TowerVision: Understanding and Improving Multilinguality in Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Despite significant advances in vision-language models (VLMs), most existing work follows an English-centric design process, limiting their effectiveness in multilingual settings. In this work, we provide a comprehensive empirical study analyzing the impact of several multilingual design choices, such as training data composition, encoder selection, and text backbones. The result is TowerVision, a family of open multilingual VLMs for both image-text and video-text tasks, built upon the multilingual text-only model Tower+. TowerVision achieves competitive performance on multiple multilingual benchmarks and shows particular strength in culturally grounded tasks and multimodal translation. By incorporating visual and cultural context during fine-tuning, our models surpass existing approaches trained on substantially larger datasets, as demonstrated on ALM-Bench and Multi30K (image tasks) and ViMUL-Bench (video tasks). Alongside the models, we release VisionBlocks, a high-quality, curated vision-language dataset. Our findings highlight that multilingual vision-language training data substantially improves cross-lingual generalization—both from high-resource to underrepresented languages and vice versa—and that instruction-tuned LLMs are not always the optimal initialization point. To support further research, we publicly release all models, data, and training recipes.

1 Introduction

The success and widespread adoption of large language models (LLMs) has naturally led to a surge of interest in adding multimodal capabilities to these models. In particular, the visual modality has recently received considerable attention, with recent releases of *frontier* vision-language models (VLMs) (Deitke et al., 2024; OpenAI et al., 2024; Comanici et al., 2025; Team et al., 2025; Bai et al., 2025b). However, despite impressive progress, the development of VLMs has been mostly built upon English-centric language models, and trained with English vision-text data, giving little consideration to performance in most other languages. A key challenge in multilingualization of 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. As such, a critical challenge remains: What are the best strategies to effectively extend these models to support multiple languages beyond English?

An effective strategy for VLM multilingualization is to let large-scale text-only multilingual data carry most of the burden. This can be achieved by continuing pretraining of the text backbone on multilingual corpora and by including multilingual content in the text-only portion of the VLM fine-tuning mixture—thereby reducing reliance on scarce multilingual multimodal data. A recent example of this approach is Pangea (Yue et al., 2025), which introduced multilinguality exclusively during the VLM fine-tuning stage using a mixture of data that combined multilingual vision-text pairs generated through synthetic data creation and machine translation of English instructions. While this strategy proved effective, it leaves open key questions: At which stages and on which modules should multilingualization be applied? Which design decisions yield the greatest impact? And how can visual grounding further enhance cross-lingual generalization?

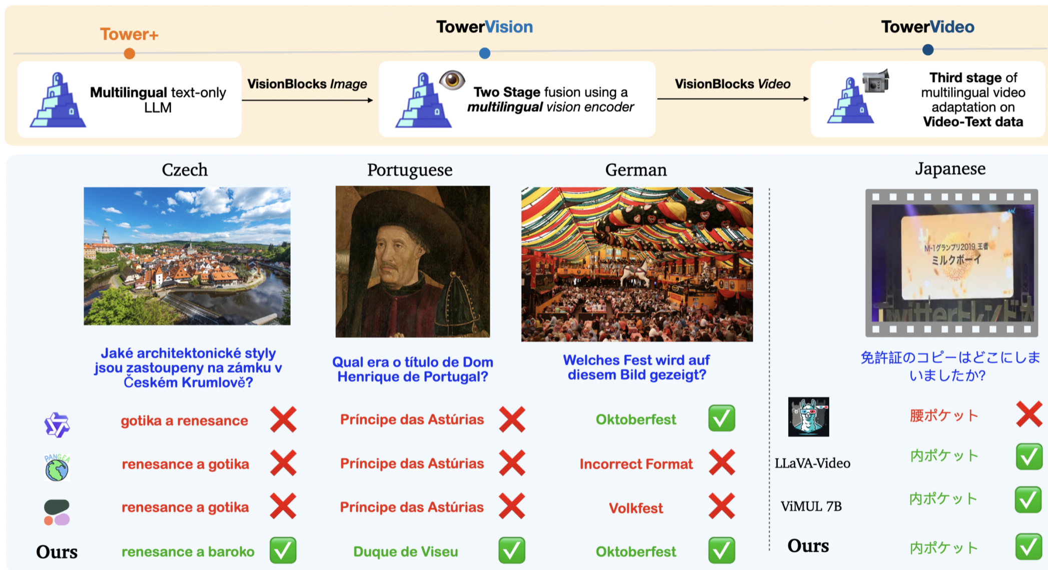

Figure 1: We present TowerVision and TowerVideo, open VLMs with enhanced cultural understanding and multilingual capabilities over leading open multimodal systems on image and video.

In this work, we introduce TowerVision,[1] a suite of open-source multilingual VLMs built on top of Tower+ models (Rei et al., 2025) for 20 languages and dialects.[2] To train TowerVision, we systematically address the challenges outlined above through comprehensive ablation studies, component-level analysis, and cross-lingual evaluation of a multilingualization recipe. Specifically, we investigate how to enhance the multilingual capabilities of VLMs from two axes: first, by exploring the impact of the underlying components (including the alignment projector, vision encoder and text-only LLM); and second, by creating better, more multilingual vision-text datasets and exploring the impact of using this data across different VLM training stages. Overall, compared to strong VLMs of similar size, TowerVision exhibits competitive or superior performance on various multilingual and multimodal benchmarks, as well as cross-lingual transfer capabilities.

In addition to image-based VLMs, we also train a separate multilingual video model, TowerVideo, built on top of TowerVision, thereby extending our analysis to the video modality. TowerVideo achieves competitive performance on ViMUL-Bench (Shafique et al., 2025), a culturally-diverse multilingual video benchmark. Taken together, these contributions provide a comprehensive and systematic study of how to best integrate multilinguality into VLMs across modalities, architectural components, and training stages. Complementing the TowerVision family, we also release VisionBlocks, a curated dataset that consolidates and filters existing vision/video-language resources, further enriched with quality-controlled translations of English textual descriptions into 20 languages and dialects.

2 TowerVision

Our approach follows a multi-stage process encompassing three key components, illustrated in Figure 1: (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 connector/adapter module that transforms these visual features to generate representations compatible with the text embedding space. These

[1] `https://huggingface.co/XXX`

[2] English, German, Dutch, Spanish (Latin America), French, Portuguese (Portugal), Portuguese (Brazilian), Ukrainian, Hindi, Chinese (Simplified), Chinese (Traditional), Russian, Czech, Korean, Japanese, Italian, Polish, Romanian, Norwegian (Nynorsk) and Norwegian (Bokmål)

modules can be selectively trained or kept frozen during different stages of development (Li et al., 2025). Although this training recipe and variations thereof are well-established and have produced several high-quality models (e.g., LLaVA (Liu et al., 2023b), Intern-VL (Chen et al., 2024), NVLM (Dai et al., 2024), Qwen2.5-VL (Bai et al., 2025b), Molmo (Deitke et al., 2024)), most of these fall short in capturing multilingual and culturally diverse nuances. We therefore introduce our multilingual adaptation, TowerVision—we first describe our carefully curated multilingual vision-text data, VisionBlocks (§2.1), and then describe the overall architecture along with an empirically derived recipe, supported by controlled ablations on data allocation, pretraining stages, and initialization strategies(§2.2). (§2.2).

2.1 VisionBlocks: Towards Better Multilingual Vision-Text Data

Creating a large-scale, high-quality, multilingual multimodal dataset for training visual language models to be helpful assistants is non-trivial for a series of intertwined reasons:

- *Human-written* vision-text data featuring user-model interactions (common in text-only alignment) is severely limited. While abundant data exists from large-scale captioning datasets (e.g., LAION-5B; Schuhmann et al. 2022), such sources over prioritize scale over quality which is not ideal for training VLMs with advanced capabilities (Dong et al., 2025; Zhou et al., 2023) like instruction-following, helpfulness, and safety.
- High-quality *multilingual* vision-text data is scarce; furthermore, the lack of open, high-quality multilingual VLMs makes controlled synthetic data challenging or restricted to closed models with limited usage licenses. The most viable alternative, also employed by Pangea (Yue et al., 2025), involves translating English vision-text interactions into target languages.
- Filtering techniques such as reward model scoring or LLM-as-judge approaches (Gu et al., 2025) are significantly more challenging to implement for vision-text data, where even state-of-the-art VLMs (both open and proprietary) struggle to provide reliable preferences (Li et al., 2024).

With this in mind, we develop and release VisionBlocks (Figure 2), which aggregates and filters data from multiple sources, enhanced with new translated and synthetic data, as described below.

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 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. The samples from PangeaIns are originally found in LAIONMulti (Schuhmann et al., 2022) that undergoes a series of automatic steps (using Gemini 1.5 Pro (Gemini Team et al., 2024)) including curating high-quality English instructions, carefully translating them to multiple languages, and adapting them for culturally-relevant multilingual contexts. CulturalGround uses a data curation pipeline that gathers culturally relevant entities from the Wikidata knowledge base, creates several questions and answers about each entity, rephrases them using an LLM, and filters low-quality samples using a VLM. In our work, we rely exclusively on CulturalGround’s filtered subsets to ensure maximum quality.

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 model (Alves 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.5.1). We also augment the dataset with synthetic captions generated by the Gemini 2.5 API. For each image, we sample multiple system prompts to elicit diverse and detailed descriptions (see §A.5.2). This augmentation is intended to improve coverage

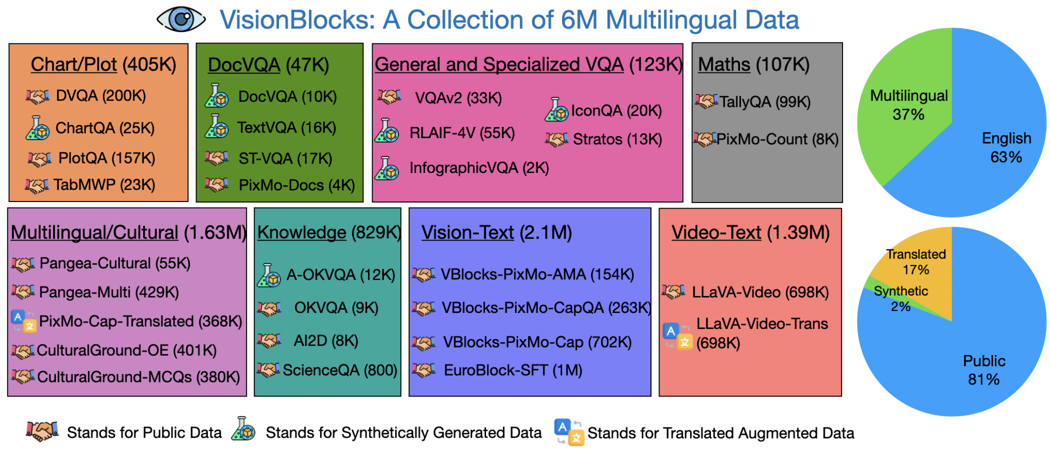

Figure 2: Overview of the VisionBlocks dataset. Synthetic data are generated with Gemini 2.5 API, while translated augmented ones use Tower (Alves et al., 2024). See Table 8, §A.1 for details.

of fine-grained visual details (e.g., spatial relations, attributes, and contextual cues) that human-authored captions often omit, and provides instruction-like supervision, aligning our model more closely with recent VLM training paradigms that leverage synthetic data to boost generalization and response quality. Similar strategies have been shown to be effective in scaling up instruction-following capabilities of VLMs such as LLaVA (Liu et al., 2023a) and InstructBLIP (Dai et al., 2023). We complete our image-text dataset by incorporating the text-only EuroBlocks set, a curated multilingual collection of high-quality synthetic data from the EuroLLM (Martins et al., 2025) synthetic post-training data. EuroBlocks provides diverse, instruction-aligned text that enriches our dataset with robust multilingual coverage and fine-grained, high-quality descriptions.

Translated Multilingual Video Data As video-text data, we employ the LLaVA-Video-178k dataset (Zhang et al., 2025c), which contains captions alongside open-ended and multiple-choice questions in English. To make the dataset multilingual, we retain a randomly sampled half of the conversations in English, and we translate the remaining half uniformly into all supported languages using Tower+9B (Rei et al., 2025), thereby ensuring balanced cross-lingual coverage.

2.2 TowerVision: Architecture & Training Details

One way to improve the multilinguality of LLMs (e.g., improving cross-lingual understanding or extending multilingual support for other languages) is to start from a strong pretrained model and continue pretraining on carefully curated data, with subsequent post-training (Xu et al., 2024; 2025; Alves et al., 2024). TowerVision follows a similar principle, starting from a strong multilingual Gemma-based backbone Tower+ 2B/9B (Rei et al., 2025), which achieves strong multilingual general-purpose performance by leveraging a curated high-quality multilingual dataset and a training recipe designed to preserve general capabilities. As shown in §4, starting from this multilingual backbone substantially improves cross-lingual performance compared to starting from Gemma indicating that strong multilingual priors tend to outperform general reasoning models.

For the vision encoder, TowerVision is initialized with the recently proposed SigLIP2-so400m/14@384px (Tschannen et al., 2025), a vision transformer operating at 384 × 384 resolution that extracts image patch representations and produces multilingually-aligned embeddings of size 729. SigLIP2 is trained on a more diverse data mixture compared to alternatives such as CLIP-ViT (Radford et al., 2021), Perception Encoder (Bolya et al., 2025), or SigLIP1 (Zhai et al., 2023), and thereby yields better multilingual understanding, as we shall see in §4. To align the vision and text modalities, we use a LLaVA-based architecture

(Liu et al., 2023b), 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:

1. A *projector pretraining* phase, where we train the model to predict captions given images on the PIXMO-Cap dataset, freezing both the vision encoder and the language model backbone (so 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).
2. A *vision finetuning* phase, where we unfreeze the full model and train it on the full VISIONBLOCKS dataset (§2.1), excluding the video-text data. 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 six tiles, which provides the best trade-off (§A.3). This phase leads to the TOWERVISION model.
3. 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. Unlike the previous stage, we omit tiling for efficiency. This phase leads to the TOWERVIDEO model.

The models were trained on a custom fork of the LLaVA-Next (Liu et al., 2024a) codebase.[3]

3 Evaluation & Main Results

We evaluate TOWERVISION and TOWERVIDEO on a comprehensive suite of benchmarks spanning multiple modalities and task types (single-image, few-image, and video) across diverse languages, both within and beyond our training set. In this section, we focus on vision-language tasks (i.e., single-image or few image), which including multilingual visual/video question answering, cultural understanding, OCR-related tasks, and visual-language understanding, as well as multilingual video-language tasks. 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

Vision-language tasks We report results on ALM-Bench (Vayani et al., 2024), a cultural understanding multilingual[4] visual QA benchmark, OCRBench (Liu et al., 2024b) and cc-OCR (Yang et al., 2024) for English and multilingual[5] OCR-centric capabilities respectively, and TextVQA (Singh et al., 2019), assessing scientific understanding. Within cc-OCR, we report results on the multilingual text reading subset, as our primary focus is to evaluate the model's multilingual text recognition capabilities.

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.

Culturally-aware multilingual video tasks We use ViMUL-Bench (Shafique et al., 2025), a multilingual video QA benchmark spanning 14 languages: Arabic (ar), Bengali

[3]The code will be released upon acceptance.

[4]German, Spanish, French, Italian, Korean, Dutch, Russian, English, Portuguese, Chinese (Simplified and Traditional), Icelandic, Czech, Ukrainian, Hindi, Japanese, Polish, Swedish, Hungarian, Romanian, Danish, Norwegian (Nynorsk), and Finnish.

[5]German, French, Italian, Russian, Spanish, Korean, Portuguese.

Table 1: **Vision-Language Model Performance.** Comparison of English and multilingual VLMs across multiple benchmarks. Reported values correspond to final accuracy (↑). Bold indicates the best score per column. TowerVision results are highlighted.

	English (↑)		**Multilingual (↑)**		
	TextVQA	OCRBench	CC-OCR	ALM-Bench (en)	ALM-Bench (multi)
Qwen2.5-VL-3B-Instruct	77.8	78.7	76.4	81.0	76.2
Qwen2.5-VL-7B-Instruct	**82.5**	**84.5**	**78.6**	83.1	83.6
Gemma3-4B-it	65.2	74.2	69.1	79.7	80.0
Gemma3-12B-it	73.2	74.7	73.8	83.5	84.5
CulturalPangea7B	69.8	63.5	51.7	61.3	65.2
Llama3-Llava-Next-8B	64.8	54.4	40.9	76.5	73.4
Aya-Vision-8B	66.9	61.0	46.3	78.2	77.3
TowerVision-2B	68.1	58.6	46.1	77.1	81.1
TowerVision-2B-OCR	69.1	63.5	55.5	76.1	77.1
TowerVision-9B	73.6	69.7	56.3	83.6	**85.2**
TowerVision-9B-OCR	76.2	72.7	65.1	**86.1**	84.8

(bn), Chinese (zh), English (en), French (fr), German (de), Hindi (hi), Japanese (ja), Russian (ru), Sinhala (si), Spanish (es), Swedish (sv), Tamil (ta), and Urdu (ur). The dataset contains both open-ended and multiple-choice questions covering culturally diverse domains such as festivals, customs, food, and heritage. Unlike prior datasets, ViMUL-Bench enables comprehensive evaluation of video-language models across both high- and low-resource languages, promoting inclusive and culturally aware research.

3.2 Baselines

For evaluation, we leverage the lmms-eval framework (Zhang et al., 2025b), which enables a systematic comparison of TowerVision against leading open VLMs. We include several multilingual multimodal models, such as *CulturalPangea-7B* (Yue et al., 2025), designed to address gaps in multilingual cultural understanding, and *Aya-Vision-8B* (Singh et al., 2024), optimized for a broad range of vision-language tasks. In addition, we evaluate models from the *Gemma3-Instruct* (*Gemma3-4B-it*, *Gemma3-12B-it*; Team et al. 2025) and the *Qwen2.5-VL-Instruct* families (*Qwen2.5-VL-3B-Instruct*, *Qwen2.5-VL-7B-Instruct*; Qwen et al. 2025), both of which have demonstrated strong performance across a variety of multimodal benchmarks. Finally, we report results for a LLaVA-based model, *Llava-Next-7B* (Liu et al., 2024a), a general-purpose VLM with strong performance across a wide range of tasks. The exact 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 (Zhang et al., 2025a), LLaVA-Video-7B (Zhang et al., 2025c)—also trained on LLaVA-Video-178k—and ViMUL-7B (Shafique et al., 2025), a multilingual video model.

3.3 Main Results

Tables 1–2 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 are strong 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. 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, suggesting that it is still performant in less seen and unseen languages within its training data. We further assess the cross-lingual generalization capabilities of TowerVision in §4.

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

	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-2B-OCR	90.1	97.5	94.7	70.0	77.3	74.2	76.9
TowerVision-9B	**95.1**	98.1	95.6	72.0	**78.8**	75.6	77.4
TowerVision-9B-OCR	94.5	**98.1**	**95.6**	**72.2**	78.3	75.6	77.3

Table 3: **Multilingual video performance per language.** Accuracy (%) on ViMUL-Bench across 14 languages averaged across multiple-choice and open-ended questions. Underlined values mark the best score within TowerVision/TowerVideo variants; **bold** indicates the best overall. Unsupported languages are marked with *.

Model	**ar**	**bn***	**zh**	**en**	**fr**	**de**	**hi**	**ja**	**ru**	**si***	**es**	**sv**	**ta***	**ur***
ViMUL-7B	41.5	35.4	37.0	48.6	48.3	43.9	**39.2**	37.8	45.7	21.2	44.3	41.4	23.3	**36.8**
LLaVA-Video-7B	38.8	30.4	43.2	**53.3**	**49.2**	45.4	34.2	33.4	38.2	18.1	45.7	39.8	21.9	33.8
VideoLLaMA3-7B	**45.6**	**36.6**	**48.0**	52.9	47.1	43.8	37.5	39.4	44.8	**25.1**	45.4	38.5	22.8	32.1
TowerVision-2B	18.9	19.5	21.7	34.2	28.9	28.3	25.1	22.2	24.8	16.3	30.4	27.1	16.1	19.9
TowerVideo-2B	23.0	18.9	35.9	45.2	39.6	39.7	37.2	34.1	38.0	17.1	37.4	38.0	17.7	18.7
TowerVision-9B	34.2	25.4	35.3	46.7	41.1	40.8	34.2	28.1	40.3	19.8	40.5	39.6	21.6	26.4
TowerVideo-9B	38.6	22.1	44.8	51.9	49.1	**47.1**	32.2	**42.3**	40.9	20.8	**46.0**	**44.8**	**24.1**	19.5

TowerVision is less competitive on OCR-related tasks. We hypothesize this is likely due to the limited amount of OCR-focused data in VisionBlocks compared against other models. Since we primarily pretrained TowerVision on large-scale image-caption datasets emphasizing natural images and language alignment, it struggles with scanned text or OCR-heavy scenarios. Despite these limitations, TowerVision does obtain superior performance compared against Aya Vision 8B and LLaVA Next 8B, the former of which has seen significant amounts of OCR-specific data (Singh et al., 2024).

TowerVision-2B is competitive multilingually with larger models. 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. For instance, on Multi30K, TowerVision-2B obtains superior scores to Qwen2.5VL 7B and CulturalPangea 7B. Similarly, on the multilingual split of ALM-Bench, TowerVision 2B is competitive with Qwen2.5VL 7B and outperforms Aya Vision 8B. These results further highlight the efficacy of TowerVision's multilinguality and design choices. We also note that scaling from 2B to 9B parameters consistently improves performance across all benchmarks, suggesting that our training recipe scales well.

Multilingual fine-tuning improves cross-lingual performance in TowerVideo. In Table 3, we report averages across multiple-choice accuracy and open-ended responses, which are automatically judged using GPT-4o (OpenAI et al., 2024), with the same evaluation prompt as Shafique et al. (2025). We compare our TowerVideo models, including the 9B variant, to strong open-source baselines. Our multilingual models are competitive across several languages despite using smaller datasets and fewer frames (for instance, VideoLLaMA3

Table 4: **Impact of backbone and instruction tuning across different benchmarks.**

Backbone Model	English (↑)		Multilingual (↑)		
	TextVQA	OCRBench	CC-OCR	ALM-Bench (en)	ALM-Bench (multi)
GEMMA2-pt-2B	69.2	61.2	45.3	74.3	76.7
TOWER+pt-2B	**70.3**	62.1	**46.3**	73.0	78.2
GEMMA2-it-2B	70.0	**63.0**	45.9	75.0	75.1
TOWER+it-2B	68.1	58.6	46.1	**77.1**	**81.1**
GEMMA2-pt-9B	72.4	66.6	49.6	79.9	79.6
TOWER+pt-9B	73.2	64.5	54.5	81.3	84.4
GEMMA2-it-9B	**74.4**	67.2	49.5	79.6	81.5
TOWER+it-9B	73.6	**69.7**	**56.3**	**83.6**	**85.2**

uses 180 frames). Specifically, ViMUL was trained with separate copies of the dataset for each language, whereas our approach uses a single copy with half in English and the other half uniformly translated into the supported languages. Overall, these results highlight the effectiveness of video-based multilingual fine-tuning in improving cross-lingual reasoning.

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 culturally-sensitive training data, a more 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?

Following the main results of TOWERVISION, we delve deeper into its design choices.

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., 2023b; Bai et al., 2025a), but whose effect on the final model remains unclear. To study these effects, we train TOWERVISION at 2B and 9B scales using three backbones: GEMMA2-pt (pretrained, not instruction-tuned), TOWER+pt (pretrained TOWER+, not instruction-tuned), and TOWER+it (instruction-tuned TOWER+), following the recipe in §2. As shown in Table 4, using TOWER+ consistently outperforms GEMMA2, confirming the importance of a 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 5).

Without additional multilingual data, SigLIP2 models consistently outperform SigLIP1, showing clear benefits in low data regimes, where training data is scarce. With multilingual fine-tuning, however, the gap narrows, showing that finetuning with sufficient multilingual data can compensate for a weaker encoder. At 9B scale, both converge to strong perfor-

Table 5: Multilingual impact of different vision encoders measure 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	84.4
SigLIP2-(En+Multi)	**77.1**	**81.1**	**83.6**	**85.2**

mance. In short, multilingual-aware encoders provide an advantage when data is scarce, but extensive multilingual training can close the gap.

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.

Table 6: Effect of using multilingual versus English-only captions during projector pretraining on ALM-Bench. Results indicate low to no gains from adding multilingual data at this stage.

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

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 §2.1, 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 6, 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.

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. Results in Figure 3 (more details in §A.4) show that broader language coverage consistently improves performance, with larger gains at the 2B scale. Zero-shot improvements for unsupported languages further support cross-lingual transfer when related languages are included. These findings highlight the value of expanding multilingual data, particularly for smaller models.

How does multilingual data affect video fine-tuning? To assess the impact of our multilingual data (see § 2.1) during video fine-tuning, we present results in Table 7 for two baselines: (i) the original TowerVision-2B model and (ii) TowerVideo-2B trained on the full English-only LLaVA-Video-178k dataset. Fine-tuning with video substantially improves the performance of TowerVision models compared to image-text-only variants, highlighting the importance of temporal information for video-language understanding. Incorporating multilingual data further enhances cross-lingual generalization, while English performance remains largely stable, indicating that adding multiple languages does not com-

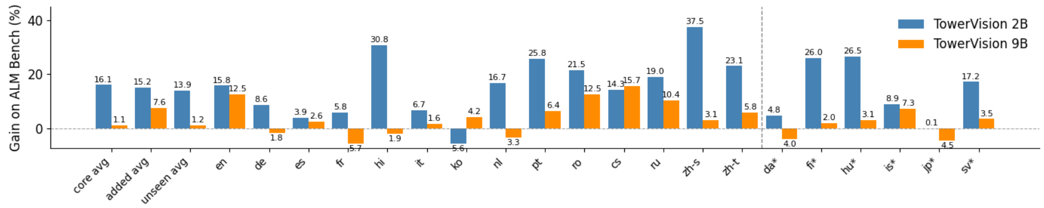

Figure 3: Performance of TowerVision models on 10 vs 20 languages/dialects at 2B and 9B scales. The bars indicate the accuracy gains by training on 20 (all) versus 10 (core) languages.

Table 7: Accuracy (%) on ViMUL-Bench across 14 languages averaged across multiple-choice and open-ended questions. Underlined values mark the best score within TowerVision/TowerVideo variants; **bold** indicates the best overall. Unsupported languages are marked with *.

Model	ar	bn*	zh	en	fr	de	hi	ja	ru	si*	es	sv	ta*	ur*
TowerVision-2B	18.9	**19.5**	21.7	34.2	28.9	28.3	25.1	22.2	24.8	16.3	30.4	27.1	16.1	**19.9**
TowerVideo-2B (english only)	**25.7**	17.8	26.7	**45.5**	**42.3**	34.8	27.8	27.7	34.4	**17.9**	**37.8**	34.0	**18.3**	19.7
TowerVideo-2B (multilingual)	23.0	18.9	**35.9**	45.2	39.6	**39.7**	**37.2**	**34.1**	**38.0**	17.1	37.4	**38.0**	17.7	18.7

promise primary-language capabilities, even though the multilingual models are trained on substantially less English data.

5 Conclusion

We introduced TowerVision, a suite of 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 research on culturally diverse multilingual multimodal language models, and accelerate progress toward narrowing the performance gap with English-centric settings.

6 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.

7 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[6], checkpoints of the TowerVision collection models[7], and the corresponding codebases for training and evaluation[8], to facilitate replication of our results.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=EHPns3hVkj.

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, 2025a. 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*, 2025b.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms, 2024. URL https://arxiv.org/abs/2409.11402.

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.

[6]Links will be made available upon acceptance.
[7]Links will be made available upon acceptance.
[8]Links will be made available upon acceptance.

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/`.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben

Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie

Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya

Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

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.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025. URL https://arxiv.org/abs/2501.02189.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. 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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

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, 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.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike

Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=csbf1p8xUq.

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.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

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, 2025b.

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*, 2025c. 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 8: 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 8 shows the full details and statistics of the VisionBlocks dataset.

A.2 Models Checkpoints

Table 9 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 9: **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 test 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. Results are reported in Table 10, highlighting the trade-offs between resolution, patch granularity, and style diversity.

A.4 Cross-Lingual Generalization

A.5 System Prompts

A.5.1 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.

Table 10: **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 11: 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 2. *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

```
# 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."
]

# 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 = [
    "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.5.2 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.