TowerVision: Understanding and Improving Multilinguality in Vision-Language Models

André G. Viveiros[*1,2], **Patrick Fernandes**[*1,2,3], **Saul Santos**[1,2], **Sonal Sannigrahi**[1,2], **Emmanouil Zaranis**[1,2], **Nuno M. Guerreiro**[4], **Amin Farajian**[5], **Graham Neubig**[3], **André F. T. Martins**[1,2,5,7]
[1]Instituto Superior Técnico, Universidade de Lisboa [2]Instituto de Telecomunicações [3]Carnegie Mellon University [4]Sword Health [5]TransPerfect [7]ELLIS Unit Lisbon
{andre.viveiros@tecnico.ulisboa.pt}

Abstract

Despite rapid progress in vision-language models (VLMs), most existing approaches remain English-centric, often relying on undisclosed training data or recipes limiting their effectiveness and reproducibility in multilingual settings. In this work, we present a systematic empirical study of how to best incorporate multilinguality across training data, encoder choices, and language models. Our results show that high-quality multilingual vision-language data substantially improves cross-lingual generalization, enabling effective transfer both from high-resource to underrepresented languages and in the opposite direction. We further find that language models with strong multilingual priors are often more effective than initializing from general-purpose language models. Guided by these findings, we design TowerVision, a family of open-source multilingual VLMs, built on the multilingual text-only model Tower+. TowerVision-9B achieves competitive performance across a range of multimodal multilingual benchmarks, with particular strength in culturally grounded tasks and multimodal translation. Notably, our models outperform existing approaches trained on substantially larger datasets, as shown on ALM-Bench and Multi30K. Along with the models, we release VisionBlocks, a high-quality, curated vision-language dataset.[1]

1 Introduction

Despite the increasing availability of vision-language models (VLMs; Comanici et al. 2025; Team et al. 2025; Bai et al. 2025a), progress has been hampered by undisclosed training data/recipes and mostly limited to English-first design choices. A key challenge in building strong multilingual VLMs stems from an asymmetric data landscape: while high-quality *text-only* multilingual corpora are relatively abundant, high-quality multilingual *vision-text* data is scarce Futeral et al. (2025). Beyond language coverage, many real-world multilingual vision-language tasks also require culturally grounded understanding. Visual concepts, symbols, practices, and social cues often carry specific cultural meaning that is not captured by vision or language alone de Dieu Nyandwi et al. (2025).

To address the imbalance above, a common strategy is to leverage large-scale text-only multilingual data while complementing it with limited high-quality multilingual multimodal supervision via translation or synthetic generation. Recent works such as Aya-Vision Dash et al. (2025) and Pangea (Yue et al., 2025) apply this strategy and extend multilinguality during the VLM fine-tuning stage. While effective, such approaches adopt specific design choices without clear guidance on 1) *which* model components most benefit from multilingual and cultural data, 2) *how* multilingual priors in the text and vision components affect

[1]All models, training recipes, and code for TowerVision will be released under open-source licenses. The VisionBlocks dataset will also be publicly available to further support open-source community research.

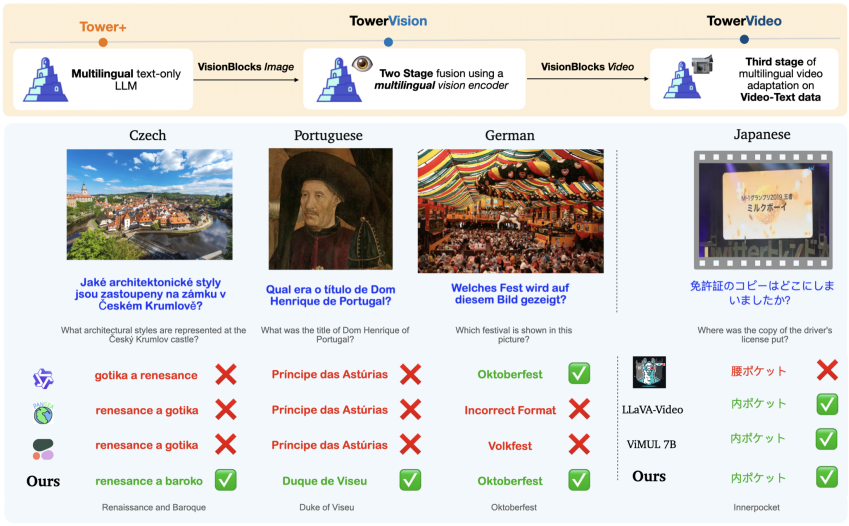

Figure 1: We present TowerVision and TowerVideo, open-source VLMs with enhanced cultural understanding and translation capabilities compared to leading open multimodal systems.

learning, and 3) *whether* expanding multilingual coverage actually improves cross-lingual generalization.

In this work, we systematically address the challenges outlined above and investigate how to improve the multilingual capabilities of VLMs from two axes: first, by performing several ablations to understand and explore the impact of the underlying components (including the alignment projector, vision encoder and text-only LLM); and second, by creating better multilingual vision-text datasets and using this data across different training stages. Guided by these findings, we introduce TowerVision[2], a suite of open-source multilingual VLMs built on top of Tower+ models (Rei et al., 2025) at 2B and 9B scales, covering 20 languages and dialects.[3] Our results demonstrate that compared to strong VLMs of similar sizes, TowerVision is competitive across several multilingual and multimodal benchmarks, and further transferring effectively to unseen languages.

We further extend our analysis to the video modality through TowerVideo, a multilingual video model built on top of TowerVision. Compared to *image-text* data, multilingual *video-text* data is substantially more scarce, making direct large-scale multilingual video training particularly challenging. We therefore investigate whether multilingual and culturally-grounded representations learned from images can transfer to video with a limited supervision signal. Our results show that this transfer is effective, with TowerVideoachieving competitive performance despite being trained with substantially less video-text data.

Complementing the TowerVision family, we also release our curated dataset, VisionBlocks, that consolidates and filters existing vision/video-language resources, further enriched with quality-filtered translations of English text descriptions into 20 languages and dialects. Overall, our work provides a unified and systematic study that offers practical guidance for future multilingual VLM development across modalities, architectures, and training stages.

[2] https://huggingface.co/collections/utter-project/towervision
[3] en, de, nl, es, fr, pt, uk, hi, zh, ru, cs, ko, ja, it, pl, ro, nn, nb.

2 TowerVision

Our approach follows the "ViT-MLP-LLM" paradigm, a multi-stage process, illustrated in Figure 1, which combines: (i) a multilingual text-only backbone model, Tower+ Rei et al. (2025); (ii) a Vision Transformer encoder (ViT; Dosovitskiy et al. 2021) that processes visual inputs and extracts meaningful features; (iii) a projector module that transforms visual features into representations compatible with the text model's embedding space. Although this training recipe and variations thereof have led to several high-quality models (e.g., LLaVA (Liu et al., 2023), Intern-VL (Chen et al., 2024), Qwen2.5-VL (Bai et al., 2025c), these models are primarily designed to optimize general, English-centric multimodal capabilities. As a result, they offer limited insight into which architectural components and training choices are responsible for improving multilingual and culturally grounded understanding, a gap we address in our work. In the remainder of this section, we present our multilingual adaptation of this framework. We first introduce the architecture and training procedure of TowerVision (§2.1), and then describe the curated multilingual vision-text dataset, VisionBlocks (§2.2).

2.1 TowerVision: Architecture & Training Details

A common approach to induce multilinguality in VLMs is to start from a strong general-purpose model optimized for English data and introduce multilinguality through continued post-training Yue et al. (2025); Shafique et al. (2025). Our ablation study in §9 challenges this practice. We find that **the choice of language backbone has a substantial impact on multilingual and cross-lingual performance**: models initialized from backbones with strong multilingual priors consistently outperform general-purpose backbones under the same training compute. For this reason, we build TowerVision on Tower+ 2B/9B (Rei et al., 2025), a multilingual Gemma-based backbone trained on high-quality curated multilingual data with a recipe designed to preserve general capabilities. As shown in §9, starting from this multilingual backbone yields substantially stronger cross-lingual performance than starting from Gemma2.

We also study the impact of multilingual priors in the *vision encoder*. Several pretrained vision backbones are available for VLMs, e.g., DINOv2 (Oquab et al., 2024), Perception Encoder (Bolya et al., 2025), or SigLIP2 (Tschannen et al., 2025), but the diversity and multilinguality of their pretraining data vary significantly. In particular, SigLIP2 is trained on more diverse and multilingual data than prior encoders, which largely rely on English-centric web data. Rather than performing a broad comparison across vision encoders, our goal is to isolate the effect of multilingual visual priors, so we directly compare SigLIP2 against SigLIP1 Zhai et al. (2023), its previous iteration without explicit multilingual training, while keeping all other components fixed. Our ablation results in §4 show that **multilingual priors in the vision encoder are especially beneficial in data constrained environments.** When multimodal multilingual data is limited, SigLIP2 consistently yields stronger multilingual and cross-lingual performance. As the amount of training data increases, the gap narrows, and models relying on less multilingual priors can partially achieve the same performance. Based on these findings and the resolution ablations reported in §A.3, we adopt SigLIP2-so400m/14@384px, a Vision Transformer operating at 384×384 resolution that extracts image patch representations and produces multilingual-aligned embeddings of size 1024.

To align the vision and text modalities, we use a LLaVA-based architecture (Liu et al., 2023), where we train a projection layer consisting of a 2-layer MLP, randomly initialized. By combining Tower+ for text and SigLIP2 for vision, TowerVision benefits from complementary multilingual strengths across both modalities. The training process consists of three stages:

- A *projector pretraining* phase, where we train the model to predict captions from images using high-quality image-caption data, while freezing both the vision encoder and the language model backbone (i.e., only the projector is trained). Each image is encoded once (downscaled to 384×384 if necessary). During this phase, **we focus exclusively on diverse, high-quality English captions**, which we show to be more effective

for aligning visual and textual representations (see §4), rather than attempting broad multilingual coverage at this stage.

- A *vision finetuning* phase, where we unfreeze the full model and train it on the VisionBlocks dataset (§2.2). In this phase, we use *high-dynamic resolution* (Liu et al., 2024a), breaking high-resolution images into a grid of smaller tiles which are then encoded with the vision encoder independently (together with a global thumbnail tile). The projected embeddings are then concatenated. We use a maximum of 6 tiles both at training and inference, which provides the best trade-off between spatial detail and downstream performance across both English and multilingual benchmarks (§A.3). This phase leads to the TowerVision model.
- A *video finetuning* phase, where the video portion of VisionBlocks is used to finetune TowerVision on 32-frame video inputs at the encoder's fixed resolution of 384×384, providing a favorable trade-off between temporal coverage and computational cost. This phase leads to the TowerVideo model.

The models were trained on a custom fork of the LLaVA-Next (Liu et al., 2024a) codebase.[4]

2.2 VisionBlocks: Towards Better Multilingual Vision-Text Data

Creating a large-scale, high-quality, multilingual multimodal dataset that captures cultural nuances for training VLMs is non-trivial for a series of intertwined reasons: (1) *human-written* vision-text data is severely limited; while large-scale captioning datasets such as LAION-5B Schuhmann et al. 2022 exist in abundance, they prioritize scale over quality, which is suboptimal for training VLMs with advanced capabilities such as instruction-following (Dong et al., 2025; Zhou et al., 2023); (2) high-quality *multilingual* vision-text data remains scarce, and the lack of open, high-quality multilingual VLMs makes controlled synthetic data generation challenging the most viable alternative, also adopted by Pangea (Yue et al., 2025), is to translate English vision–text interactions into target languages; and (3) filtering techniques such as reward model scoring or LLM-as-judge approaches (Gu et al., 2025) are significantly harder to apply to vision–text data, where even frontier VLMs struggle to provide reliable preference judgments (Li et al., 2024). With this in mind, we develop and release VisionBlocks (see §A.1), a dataset comprising 6M vision-text instances, aggregated and filtered from multiple sources and further enhanced with new translated and synthetic data.

Collection of existing VLM data For English vision-text data, we use the mixture created in PixMo (Deitke et al., 2024) with a few minor changes: we exclude the Android-Control, Points, and PointQA datasets, as they do not provide additional multilingual value at this stage. For culturally and multilingual vision-text data, we leverage a subset of "Open-Ended" and "Multiple-Choice" questions from CulturalGround (de Dieu Nyandwi et al., 2025) and the "Cultural" split of PangeaIns (Yue et al., 2025) for our languages of interest.

Translated and synthetic generated vision-language data In addition to the original English and multilingual captions, we translate the highly curated PixMo-Cap caption data Deitke et al. (2024) to our target languages using a Tower+ 9B model (Rei et al., 2024). These translations are scored using CometKiwi (Rei et al., 2022) and filtered with a high threshold of 0.85 to ensure maximum quality. To further enhance diversity, we pair the remaining high-quality translations with a variety of language-specific captioning prompt templates (§A.7.1). We also augment the dataset with synthetic captions generated by Gemini 2.5. For each image, we sample multiple prompts to elicit diverse and detailed descriptions (see §A.7.2).

Text-only data We include a fixed proportion of text-only data in the multimodal SFT mixture to preserve the text-only capabilities of the backbone LLM. In all experiments, we use a 20% text-only ratio, which we found to provide the best trade-off between text-only and multimodal benchmark performance. We use EuroBlocks (Martins et al., 2025), a curated multilingual collection of high-quality instruction-aligned synthetic data.

[4] https://github.com/deep-spin/LLaVA-NeXT

Translated multilingual video data As video-text data, we employ the LLaVA-Video-178k dataset (Zhang et al., 2025b), which contains captions alongside open-ended and multiple-choice English questions. To construct a multilingual version, we randomly split the dataset into two halves: one half is kept in English, while the other is uniformly translated into all supported languages using TOWER+9B (Rei et al., 2025). All translations are scored with COMETKIWI and filtered using a high-quality threshold of 0.85.

3 EVALUATION & MAIN RESULTS

We evaluate TOWERVISION and TOWERVIDEO on a comprehensive suite of benchmarks spanning multiple modalities and task types (single-image, multi-images, and video) across many different languages, both within and beyond our training set. In this section, we focus on vision-language tasks, which includes multilingual visual/video and visual reasoning. Our assessment relies primarily on closed-form tasks, complemented by large language models serving as judges for video-based evaluations.

3.1 TASKS & EVALUATION BENCHMARKS

For **Vision-language tasks** we report results on ALM-Bench (Vayani et al., 2024), a cultural understanding multilingual[5] visual QA benchmark, OCRBench (Liu et al., 2024b) for OCR-centric capabilities, TextVQA (Singh et al., 2019), assessing scene-text reasoning capabilities, and XGQA[6], which evaluates cross-lingual visual reasoning. For **Multimodal translation** We report results on CoMMuTE (Futeral et al., 2023), a specialized multimodal translation benchmark that uses the visual content to resolve lexical ambiguities present in the source language, and Multi30K (Elliott et al., 2016), a standard benchmark for multimodal machine translation (MT) of image captions. For **Culturally-aware multilingual video tasks** We use the open-ended split of ViMUL-Bench (Shafique et al., 2025), a multilingual video QA benchmark [7]. The dataset contains questions covering culturally diverse domains such as festivals, customs, food, and heritage.

3.2 BASELINES

For evaluation, we leverage lmms-eval (Zhang et al., 2025a), a framework which enables a systematic comparison of TOWERVISION against leading open VLMs. We include several multilingual multimodal models, such as *CulturalPangea-7B*, designed to address gaps in multilingual cultural understanding, and *Aya-Vision-8B*, optimized for a broad range of vision-language tasks. In addition, we evaluate models from the *Gemma3-Instruct* and *Qwen2.5-VL-Instruct* families, both of which have demonstrated strong performance across a variety of multimodal benchmarks. Finally, we report results for a LLaVA-based model, *Llama3-Llava-Next-8B*, a general-purpose VLM with strong performance across a wide range of tasks. The checkpoints for all models are listed in §A.2. For TOWERVIDEO, we consider several competitive open-source video models of comparable scale, including *VideoLLaMA3-7B*, *LLaVA-Video-7B* also trained on LLaVA-Video-178k, *ViMUL-7B*, a multilingual video model, and *Qwen2.5-VL-7B-Instruct*, a strong general-purpose vision model.

3.3 MAIN RESULTS

Tables 1 and 2 report the performance of TOWERVISION on vision-language benchmarks as well as multimodal translation benchmarks, while Table 3 reports the results on the multilingual video-language benchmark. We summarize the main findings below.

TowerVision models have strong performance in cultural-aware tasks. Within our suite of vision-language benchmarks, we achieve state-of-the-art results on ALM-Bench (Table 1), a culturally diverse benchmark, in both the English and multilingual split.

[5] de, es, fr, it, ko, nl, ru, en, pt, zh, zh, is, cs, uk, hi, ja, pl, sv, hu, ro, da, nn, fi

[6] en, de, ko, pt, zh

[7] en, fr, de, hi, ru, es, sv, ta.

Table 1: **Vision-Language Model Performance.** English and multilingual VLMs results across multiple benchmarks. Reported values correspond to final accuracy (↑).

	English (↑)		**Multilingual** (↑)		
	TextVQA	OCRBench	XGQA	ALM-Bench (en)	ALM-Bench (multi)
Qwen2.5-VL-3B-Instruct	77.8	78.7	42.1	81.0	76.2
Qwen2.5-VL-7B-Instruct	**82.5**	**84.5**	47.2	83.1	83.6
Gemma3-4B-it	65.2	74.2	44.2	79.7	80.0
Gemma3-12B-it	73.2	74.7	**48.1**	83.5	84.5
CulturalPangea7B	69.8	63.5	37.9	61.3	65.2
Llama3-Llava-Next-8B	64.8	54.4	40.1	76.5	73.4
Aya-Vision-8B	66.9	61.0	36.2	78.2	77.3
TowerVision-2B	68.1	58.6	39.1	77.2	81.1
TowerVision-9B	76.6	72.7	47.2	**86.1**	**85.0**

Table 2: **Multimodal Translation Benchmarks.** We report xComet (Guerreiro et al., 2024) for Multi30K and contrastive pairwise accuracy for CoMMuTE.

	Multi30K (↑)			CoMMuTE (↑)			
	en→cs	en→de	en→fr	en→de	en→fr	en→ru	en→zh
Qwen2.5-VL-3B-Instruct	83.3	96.7	92.6	71.6	74.4	77.5	81.5
Qwen2.5-VL-7B-Instruct	83.9	97.1	93.2	74.7	76.9	77.2	**82.4**
Gemma3-4B-it	33.4	44.0	33.2	**76.7**	78.2	**79.0**	74.4
CulturalPangea7B	80.0	95.8	92.1	68.3	77.3	75.3	79.3
Llama3-Llava-Next-8B	79.1	93.3	88.1	72.0	74.4	74.4	73.5
Aya-Vision-8B	94.4	97.9	95.3	69.3	76.9	74.4	76.2
TowerVision-2B	90.3	97.5	94.7	70.0	74.3	73.2	76.6
TowerVision-9B	**95.1**	**98.1**	**95.6**	72.0	**78.8**	75.6	77.4

Qwen2.5VL-7B and Gemma3-12B are the closest competitors, while other baselines lag behind. In the multilingual split, we evaluate on a diverse set of 23 languages covering several language families and scripts. TowerVision is able to exhibit enhanced cultural multimodal understanding, even in unseen languages. To better contextualize these gains, we provide side-by-side qualitative comparisons that highlight differences in cultural grounding in §A.6 that illustrate cultural awareness across languages and regions. We further assess the cross-lingual generalization capabilities of TowerVision in §4.

TowerVision-2B is competitive with larger models in multilingual settings. In multimodal translation benchmarks, TowerVision consistently demonstrates strong performance on Multi30K and is competitive on CoMMuTE (Table 2). Our 9B variant achieves state-of-the-art results on Multi30k across all language pairs, and we observe that even our smaller 2B variant is a competitive model against the larger baselines on translation-specific, as well as vision-language benchmarks.

Multilingual fine-tuning improves cross-lingual performance in TowerVideo. In Table 3, we report the GPT-4o–judged open-ended responses (OpenAI et al., 2024), following the same absolute scoring evaluation protocol as Shafique et al. (2025). In particular, the image–text model TowerVision, without any video fine-tuning, already achieves competitive performance, indicating that multilingual and cultural representations learned at the image level generalize directly to video understanding. With the addition of limited video supervision, using substantially smaller datasets and no culturally grounded video annotations TowerVideo attains similar and even superior performance across several languages, surpassing models trained with an order of magnitude more data. These results highlight that our multilingual recipe learns transferable, culturally grounded representations that generalize across modalities. [8]

Overall, our results demonstrate the effectiveness of our design choices in endowing our model with strong multilingual capabilities due to a combination of increased multilingual

[8] All GPT-4o evaluation prompts used in our experiments are available in Appendix §A.8.

Table 3: **Multilingual video performance.** Percentage share (pct% of preference shares across responses) and average score (avg) on ViMUL-Bench across selected languages.

Model	en	fr	de	hi	ru	es	sv*	ta*
Qwen2.5-VL-7B	**17.8%/2.8**	14.9%/2.5	**16.3%/2.6**	13.5%/1.7	**16.9%/2.6**	15.0%/2.4	15.6%/2.2	**20.1%/1.0**
VideoLLaMAA3-7B	12.9%/2.2	11.3%/1.9	10.2%/1.6	10.3%/1.3	11.9%/1.9	12.1%/2.1	9.1%/1.3	9.7%/0.5
ViMUL-7B	10.8%/1.7	11.7%/2.0	11.4%/1.8	11.7%/1.6	12.7%/2.0	11.5%/2.0	11.7%/1.7	14.7%/0.8
LLaVA-Video-7B	10.2%/1.9	12.5%/2.2	11.1%/1.8	9.7%/1.4	10.3%/1.6	10.9%/2.0	9.1%/1.3	12.2%/0.6
towervision-2B	5.7%/1.7	9.6%/1.7	9.7%/1.6	5.5%/1.1	8.2%/1.3	10.4%/1.8	9.7%/1.4	10.0%/0.5
towervideo-2B	12.9%/2.1	13.0%/2.2	13.1%/2.1	16.6%/2.1	13.0%/2.0	12.3%/2.1	14.8%/2.1	9.8%/0.5
towervision-9B	14.3%/2.3	11.9%/2.1	13.6%/2.2	13.8%/1.8	12.6%/2.0	12.5%/2.1	13.7%/2.0	13.6%/0.7
towervideo-9B	15.2%/2.4	**15.1%/2.6**	14.7%/2.4	**19.0%/2.4**	14.5%/2.3	**15.2%/2.5**	**16.4%/2.3**	9.9%/0.5

culturally-sensitive training data, a stronger multilingual text backbone (Tower+), and a multilingual vision encoder. We detail these choices in §4 with a carefully conducted set of ablation experiments.

4 Where and How Does Multilinguality Matter?

In this section, we analyze where and how multilinguality affects different components and training stages of TowerVision.

Multilingual backbones improve cross-modal performance. The choice of backbone in TowerVision can substantially influence performance across multilingual and multimodal tasks. We focus on two complementary aspects. First, we examine the significance of multilingual capacity by comparing the Tower+ backbone, which is highly multilingual and designed for general-purpose multilingual text tasks, against Gemma2, the model on which Tower+ was built. Second, we investigate the impact of instruction tuning before modality fusion, which is widely applied in modern VLMs from the start (Liu et al., 2023; Bai et al., 2025b), but whose effect on the final model remains unclear. To study these effects, we train TowerVision at 2B and 9B scales using four backbones: Gemma2-pt (pretrained), Gemma2-it (instruction-tuned), Tower+pt (pretrained Tower+), and Tower+it (instruction-tuned Tower+), following the recipe in §2. As shown in §A.4, using Tower+ consistently outperforms Gemma2, confirming the importance of a strong multilingual backbone for robust cross-modal understanding. At smaller scales, non-instructed models (Gemma2-pt, Tower+pt) retain stronger raw visual extraction, while instruction-tuned variants excel in cultural knowledge and reasoning. By the 9B scale, this gap narrows, with instruction-tuned models integrating both skills and achieving state-of-the-art performance. These findings underscore the complementary roles of multilingual pretraining and instruction tuning, and the need for careful backbone selection in VLMs.

Multilingual-aware vision encoders improve performance in low-data regimes.

Effectively leveraging multilingual data is crucial for VLMs, yet it is unclear whether the vision encoder's own multilingual capacity plays an important role. We compare SigLIP2, trained on diverse multilingual data, with SigLIP1, an earlier English-centric version, to test whether multilingual-aware encoders are essential or if sufficient fine-tuning can compensate. We train TowerVision with both encoders on English-only and multilingual data at 2B and 9B scales (results in Table 4). Without additional multilingual data, SigLIP2 models consistently outperform SigLIP1, showing clear benefits in low data regimes. With multilingual finetuning, however, the gap narrows, showing that finetuning with sufficient multilingual data can compensate for a weaker encoder. At 9B scale, both converge to strong performance. In short, multilingual-aware encoders provide an advantage when data is scarce, but extensive multilingual training can close the gap.

Table 4: Multilingual impact of different vision encoders measured on ALM-Bench.

TowerVision	2B		9B	
Variant	En	Multi	En	Multi
SigLIP1-En	67.4	60.2	78.3	81.2
SigLIP2-En	69.3	67.1	77.2	81.1
SigLIP1-(En+Multi)	76.6	80.7	83.6	84.9
SigLIP2-(En+Multi)	**77.1**	**81.1**	**83.6**	**85.2**

Expanding languages improves cross-lingual generalization in VLMs. We study how language coverage in training data impacts performance on both included and excluded languages. Specifically, we compare training on 10 high-resource "core languages" versus the full set of languages, while controlling for dataset size. Our questions are: (i) whether adding balanced multimodal data for more languages improves performance on core languages (Conneau et al., 2020; Hu et al., 2020), and (ii) whether unsupported languages benefit in zero-shot fashion if related languages are present (Ni et al., 2021). We train TowerVision at 2B and 9B scales using the recipe in §2, first on 10 "core" languages (English, German, Dutch, Portuguese, Russian, Simplified and Traditional Chinese, Spanish, French, Italian), then on all available languages. Detailed per-language results reported in Appendix A.5 show that broader language coverage consistently improves performance, with larger gains at the 2B scale. Importantly, English performance is not degraded and often exhibits positive gains across both scales, indicating positive transfer rather than interference. We also observe zero-shot improvements for languages not explicitly included in the training set, further supporting cross-lingual transfer when related languages are included. These findings highlight the value of expanding multilingual data, particularly for smaller models.

High-quality English captions are enough to ensure strong alignment. To assess whether multilingual supervision is necessary during alignment pretraining, we train two versions of TowerVision on both scales, 2B and 9B. The first version uses only English-only captions from PixMo-Cap, comprising 702, 205 text-image pairs. The second version uses the same English captions combined with a high-quality translated subset from PIXMO-CAP, where data was uniformly translated into the supported languages as described in §2.2, comprising 367,779 samples. We evaluate the models in ALM-Bench to measure TowerVision performance both in English and across multiple non-English languages, providing insights into how well cross-lingual generalization is preserved or improved. As shown in Table 5, adding high-quality multilingual captions during the projector alignment stage has little to no positive effect and, in some cases, slightly decreases performance on the multilingual subset. This suggests that the most effective strategy is to focus on diverse and high-quality captions, ensuring strong alignment between visual and textual modalities, rather than prioritizing extensive multilingual coverage at this stage.

Table 5: Effect of using multilingual versus English-only captions during projector pretraining on ALM-Bench.

TowerVision	**2B**		**9B**	
Projector	**En**	**Multi**	**En**	**Multi**
En	77.1	**81.1**	**83.6**	**85.2**
En+Multi	**77.9**	79.3	83.0	84.1

5 Conclusion

We provided a detailed training recipe for developing strong multilingual vision-language models, covering data, encoders, and text backbones, complemented by an extensive ablation study on key components of our approach. Our systematic empirical study details dataset curation and various design decisions, with the results of our ablations challenging several current practices. As a result of this work, we introduced TowerVision, a suite of open-source multimodal models for image-text and video-text tasks, designed with a strong emphasis on cultural understanding and multilinguality. Our models demonstrate competitive, and in several cases improved, multilingual performance across a range of benchmarks when compared with existing open multimodal systems. Alongside this, we released VisionBlocks, a high-quality vision-language dataset, and provided a detailed training recipe covering data, encoders, and text backbones, complemented by an extensive ablation study on key components of our approach.

We hope that these contributions, spanning models, data, and methodology help advance open-source research on culturally diverse multilingual multimodal models, and accelerate progress toward narrowing the performance gap with closed-source English-centric settings.

Acknowledgments

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for ResponsibleAI), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

Ethics Statement

This work develops and evaluates multilingual vision-language models using publicly available datasets as well as our own synthetic and translated data. We acknowledge potential risks, including biased model outputs and unintended misuse of generated content. While we have taken steps to ensure diversity and maximum data quality, we always encourage careful evaluation and responsible deployment of these models in real-world scenarios. Our research does not involve sensitive personal data or tasks with direct safety-critical impact.

Reproducibility Statement

This work provides detailed descriptions of the data, model architectures, training procedure (including the codebase), and evaluation benchmarks used. All datasets used are either publicly available or created by our team (synthetic and translated), with the respective system prompts shared for maximum transparency. Additionally TowerVision all the collection of models, code for data preprocessing, training, and evaluation will be released to facilitate replication of our results. We aim to ensure that other researchers can reproduce our findings with minimal effort.

We ensure reproducibility by providing detailed descriptions of the data, model architectures, training procedures, and evaluation benchmarks. Upon acceptance, we will release the VisioBlocks dataset, checkpoints of the TowerVision collection models[9], and the corresponding codebases for training and evaluation[10], to facilitate replication of our results.

References

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, 2025a. URL `https://arxiv.org/abs/2511.21631`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL `https://arxiv.org/abs/2502.13923`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

[9] https://huggingface.co/collections/utter-project/towervision

[10] https://github.com/deep-spin/LLaVA-NeXT

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL https://arxiv.org/abs/2504.13181.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. URL https://arxiv.org/abs/1911.02116.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya vision: Advancing the frontier of multilingual multimodality, 2025. URL https://arxiv.org/abs/2505.08751.

Jean de Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. Grounding multilingual multimodal llms with cultural knowledge, 2025. URL https://arxiv.org/abs/2508.07414.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation, 2025. URL https://arxiv.org/abs/2501.05952.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In Anya Belz, Erkut Erdem, Krystian Mikolajczyk, and Katerina Pastra (eds.), *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL https://aclanthology.org/W16-3210/.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5394–5413, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.295. URL https://aclanthology.org/2023.acl-long.295/.

Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. moscar: A large-scale multilingual and multimodal document-level corpus, 2025. URL https://arxiv.org/abs/2406.08707.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020. URL https://arxiv.org/abs/2003.11080.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. Vlrewardbench: A challenging benchmark for vision-language generative reward models, 2024. URL https://arxiv.org/abs/2411.17451.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, et al. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*, 2025.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Jianfeng Gao, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training, 2021. URL https://arxiv.org/abs/2006.02635.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela

Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry,

Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60/.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 185–204, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.12. URL https://aclanthology.org/2024.wmt-1.12/.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. Tower+: Bridging generality and translation specialization in multilingual llms, 2025. URL https://arxiv.org/abs/2506.17080.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, Leon Schaller, Roman Tymtsiv, Simon Weber, Hisham Cholakkal, Ivan Laptev, Shin'ichi Satoh, Michael Felsberg, Mubarak Shah, Salman Khan, and Fahad Shahbaz Khan. A culturally-diverse multilingual multimodal video benchmark & model. *arXiv preprint arXiv:2506.07032*, 2025. URL https://arxiv.org/abs/2506.07032.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL `https://arxiv.org/abs/2502.14786`.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*, 2024.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy, 2024. URL `https://arxiv.org/abs/2412.02210`.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham

Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages, 2025. URL https://arxiv.org/abs/2410.16153.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2025b. URL https://arxiv.org/abs/2410.02713.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023. URL https://arxiv.org/abs/2305.11206.

Table 6: Overview of dataset composition across categories. Each dataset lists its sample size with the proportion of the total in parentheses, along with its collection type tag (Public Data, Synthetic (Generated), or Translated (Augmented)). Totals are shown for English-only and Multilingual subsets, as well as the overall dataset size.

Category	Dataset	Samples (%)	Tag
Chart/Plot	DVQA	199,995 (3.17%)	Public Data
	ChartQA	25,055 (0.40%)	Synthetic (Generated)
	PlotQA	157,070 (2.49%)	Public Data
	TabMWP	22,717 (0.36%)	Public Data
General VQA	VQAv2	428,708 (6.79%)	Public Data
	RLAIF-4V	59,408 (0.94%)	Synthetic (Generated)
Doc VQA	DocVQA	9,664 (0.15%)	Synthetic (Generated)
	TextVQA	15,690 (0.25%)	Synthetic (Generated)
	ST-VQA	17,242 (0.27%)	Public Data
	PixMo-Docs	3,634 (0.06%)	Public Data
Reasoning/Knowledge	A-OKVQA	11,853 (0.19%)	Synthetic (Generated)
	OKVQA	9,009 (0.14%)	Public Data
	AI2D	7,791 (0.12%)	Public Data
	ScienceQA	758 (0.012%)	Public Data
Multilingual/Cultural	Pangea-Cultural	55,438 (0.88%)	Public Data
	Pangea-Multi	428,838 (6.79%)	Public Data
	PixMo-Cap-Translated	367,779 (5.83%)	Translated (Augmented)
	CulturalGround-OE	401,149 (6.35%)	Public Data
	CulturalGround-MCQs	379,834 (6.02%)	Public Data
Specialized VQA	IconQA	19,543 (0.31%)	Synthetic (Generated)
	InfographicVQA	2,049 (0.03%)	Synthetic (Generated)
	Stratos	12,585 (0.20%)	Public Data
Counting/Math	TallyQA	98,675 (1.56%)	Public Data
	PixMo-Count	8,128 (0.13%)	Public Data
Vision/Text	VBlocks-PixMo-AMA	154,336 (2.44%)	Public Data
	VBlocks-PixMo-Cap	702,205 (11.12%)	Public Data
	VBlocks-PixMo-CapQA	262,862 (4.16%)	Public Data
	EuroBlocks-SFT	1,094,265 (17.34%)	Public Data
Video/Text	LLaVA-Video-178k-subset	697,618 (11.05%)	Public Data
	LLaVA-Video-178k-translated	697,617 (11.05%)	Translated (Augmented)
	Total (English)	3,982,630 (63.1%)	
	Total (Multilingual)	2,330,656 (36.9%)	
	Overall Total	6,313,286 (100%)	

A Appendix

A.1 Full Description of VisionBlocks

Table 6 shows the full details and statistics of the VisionBlocks dataset.

A.2 Models Checkpoints

Table 7 lists all model checkpoints used for comparative baselines. We use checkpoints released HuggingFace when possible.

A.3 Vision Encoder Variants

Beyond selecting a more multilingual vision encoder, several other factors significantly influence its performance. These include the input image resolution supported by the encoder, the number of patches it uses, which determines the total number of visual tokens for a

Model	Params	Checkpoint Link
Qwen2.5-VL-Instruct	3B	https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-Instruct	7B	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
Gemma2-it	2B	https://huggingface.co/google/gemma-2-2b-it
Gemma2-pt	2B	https://huggingface.co/google/gemma-2-2b
Gemma2-it	9B	https://huggingface.co/google/gemma-2-9b-it
Gemma2-pt	9B	https://huggingface.co/google/gemma-2-9b
Gemma3-it	4B	https://huggingface.co/google/gemma-3-4b-it
Gemma3-it	12B	https://huggingface.co/google/gemma-3-12b-it
CulturalPangea	7B	https://huggingface.co/neulab/CulturalPangea-7B
LLaVA-Next	7B	llava-hf/llava-v1.6-mistral-7b-hf
Aya-Vision	8B	https://huggingface.co/CohereForAI/aya-vision-8b
Pixtral	12B	https://huggingface.co/mistralai/Pixtral-12B-2409
Phi-4-Multimodal	14B	https://huggingface.co/microsoft/Phi-4-multimodal-instruct

Table 7: **Model checkpoints.** Parameters and HuggingFace links for models included in our evaluation suite.

given image resolution (e.g, for an img resolution of 224×224 using patch size of 14 we obtain 256 visual tokens) and the number of tiles.

Our goal is to empirically identify the optimal configuration for processing visual inputs, focusing on these three factors.

Specifically, we perform experiments using the TowerVision 2B version with variants of SigLIP2 framework:

1. Image resolution: We vary the input image size between 384×384, 224×224, and 512×512 to examine its effect on feature extraction quality.
2. Patch numbers: We evaluate different patch sizes (14 and 16) to assess how granularity impacts the learned representations. Smaller patches capture finer details but increase the number of tokens, affecting the context length the model must handle.
3. Number of tiles: Beyond the default 6 tiles, we also experiment with 4 and 22 tiles. The number of tiles is adjusted to the image resolution: lower-resolution images (e.g, 224×224) require more tiles to cover the same amount of visual information as a higher-resolution encoder (e.g., 512×512). For example, an image with resolution $(1024, 1024)$ processed with a 512×512 encoder requires roughly 4 tiles to cover the full image, whereas a 224×224 encoder would need at least 25 tiles (including padding) to achieve similar coverage. This creates a trade-off between capturing detailed local information and maintaining manageable context length.

These experiments allow us to systematically compare variations while keeping other components constant, providing insights into which configuration yields the best overall performance. The results are reported in Table 8, which highlights the trade-offs between resolution, patch granularity, and style diversity. We included an additional multilingual OCR benchmark (cc-OCR Yang et al. (2024)) to evaluate the impact of image resolution, as OCR performance is particularly sensitive to image resolution.

A.4 Backbone ablations

A.5 Cross-Lingual Generalization

Looking at Table 10, we observe that cross-lingual transfer gains at larger scales tend to be lower. This finding is somewhat counter-intuitive, as one would expect the transfer to be

Table 8: **Impact of Vision Encoder Configuration and Instruction Tuning.** Evaluation of TOWER+ models across English and multilingual tasks with varying image resolution, patch size, and number of tiles. Results highlight how these design choices affect overall performance.

Resolution	Patch Size	Tiles	English		Multilingual	
			TextVQA	OCRBench	CC-OCR	ALM-Bench
224x224	14	22	59.1	53.3	37.2	70.5
224x224	16	20	68.6	57.8	44.3	75.2
384x384	14	6	**70.3**	**62.1**	**46.1**	**75.6**
512x512	16	4	64.0	55.7	39.6	74.7

Table 9: **Impact of backbone and instruction tuning.** Performance of VLMs with different backbones on English and multilingual tasks. TOWER+ consistently outperforms GEMMA2, demonstrating the benefit of a highly multilingual backbone.

Backbone Model	English (↑)		Multilingual (↑)	
	TEXTVQA	OCRBENCH	ALM-BENCH (EN)	ALM-BENCH (MULTI)
GEMMA2-PT-2B	69.2	61.2	74.3	76.7
TOWER+PT-2B	**70.3**	62.1	73.0	78.2
GEMMA2-IT-2B	70.0	**63.0**	75.0	75.1
TOWER+IT-2B	68.1	58.6	**77.1**	**81.1**
GEMMA2-PT-9B	72.4	66.6	79.9	79.6
TOWER+PT-9B	73.2	64.5	81.3	84.4
GEMMA2-IT-9B	**74.4**	67.2	79.6	81.5
TOWER+IT-9B	73.6	**69.7**	**83.6**	**85.2**

on-par or even higher for larger models, since they can model more complex distributions and generalize better. However, our results indicate increased interference at larger scales, resulting in lower average performance and more regressions. A more careful analysis of this phenomenon is left for future work.

A.6 Qualitative Analysis of Cultural Grounding

Across diverse cultural and linguistic contexts, TOWERVISION seems to outperform strong baselines by leveraging culturally grounded visual and textual cues rather than relying on generic visual patterns or surface-level correlations. We highlight several representative examples below.

A.7 System Prompts

A.7.1 Tower System Prompts used for Translation

The prompts vary in style and specificity to improve diversity and capture nuanced meaning from the original English captions. They are grouped by language and include multiple phrasings for the same instruction to encourage robust translations.

```
# English prompts
EN_PROMPTS = [
    "Describe this image.",
    "What can you see in this picture?",
    "Tell me what's in this image.",
    "Explain what this image shows.",
    "Caption this image.",
    "What's happening in this picture?",
    "Provide a description of this image."
```

Table 10: Cross-lingual performance of TowerVision models at 2B and 9B scales, evaluated on the ALM-Bench benchmark. *Core Langs* refers to a set of 10 languages: English, German, Dutch, Portuguese, Russian, Simplified and Traditional Chinese, Spanish, French and Italian. *Core+Added Langs* includes all languages supported by TowerVision as indicated in footnote 3. *Unseen* languages are those not encountered during training and are marked with an asterisk (*). Bold values indicate the better result within each scale. Positive gains from adding languages are highlighted in light green, negative gains in light red.
Overall, adding more languages tends to improve performance across the board, demonstrating strong cross-lingual transfer capabilities, even for unseen languages.

	TowerVision-2B			**TowerVision-9B**		
Metric / Lang	Core Langs	Core + Added Langs	Gain	Core Langs	Core + Added Langs	Gain
English (en)	60.9	**76.6**	+15.8	70.3	**82.8**	+12.5
Core Avg	65.3	**81.3**	+16.1	81.5	**82.6**	+1.1
Added Avg	60.2	**75.4**	+15.2	76.3	**84.3**	+7.6
Unseen Avg	69.2	**83.0**	+13.9	81.2	**82.5**	+1.2
German (de)	75.9	**84.5**	+8.6	**89.7**	87.9	-1.8
Spanish (es)	56.6	60.5	+3.9	73.7	**76.3**	+2.6
French (fr)	76.9	**82.7**	+5.8	**86.5**	80.8	-5.7
Hindi (hi)	44.2	**75.0**	+30.8	82.7	80.8	-1.9
Italian (it)	75.0	**81.7**	+6.7	**96.7**	**98.3**	+1.6
Korean (ko)	**76.4**	70.8	-5.6	75.0	**79.2**	+4.2
Dutch (nl)	70.0	**86.7**	+16.7	**90.0**	86.7	-3.3
Portuguese (pt)	64.5	**90.3**	+25.8	85.5	**91.9**	+6.4
Romanian (ro)	58.9	**80.4**	+21.5	75.0	**87.5**	+12.5
Czech (cs)	61.4	**75.7**	+14.3	74.3	**90.0**	+15.7
Russian (ru)	65.5	**84.5**	+19.0	65.5	**75.9**	+10.4
Chinese (simp.) (zh-hans)	50.0	**87.5**	+37.5	68.8	71.9	+3.1
Chinese (trad.) (zh-hant)	53.8	**76.9**	+23.1	61.5	67.3	+5.8
Danish (da)*	66.1	70.9	+4.8	**90.3**	86.3	-4.0
Finnish (fi)*	56.0	**82.0**	+26.0	70.0	72.0	+2.0
Hungarian (hu)*	68.8	**95.3**	+26.5	79.7	**82.8**	+3.1
Icelandic (is)*	67.6	**76.5**	+8.9	76.5	**83.8**	+7.3
Japanese (jp)*	**78.8**	78.9	0.1	**84.8**	80.3	-4.5
Swedish (sv)*	77.6	**94.8**	+17.2	86.2	**89.7**	+3.5

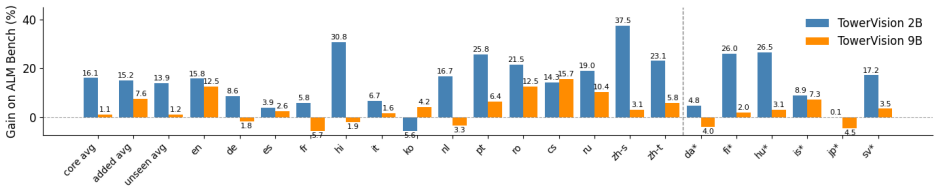

Figure 2: Performance of TowerVision models on 10 vs 20 languages/dialects at 2B and 9B scales. The bars indicate the accuracy gains by training on 20 (all) versus 10 (core) languages (more details in §A.5). Unsupported languages are marked with *.

```
]

# European Portuguese prompts
PT_PROMPTS = [
    "Descreva esta imagem.",
    "O que consegue ver nesta fotografia?",
    "Diga-me o que está nesta imagem.",
    "Explique o que esta imagem mostra.",
    "Legende esta imagem.",
    "O que se passa nesta fotografia?",
    "Forneça uma descrição desta imagem."
]

# French prompts
FR_PROMPTS = [
```

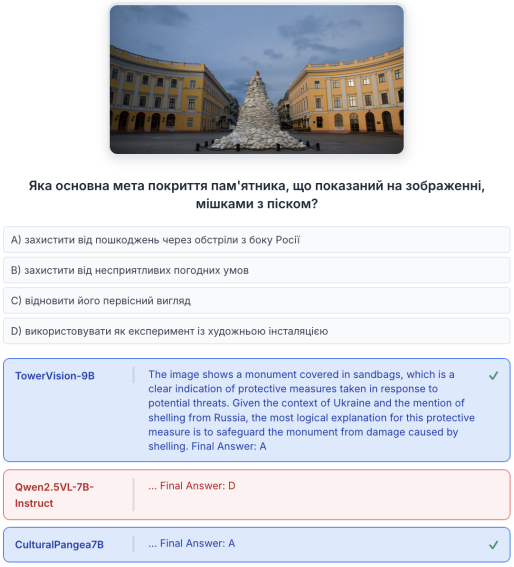

Figure 3: Ukrainian cultural context; Task: Identify the purpose of covering a monument with sandbags. TowerVision correctly identifies the sandbagging as a protective measure against wartime shelling, while Qwen2.5-VL misinterprets the scene as an artistic installation. This highlights TowerVision's ability to integrate visual cues with region-specific geopolitical context.

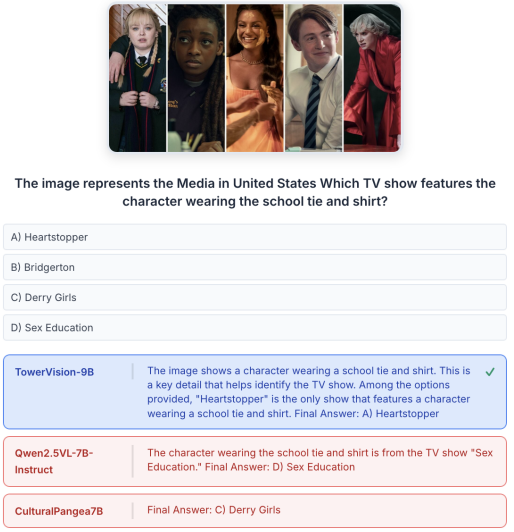

Figure 4: Popular culture grounding; Task: Identify the TV series featuring the character wearing a school tie and shirt. TowerVision correctly identifies *Heartstopper* using fine-grained visual and stylistic cues, while Qwen2.5-VL and CulturalPangea default to culturally adjacent but incorrect series. This illustrates TowerVision's ability to resolve pop-cultural ambiguities through contextual grounding.

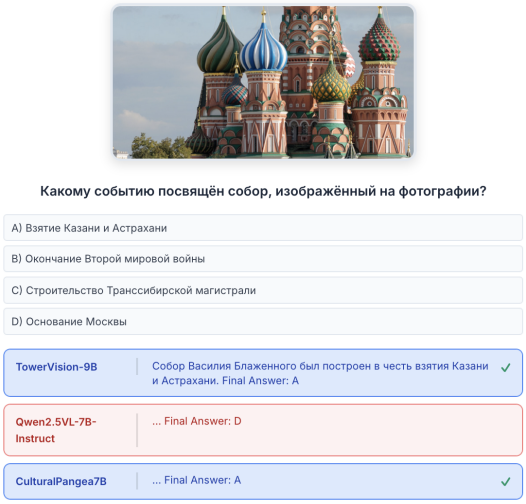

Figure 5: Russian historical context; Task: Identify the historical event commemorated by the cathedral. TowerVision correctly associates St. Basil's Cathedral with the conquest of Kazan and Astrakhan, whereas Qwen2.5-VL incorrectly attributes it to the founding of Moscow

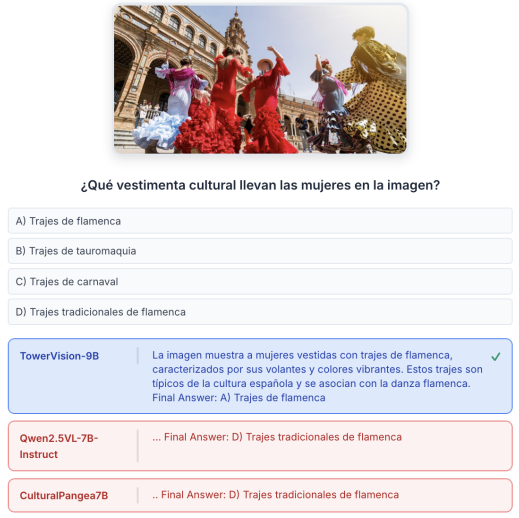

Figure 6: Spanish cultural context; Task: Identify the cultural attire worn by the women in the image. TowerVision correctly identifies *trajes de flamenca*, while baselines select semantically similar but less contextually appropriate alternatives. This highlights TowerVision's sensitivity to culturally terminology.

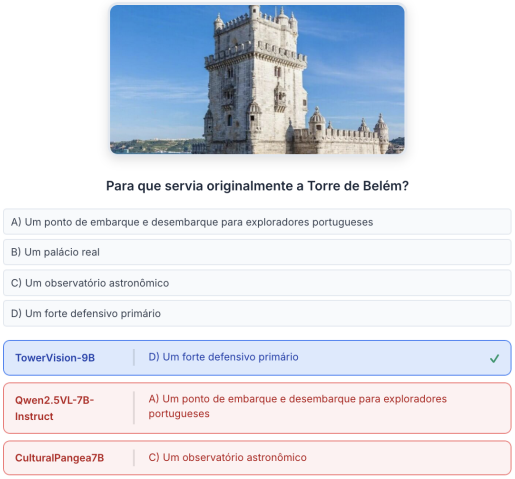

Figure 7: Portuguese historical context; Task: Identify the original purpose of the Torre de Belém. TowerVision correctly identifies the monument as a primary defensive fortification, while baselines rely on stereotypical narratives related to exploration. This demonstrates TowerVision's culturally grounded understanding of national heritage.

```
    "Décrivez cette image.",
    "Que pouvez-vous voir sur cette photo?",
    "Dites-moi ce qu'il y a dans cette image.",
    "Expliquez ce que cette image montre.",
    "Légendez cette image.",
    "Que se passe-t-il sur cette photo?",
    "Fournissez une description de cette image."
]

# Dutch prompts
NL_PROMPTS = [
    "Beschrijf deze afbeelding.",
    "Wat zie je op deze foto?",
    "Vertel me wat er op deze afbeelding staat.",
    "Leg uit wat deze afbeelding laat zien.",
    "Onderschrift deze afbeelding.",
    "Wat gebeurt er op deze foto?",
    "Geef een beschrijving van deze afbeelding."
]

# German prompts
DE_PROMPTS = [
    "Beschreiben Sie dieses Bild.",
    "Was können Sie auf diesem Foto sehen?",
    "Sagen Sie mir, was auf diesem Bild zu sehen ist.",
    "Erklären Sie, was dieses Bild zeigt.",
    "Beschriften Sie dieses Bild.",
    "Was passiert auf diesem Foto?",
    "Geben Sie eine Beschreibung dieses Bildes."
]

# Spanish prompts
ES_PROMPTS = [
    "Describe esta imagen.",
    "¿Qué puedes ver en esta foto?",
    "Dime qué hay en esta imagen.",
    "Explica qué muestra esta imagen.",
    "Pon un título a esta imagen.",
```

```
    "¿Qué está pasando en esta foto?",
    "Proporciona una descripción de esta imagen."
]

# Italian prompts
IT_PROMPTS = [
    "Descrivi questa immagine.",
    "Cosa puoi vedere in questa foto?",
    "Dimmi cosa c'è in questa immagine.",
    "Spiega cosa mostra questa immagine.",
    "Dai un titolo a questa immagine.",
    "Cosa sta succedendo in questa foto?",
    "Fornisci una descrizione di questa immagine."
]



# Korean prompts
KO_PROMPTS = [
    "이 이미지를 설명해주세요.",
    "이 사진에서 무엇을 볼 수 있나요?",
    "이 이미지에 무엇이 있는지 알려주세요.",
    "이 이미지가 보여주는 것을 설명해주세요.",
    "이 이미지에 캡션을 달아주세요.",
    "이 사진에서 무슨 일이 일어나고 있나요?",
    "이 이미지에 대한 설명을 제공해주세요."
]


# Chinese prompts
ZH_PROMPTS = [
    "描述这张图片。",
    "你能在这张照片中看到什么？",
    "告诉我这张图片里有什么。",
    "解释这张图片展示了什么。",
    "为这张图片添加说明。",
    "这张照片中发生了什么？",
    "提供这张图片的描述。"
]
```

A.7.2 Gemini 2.5 System Prompts

We generate synthetic captions using the Gemini 2.5 API with a diverse set of system prompts. These prompts are designed to produce varied response formats, including direct answers, caption-plus-answer pairs, and structured final-answer formats.

```
# Direct answer formats
    "Answer the question concisely.",
    "Provide a brief, direct answer to the question.",
    "Keep your response short and to the point.",
    "Give a concise answer based on what you see in the image.",
    "Answer directly based on the visual information.",
    "Respond with a short, clear answer to the question.",
    "Be brief and direct in your response."

# Simple caption + answer formats
    "First provide a caption of what you see, then give your answer.",
    "Write a brief caption describing the image, followed by your answer to the question.",
    "Start with a description of the image, then provide your answer clearly marked as 'Answer:'.",
    "First write 'Caption: <brief image description>' then answer the question.",
    "Begin with 'Caption: [what you see in the image]' followed by your response to the question.",
    "Start by writing 'CAPTION: {description}' before answering the question."
```

```
# Final Answer formats
    "End your response with 'Final Answer: <your answer>'.",
    "Conclude with 'Final Answer: <your answer>'.",
    "After looking at the image, provide 'Final Answer: <your answer>'.",
    "Your response should end with 'Final Answer: <your answer>'.",
    "First describe what you see, then provide 'Final Answer: <your answer>'.",
    "Always end your response with 'Final Answer: <your answer>' after analyzing the image.",
    "Provide a concise answer. End with 'Final Answer: <your answer>'."

# Naive formats (simple, direct)
    "Describe the image and answer the question.",
    "Begin by describing the image and then answer the question.",
    "Provide a brief description of the image and then answer the question.",
    "Answer the question in a helpful and informative manner.",
    "Start by describing the image and then answer the question.",
    "You are a helpful assistant. Describe the image and answer the question."

# Simple formatted caption/answer pairs
    "Caption: <description> → Answer: <response>",
    "Image shows: <description> | My answer: <response>",
    "[CAPTION] <description> [ANSWER] <response>",
    "# Image: <description>\n# Answer: <response>",
    "First 'Image Description: <what you see>' then 'Answer: <your response>'"

# With specific markers
    "<description><answer>",
    "Image: <description> → Answer: <conclusion>",
    "<IMAGE> describe what you see </IMAGE> <ANSWER> provide your response </ANSWER>"
"Begin with '{IMAGE DESCRIPTION}' and end with '{FINAL ANSWER}'."
```

These prompts are used to generate high-quality captions that improve instruction-following and visual description diversity.

A.8 GPT-4o Evaluation Prompts

```
    system_prompt = (
    "You are an evaluation assistant tasked with assessing the correctness of "
    "AI-generated answers to question-answer pairs.\n\n"
    "Your goal is to compare a predicted answer against the ground-truth answer "
    "and judge its correctness.\n\n"
    "Evaluation guidelines:\n"
    "• Correctness focus: Determine whether the predicted answer is semantically "
    "consistent with the ground-truth answer.\n"
    "• Partial correctness: Answers that capture the core meaning of the ground-truth "
    "should be considered partially correct, even if minor details are missing, "
    "unless those details are explicitly required by the question.\n"
    "• Scoring: Assign an integer score between 0 and 5, where 0 indicates a fully "
    "incorrect answer, 5 indicates a fully correct answer, and intermediate values "
    "reflect varying degrees of partial correctness."
)

user_prompt = (
    f"Please evaluate the following video-based question-answer pair:\n\n"
    f"Question: {question}\n"
    f"Ground-truth answer: {answer}\n"
    f"Predicted answer: {pred}\n\n"
    "Indicate whether the predicted answer is correct or incorrect and assign a score "
    "between 0 (fully incorrect) and 5 (fully correct), where intermediate scores "
    "represent partial correctness.\n\n"
    "Response format:\n"
    "Return your evaluation as a Python dictionary string with the following keys:\n"
```

```
    "• 'pred': either \"correct\" or \"incorrect\".\n"
    "• 'score': an integer between 0 and 5.\n"
    "• 'reason': a brief justification for the assigned score.\n\n"
    "Only output the Python dictionary string.\n\n"
    "Example:\n"
    '{"pred": "correct", "score": 4, "reason": "The predicted answer captures the main intent of
)
```