforms SFT under both English-only and multilingual regimes; the combination of predictive alignment and multilingual data yield strong EMC and a small HR–LR gap.

Method & Data	HR ↑	LR ↑	Gap ↓	EMC ↑
SFT (En only)	63.6	57.8	5.8	33.8
MAPLE (En only)	72.1	68.6	3.5	62.7
SFT (Multilingual)	67.6	62.2	5.4	45.2
MAPLE (Multilingual)	**76.8**	**74.6**	**2.2**	**74.5**

Table 12: Effect of multilingual trajectory supervision on m-LVR.

K Inference Efficiency

Table 13 compares inference-time characteristics. MAPLE preserves the standard VLM decoding path without external tools, latent-token decoding, and no extra test-time tokens.

Method	Tools	Latent Dec.	Extra Tok.	Rel. Lat.
Zero-Shot / SFT	No	No	0	1.00×
LVR	No	Yes	4–16	1.12×
CoVT	No	Yes	fixed chain	1.18×
PixelReasoner	Yes	No	tool-dependent	1.85×
MAPLE	**No**	**No**	**0**	**1.00×**

Table 13: Inference-time comparison. MAPLE matches vanilla VLM decoding latency.

L Correct Consistency

EMC measures whether the model produces the same answer across translated variants regardless of correctness. To distinguish stable correctness from stable errors, we additionally compute *Correct Consistency* (CC), which requires all language variants of an instance to receive the same *correct* answer. MAPLE improves both EMC and CC, showing that consistency gains do not simply reflect repeating incorrect predictions across languages. The reduction in *Stable Wrong* predictions further suggests that predictive alignment improves cross-lingual grounding.

Method	EMC ↑	CC ↑	Stable Wrong ↓
Zero-Shot	41.3	33.2	8.1
SFT	45.2	37.6	7.6
LVR	48.6	40.5	8.1
MAPLE	**74.5**	**68.9**	**5.6**

Table 14: Correct Consistency on m-LVR (translation of an instance to receive the same correct answer).

M Hyper-parameter Sensitivity

Tables 15 and 16 analyse sensitivity to the number of prediction tokens K and the latent-objective weight λ. MAPLE is stable around $K{=}4$ and $\lambda{=}0.20$: using no prediction tokens ($K{=}0$) reduces EMC, and increasing beyond 4, yields no clear benefit. Very small λ values under-weight latent alignment; larger values slightly reduce English accuracy, reflecting a trade-off between generation quality and latent consistency.

K	En ↑	HR ↑	LR ↑	m-Avg ↑	EMC ↑
0	78.3	74.1	70.6	72.7	64.8
2	78.6	75.3	72.9	74.3	69.9
4	**78.9**	**76.8**	**74.6**	**75.9**	**74.5**
8	78.7	76.2	73.7	75.2	72.4

Table 15: Sensitivity to prediction-token count K on m-LVR.

λ	En ↑	HR ↑	LR ↑	m-Avg ↑	EMC ↑
0.05	78.5	73.2	69.8	71.8	61.9
0.10	78.8	75.4	72.9	74.4	70.8
0.20	**78.9**	**76.8**	**74.6**	**75.9**	**74.5**
0.40	77.9	75.6	73.9	74.9	73.1

Table 16: Sensitivity to latent-objective weight λ on m-LVR.

N SEED VARIANCE

Table 17 repeats the main m-LVR experiments across 3 random seeds. MAPLE's gains are consistent, especially for EMC and the HR–LR gap.

Method	En ↑	m-Avg ↑	Gap ↓	EMC ↑
SFT	77.4±0.3	65.2±0.4	5.5±0.2	45.0±0.6
LVR	78.5±0.2	67.1±0.3	5.3±0.3	48.4±0.5
MAPLE	**78.8**±0.2	**75.7**±0.3	**2.3**±0.2	**74.2**±0.4

Table 17: Seed variance on m-LVR (mean ± std 3 runs).

O TRAJECTORY-LENGTH DIAGNOSTICS

Table 18 verifies that multilingual performance is not a by-product of shorter or easier translated traces. Although several lower-resource languages produce longer textual realisations, MAPLE maintains comparatively stable accuracy.

Language	Avg. Traj. Len	MAPLE Acc. ↑	Δ_{en} ↓
English	42.3	78.9	0.0
Italian	50.6	77.1	1.8
Spanish	49.8	76.8	2.1
French	52.4	77.3	1.6
German	57.8	76.1	2.8
Russian	55.2	75.5	3.4
Chinese	38.4	77.9	1.0
Arabic	61.7	74.8	4.1
Bengali	64.1	73.6	5.3
Swahili	53.9	75.2	3.7
Telugu	66.8	74.9	4.0

Table 18: Translated trajectory length vs. downstream MAPLE accuracy in the m-LVR regime.

P LATENT-DECODING STEP SENSITIVITY

Table 19 provides the numerical values underlying Figure 3. Increasing LVR's decoding steps yields only marginal improvements in non-English accuracy and does not resolve the low EMC, meaning that the instability is structural.

Method / Steps	En ↑	HR ↑	LR ↑	Non-En ↑	EMC ↑
LVR, 1 step	77.9	68.7	63.8	66.7	45.1
LVR, 2 steps	78.4	69.1	64.0	67.1	47.3
LVR, 4 steps	**78.7**	**69.5**	**64.1**	**67.3**	**48.6**
LVR, 8 steps	78.5	69.2	63.5	66.9	46.8
LVR, 16 steps	78.0	68.6	62.8	66.3	44.0
MAPLE	**78.9**	**76.8**	**74.6**	**75.9**	**74.5**

Table 19: Effect of LVR latent decoding length on multilingual stability. Increasing steps does not substantially improve accuracy or EMC.

Q Latent Space Diagnostics

Table 20 provides geometric diagnostics supporting the visualisations in § 4.5. A lower Language Silhouette indicates less language-based clustering; a higher Task Silhouette score indicates a stronger task-based structure. View Align. measures input–trajectory cosine similarity; XL Retrieval measures whether nearest neighbours recover translated variants of the same visual instance.

Model	Lang. Silh.↓	Task Silh.↑	View Align.↑	XL Retr.↑
SFT multilingual	0.31	0.18	0.42	44.7
MAPLE w/o $\mathcal{L}_{\text{CLLC}}$	0.18	0.34	0.61	68.9
MAPLE	**0.07**	**0.46**	**0.74**	**82.3**

Table 20: Latent-space diagnostics on m-LVR. MAPLE reduces language-based clustering while increasing task-based structure, input–trajectory alignment.

R Failure Cases

A manual inspection of MAPLE errors identifies 5 recurring failures (Tables 21 and 28). The model still fails when answers depend on subtle spatial distinctions (small left/right offsets, partially occluded objects). Because MAPLE internalises visual trajectories, questions that require reading small embedded text remain challenging, especially in non-Latin scripts. Errors also persist for culturally specific objects underrepresented in backbone pretraining. Some translated questions admit broader or narrower interpretations than their English sources (colour, material, or relational terms). Finally, MAPLE sometimes confuses visually similar attributes such as patterns or textures, confirming that predictive alignment does not fully eliminate fine-grained perception limitations inherited from the underlying VLM.

Failure Type	Share	Benchmarks
Fine-grained spatial ambiguity	31.4%	V*, BLINK
OCR / text-heavy images	23.7%	MMVP, M5
Culturally specific concepts	18.6%	MARVEL, M5
Translation-induced ambiguity	14.2%	V*, BLINK
Subtle attribute distinction	12.1%	MMVP, BLINK

Table 21: Manual categorisation of MAPLE failure cases.

S Baseline Configuration

We summarise the baselines used in each operational regime. For the *single-type, single-call* setting, we compare against reconstruction-based latent reasoning methods (LVR with 4 latent decoding steps and CoVT). For the *multiple-type, single-call* setting, we compare against an SFT baseline trained on the same multilingual ThinkMorph trajectories. For the *single-type, multiple-call* setting, we compare against PixelReasoner, which explicitly invokes tools at inference time. Across all regimes, the SFT baseline uses the same trajectory data as MAPLE but without the predictive alignment, cross-lingual consistency, and latent-dynamics objectives. The zero-shot baseline is the instruction-tuned backbone evaluated without any task-specific fine-tuning.

Regime	MAPLE Training Data	Baselines	Purpose
m-LVR	Single-type, single-call trajectories	SFT, LVR, CoVT	Predictive alignment vs. reconstruction-based latent reasoning
m-ThinkMorph	Multiple-type, single-call trajectories	SFT	Heterogeneous visual-operation supervision
m-PixelReasoner	Single-type, multi-call trajectories	PixelReasoner	Sequential tool-use internalisation without inference-time tools

Table 22: Baseline configuration by operational regime.

Regime	En	It	Es	Fr	De	Ru	Zh	Ar	Bn	Sw	Te
m-LVR	18,920	18,920	18,920	18,920	18,920	18,920	18,920	18,920	18,920	18,920	18,920
m-ThinkMorph	12,480	12,480	12,480	12,480	12,480	12,480	12,480	12,480	12,480	12,480	12,480
m-PixelReasoner	9,750	9,750	9,750	9,750	9,750	9,750	9,750	9,750	9,750	9,750	9,750

Table 23: Per-language training realisations. Each translated trajectory preserves the same visual states and tool-use sequence as the English source.

Benchmark	Source	Task Focus	Base Items	Lang.	Total Items	Metric
V*	English	Object-attribute and relative-position visual search	1,024	11	11,264	Acc.
MMVP	English	CLIP-blind visual pattern robustness	300	11	3,300	Acc.
BLINK	English	Fine-grained visual perception and comparison	1,200	11	13,200	Acc.
MARVEL	Multicultural	Culturally grounded visual reasoning	2,400	5	2,400	Acc.
M5	Multilingual	General multilingual vision-language competence	1,500	8	1,500	Acc.

Table 24: Evaluation benchmark statistics.

T MULTILINGUAL TRAJECTORY EXAMPLE

Table 25 shows one multilingual training instance. The visual states are shared across languages, whereas the question, reasoning trace, and answer are realised in different linguistic forms, illustrating the construction procedure described in Appendix D.

m-LVR Trajectory Realisation Across Languages

1. Shared Visual Environment (language-agnostic)



Input Frame $\mathcal{I}_0$



Tool Visual Crop $\mathcal{I}_1$

2. Trajectory Realisation: English (anchor)

Query $\mathcal{Q}$**:** What is the young man doing with his hands?
Reasoning $\mathcal{T}_1$**:** To determine what the young man is doing with his hands, I focus on the hand region visible in the image and draw a bounding box around it.
Answer $\mathcal{T}_2$**:** He is looking at his hands.

3. Trajectory Realisation: Italian (Romance)

Query $\mathcal{Q}$**:** Cosa sta facendo il giovane con le mani?
Reasoning $\mathcal{T}_1$**:** Per determinare cosa sta facendo il giovane con le mani, mi concentro sulla regione delle mani visibile nell'immagine e disegno un bounding box attorno a essa.
Answer $\mathcal{T}_2$**:** Sta guardando le sue mani.

4. Trajectory Realisation: Swahili (Bantu/Niger-Congo)

Query $\mathcal{Q}$**:** Kijana anafanya nini na mikono yake?
Reasoning $\mathcal{T}_1$**:** Ili kujua kijana anafanya nini na mikono yake, ninazingatia eneo la mikono la mtu anayeonekana kwenye picha na kuchora kisanduku cha mipaka kulizunguka eneo hilo.
Answer $\mathcal{T}_2$**:** Anatazama mikono yake.

Table 25: Example of a multilingual training trajectory. Visual states remain fixed while the linguistic realisation varies across languages. MAPLE aligns these diverse inputs with a shared predictive latent representation.

U Detailed Multilingual Results

Table 26 provides disaggregated per-language accuracy across the 11 tested languages. MAPLE yields the largest improvements on typologically distant and lower-resource languages, including Arabic, Bengali, Swahili, and Telugu.

Regime	Method	En	It	Es	Fr	De	Ru	Zh	Ar	Bn	Sw	Te
m-LVR	Zero-Shot	73.6	66.2	65.8	66.5	64.9	62.4	68.1	61.7	58.9	60.5	59.8
	LVR	78.7	70.4	69.8	70.1	68.9	66.5	71.2	65.9	62.8	64.1	63.6
	SFT	77.6	68.5	67.9	68.3	66.4	64.1	70.2	63.5	60.7	62.9	61.8
	MAPLE	**78.9**	**77.1**	**76.8**	**77.3**	**76.1**	**75.5**	**77.9**	**74.8**	**73.6**	**75.2**	**74.9**
m-ThinkMorph	SFT	63.8	49.4	48.8	50.1	47.5	45.2	51.7	43.9	41.8	45.4	43.6
	MAPLE	**75.9**	**72.8**	**72.1**	**73.4**	**71.5**	**70.2**	**74.1**	**69.5**	**68.3**	**70.8**	**69.6**
m-PixelReasoner	PixelReasoner	**81.2**	71.1	70.4	71.8	69.3	67.5	72.9	66.2	64.8	67.1	65.9
	MAPLE	78.0	**75.8**	**75.4**	**76.1**	**74.9**	**74.1**	**76.7**	**73.6**	**72.8**	**74.0**	**73.5**

Table 26: Per-language accuracy across the multilingual perception benchmark suite. MAPLE yields more stable performance, particularly for typologically distant languages.

V Qualitative Prediction Examples

Lang.	Question	SFT	LVR/Tool	MAPLE	Gold
En	Which object is immediately to the left of the red suitcase?	B: backpack	B: backpack	C: umbrella	C: umbrella
Fr	Quel objet se trouve immédiatement à gauche de la valise rouge ?	A: sac à main	B: sac à dos	C: parapluie	C: parapluie
De	Welcher Gegenstand befindet sich unmittelbar links vom roten Koffer?	D: Flasche	A: Handtasche	C: Regenschirm	C: Regenschirm
Es	¿Qué objeto está inmediatamente a la izquierda de la maleta roja?	B: mochila	D: botella	C: paraguas	C: paraguas
En	What pattern is visible on the child's purple shoes?	B: stripes	C: flowers	D: rabbit	D: rabbit
Sw	Ni mchoro gani unaoonekana kwenye viatu vya zambarau vya mtoto?	A: nyota	C: maua	D: sungura	D: sungura
Zh	孩子的紫色鞋子上能看到什么图案?	B: 条纹	B: 条纹	D: 兔子	D: 兔子

Table 27: Qualitative examples of cross-lingual prediction stability. Questions and answer texts are shown in the evaluated language, while option labels are preserved across languages.

Failure Type	Question	Issue	MAPLE	Gold
Spatial ambiguity	Which cup is closest to the plate?	Two cups partially overlap with only a small spatial offset.	Blue cup	Green cup
OCR	What word appears on the sign above the shop?	Reading small embedded text, not only locating the sign.	Market	Bakery
Cultural concept	What object is the person holding during the ceremony?	Culturally specific object visually similar to a generic bowl.	Bowl	Ritual lamp
Translation ambiguity	Which object is behind the chair?	Spatial term can mean "after" or "beyond" depending on context.	Table	Bag
Attribute distinction	What design appears on the purple shoes?	Pattern is small and visually similar to another answer option.	Flower	Rabbit

Table 28: Representative MAPLE failure cases.