

# A Large-Scale Instruction-Tuning Dataset and Models for Slovenian Vision-Language Tasks

Matej Martinc<sup>(1)</sup> and Domen Vreš<sup>(2)</sup>

<sup>(1)</sup>Jožef Stefan Institute; <sup>(2)</sup>Faculty of Computer and Information Science, University of Ljubljana  
Ljubljana, Slovenia  
matej.martinc@ijs.si, domen.vres@fri.uni-lj.si

## Abstract

Vision-language models (VLMs) represent a significant leap forward in artificial intelligence, yet their development has been predominantly focused on English, creating a digital divide for speakers of less-resourced languages. This paper addresses this gap by introducing the first large-scale, general instruction-tuning dataset for the less-resourced Slovenian language. Comprising over one million text-image pairs, the dataset was constructed through a multi-pronged approach: automatic curation from Slovenian news media and Wikipedia, and machine translation of the English LLaVA-665k dataset. To demonstrate the dataset's efficacy, we fine-tuned two pre-trained, multilingual Gemma-3 models (4B and 12B parameters) on this new resource. Our evaluation, conducted on a new manually curated test set, reveals that the fine-tuned models named SVILA (Slovenian Vision Language Assistant) exhibit substantial performance gains on a variety of vision question answering, visual grounding, and optical character recognition tasks when compared to their baseline counterparts. This establishes our methodology as an effective blueprint for enhancing VLM capabilities in other less-resourced languages. The dataset is publicly available in the Slovenian language resource repository CLARIN.SI (<http://hdl.handle.net/11356/2050>) and both fine-tuned models are published on the Hugging Face platform (<https://huggingface.co/GaMS-Beta/SVILA-1-12B> and <https://huggingface.co/GaMS-Beta/SVILA-1-4B>).

**Keywords:** Vision-language model, Instruction-tuning dataset, Slovenian

## 1. Introduction

Vision-language models (VLMs) are a class of artificial intelligence systems designed to integrate and process information from both visual and textual modalities (Li et al., 2025). By combining principles from computer vision and natural language processing, VLMs can perform complex, cross-modal tasks that were previously impossible for single-modality models (Islam et al., 2023). This capability to bridge the gap between visual and linguistic information represents a significant advancement. Unlike traditional computer vision systems that perform narrow tasks like object classification, VLMs can engage in flexible, human-like reasoning about the world. This shift enables a wide range of new applications, from accessibility tools that describe images for the visually impaired to sophisticated chatbots that can understand and respond to visual cues (Wang et al., 2024). This versatility makes VLM development a critical area of research for advancing equitable access to technology and enhancing digital literacy.

Despite the transformative potential of VLMs, their development is heavily biased toward a handful of well-resourced languages, primarily English. This bias stems from their dependence on vast, diverse datasets, which are often non-existent for languages with smaller speaker populations. This creates a significant “digital divide” where speakers of less-resourced languages are excluded from

the benefits of modern AI (Holdt et al., 2025). In the European context, this disparity is particularly relevant. The European Union (EU) recognizes 24 official languages, but many, including Slovenian, are considered less-resourced. Slovenian, a South Slavic language with approximately 2.5 million speakers, faces a double data scarcity: a limited supply of high-quality text data and an even more pronounced lack of paired image-text data. The absence of these foundational resources presents a fundamental challenge, as training models on small datasets can lead to poor generalization, low accuracy, and a lack of robustness. The development of language resources for Slovenian VLM is therefore a crucial effort to ensure linguistic diversity and equitable access to advanced AI capabilities within the EU.

To address the existing digital divide in Slovenian language technologies, this study presents a three-fold contribution to the development of VLMs for the Slovenian language:

**A large, publicly available instruction-tuning dataset in a dialog format:** We have created the first instruction-tuning dataset for Slovenian VLM training, which is structured in a conversational format. This extensive dataset contains over 1 million text-image pair examples. The data was obtained through the automatic scraping of content from Slovenian news media portals and



**Q: Kaj je na sliki?**

*En transl.: What is on the image?*

**A: Na sliki je Tromostovje v Ljubljani. V ozadju je Prešernov trg. Vidim tudi reko Ljubljanico. V ospredju so kamnite ograje mostu.**

*En transl.: The picture shows Tromostovje in Ljubljana. In the background is Prešeren square. I can also see the river Ljubljanica. In the foreground are the stone railings of the bridge.*



**Q: Katero mesto je to?**

*En transl.: Which city is this?*

**A: To je staro mestno jedro Maribora.**

*En transl.: This is the old city center of Maribor.*

Figure 1: Qualitative examples generated by the SVILA-12b model.

Slovenian Wikipedia, and was further augmented by the machine translation of the English VLM instruction-tuning dataset, LLaVA-665k (Liu et al., 2023, 2024). The dataset is publicly available to foster further research and development.

**A manually curated test set:** To ensure robust evaluation, we have developed a small, manually checked test dataset. This benchmark is crucial for accurately assessing the performance of Slovenian VLMs on a variety of tasks.

**Two open-source Slovenian VLMs:** We have developed and publicly released two VLMs for

Slovenian, with 4 billion and 12 billion parameters, which are available on Hugging Face. These models were created by fine-tuning pre-trained, multilingual Gemma-3 backbone models (Kamath et al., 2025) on our newly created instruction-tuning dataset. We named them SVILA (Slovenian Vision Language Assistant). The code for training the models is available at <https://github.com/matejMartinc/SVILA>. Figure 1 shows examples generated by the SVILA-12b model.

To showcase the effectiveness of our contributions, the models were evaluated on the newly created manually checked test dataset. The results demonstrate substantial improvements over the baseline models on a wide range of vision grounding, vision question answering (VQA) and optical character recognition (OCR) tasks, confirming the value of the new dataset and the fine-tuned models.

## 2. Related Work

### 2.1. VLMs

The architectural landscape of VLMs can be broadly categorized into several paradigms. One of the foundational approaches is the “dual-encoder architecture”, exemplified by models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). These models employ separate encoders for visual and text data, and project both into a shared, high-dimensional embedding space. The core training mechanism, known as contrastive learning, works by bringing the embeddings of matching image-text pairs closer together while pushing apart the embeddings of mismatched pairs. This enables models to learn a robust understanding of the relationship between visual concepts and natural language descriptions. This is in contrast to the paradigm known as “fusion models”, which merges visual and textual information earlier in the processing pipeline. The BLIP (Bootstrapping Language-Image Pre-training) framework is a prime example (Li et al., 2022).

Recently, due to the prohibitive costs of training large-scale multimodal models from scratch, a new architectural paradigm emerged: the “frozen backbone” approach, pioneered by the Flamingo VLM family (Alayrac et al., 2022). This approach leverages powerful, pre-trained, and crucially, frozen vision and language models, and connects them with a small number of learnable, bridging components. Latest VLM models, such as the Qwen-VL series (Wang et al., 2024), also build on pre-trained foundations but employ different integration techniques. Qwen-VL utilizes a pre-trained Vision Transformer (ViT) and connects it to a large language model using a position-aware vision-language adapter. This

adapter helps to effectively align and fuse the visual features with the language model’s embedding space. Qwen-VL is notable for its strong multilingual capabilities and its ability to process multiple image inputs in conversational contexts.

In contrast, the Gemma-3 family of models (Kamath et al., 2025), which is used in this study, represents a more unified architectural design. Gemma-3’s multimodal versions (4B, 12B, and 27B) employ a pre-trained SigLIP vision encoder (Zhai et al., 2023) and the model is designed for joint, end-to-end training. It directly integrates visual information by treating images as a sequence of compact “soft tokens” that are fed into the decoder-only transformer along with text tokens. This allows for a deep fusion of modalities from the earliest layers, distinguishing its methodology from the modular, adapter-based conditioning seen in Flamingo. Nevertheless, same as in the frozen backbone paradigm, the SigLIP vision encoder is frozen during training.

Recently, the trend of adapting VLMs for less-resourced languages has been gaining traction. In a survey of 106 studies covering 75 low-resource languages, Lupascu et al. (2025) identified key strategies for overcoming data scarcity, highlighting data curation, machine translation, and cross-modal transfer as essential techniques. While the studies that would cover the development of VLMs for less-resourced EU languages are still scarce, there is an on-going *EuroVLM* project, which is trying to develop VLMs for all EU languages. The developed VLM is nevertheless at the moment still in a preview stage<sup>1</sup>. Complementing the trend of developing VLMs for low-resource languages is the emergence of small VLMs (Patnaik et al., 2025). This movement in the research community also addresses the challenge of accessibility by focusing on achieving a balance between performance, computational efficiency, and scalability. Techniques such as knowledge distillation and lightweight attention mechanisms are central to their design.

When it comes to the Slovenian language technologies, several LLMs have been recently developed in the scope of the PoVeJMo project, a national program funded by the Slovenian Research and Innovation Agency and the EU. The project has already yielded models such as the GaMS-1B-Chat and other instruction-tuned GaMS models (2B, 9B, and 27B) (Vreš et al., 2024; Holdt et al., 2025). While these models provide valuable insights, they as of yet do not support non-textual input and are still prone to errors, underscoring the ongoing need for more training data. In addition to these local efforts, researchers also leverage multilingual models like Llama (Grattafiori et al., 2024) and Gemma (Kamath et al., 2025), which have shown good per-

formance on Slovenian benchmarks. However, the closed-source proprietary model Gemini-2.5-pro (Comanici et al., 2025) is still reported to be the best on the new Slovene LLM Arena Benchmark<sup>2</sup>.

## 2.2. Slovenian Multimodal Datasets

While some Slovenian multimodal resources are available, none provide a ready-made solution for developing a robust VLM for Slovenian. The WIT (Wikipedia-based Image Text) Dataset is a notable resource, containing over 37.6 million image-text pairs across 108 languages (Srinivasan et al., 2021). The dataset explicitly includes Slovenian, providing at least 12,000 examples, though the exact number of high-quality, well-formed captions is not publicly available in the source material. Another available resource is the GLAMI-1M dataset (Kosar et al., 2022), a high-quality, human-labeled resource that contains 1 million fashion items across 13 languages, including Slovene (Kosar et al., 2022). Unfortunately its narrow focus makes it unsuitable for training a general-purpose VLM. Other well-known datasets like Conceptual Captions are primarily English-centric and lack extensive multilingual coverage (Sharma et al., 2018). It is also important to note that the existing CLARIN.SI multimodal corpus for Slovenian (Mlakar et al., 2020) is not an image-text dataset but rather a collection of audio and video recordings with transcriptions, making it unsuitable for the intended task. As of yet, there is no “off-the-shelf” dataset for Slovene VLM training.

## 3. Dataset Construction

As stated above, the central hurdle in developing advanced AI for languages like Slovenian is the lack of data. Large-scale models depend on vast datasets for training, a resource that is inherently limited for languages with a smaller speaker base. For a VLM to be able to follow instructions, it requires a large corpus of high-quality, paired image-text data, a resource that is even more difficult and expensive to collect than monolingual text. While public multilingual datasets like WIT and GLAMI-1M can serve as a starting point, they are either too sparse or too domain-specific for a general-purpose VLM. Therefore, in this research we prioritized a multi-pronged approach to creating a high-quality, native Slovene image-text dataset.

The constructed SLO-VLM-IT-Dataset 1.0 is a comprehensive resource specifically engineered for the instruction-tuning of VLMs in the Slovenian language. Comprising a total of 1,128,228 diverse examples, this dataset employs a crucial dual strategy:

<sup>1</sup><https://huggingface.co/utter-project/EuroVLM-9B-Preview>

<sup>2</sup>See the current ranking at: <https://arena.cjvt.si/sl/leaderboard>

the inclusion of a large, machine-translated corpus to establish a diverse foundation of instruction-following tasks, combined with a significant volume of culturally-specific Slovenian data to ensure adaptation to the local context and cultural sphere. The dataset aims to address the scarcity of high-quality, large-scale VLM training data in low-resource languages and is designed to significantly enhance a VLM’s capacity for cross-modal understanding, visual grounding, VQA, OCR, and instruction following within a Slovenian linguistic context.

The dataset was constructed from five primary sources (see dataset statistics in Table 1). The foundational component is a machine-translated version of the popular English LLaVA mix665k v1.5 language-image instruction-following data (LLaVA-665k) (Liu et al., 2023, 2024)<sup>3</sup>. This initial component is vital for task diversity, ensuring the model learns a broad spectrum of general instruction-following behaviors. The original dataset, which was generated with GPT-4 (Achiam et al., 2023), was translated to Slovenian by the proprietary Gemini 1.5 Pro model (Georgiev et al., 2024). The composition of the translated LLaVA-665k dataset exhibits substantial task diversity, including roughly 20K examples focusing on VQA based on OCR tokens, 66K examples which demand a concise one-word or one-phrase response, roughly 80K OCR examples targeting explicit text extraction from images, and roughly 80K examples of visual grounding, where instructions prompt the generation of bounding box coordinates from a specified description. The largest translated component is about 350K examples derived from roughly 120K COCO 2017 images, covering tasks such as long description generation and multiple-choice questions. The remaining examples are text-only. Note that while the machine translation model was instructed to translate to Slovene both prompts and instruction, for the OCR examples, only the translated user prompts were used, while the model responses, the actual text extracted from the images, were intentionally left in their original English to train the VLM’s to extract content from images following Slovenian instructions.

The remaining dataset examples introduce uniquely Slovenian content, providing a large corpus of roughly 440K newly generated, culturally and temporally specific examples. This includes 140K instruction-tuning examples derived from a March 14th, 2025 Slovenian Wikipedia database dump, and a total of 300K examples (100K each) sourced from contemporary images scraped from the RTV, Siol, and 24ur news portals between February and March 2025. This body of native Slovenian data is

---

<sup>3</sup>Available at: [https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K/blob/main/llava\\_v1\\_5\\_mix665k.json](https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K/blob/main/llava_v1_5_mix665k.json)

crucial for adapting the VLM to the specific cultural sphere and contemporary topics, going beyond the general-purpose capabilities learned from the translated dataset. All these examples were curated into an instruction-tuning format using the Gemini 1.5 Pro model (see example prompt and the transformation from the Wikipedia article to the dialogue format in Figure 2)<sup>4</sup>. This approach enables a form of knowledge distillation (Patnaik et al., 2025), where a smaller “student” VLM (in our case Gemma-3 models) can be trained to mimic the high-quality, complex instruction-following behavior generated by the large-scale “teacher” model (Gemini 1.5 Pro), thereby significantly reducing the computational resources required for achieving state-of-the-art performance.

To ensure a task-diverse evaluation of the trained VLM, a test set was created<sup>5</sup>. This test set, summarized in Table 2, comprises a total of 3,780 examples, drawn from the initial pool of machine-translated LLaVA-665k examples and not included in the training set. This strategic composition allows for the direct evaluation of the model’s instruction-following capabilities across tasks learned from the translated data, which forms the foundational component of the training corpus. The test set is organized into four distinct files corresponding to the major task types present in the LLaVA-665k data. This includes 995 examples focused on visual grounding, where the model is prompted to extract bounding box coordinates for a specific object in the image; 974 examples for multi-choice VQA, 817 examples for general VQA tasks, such as generating image and object descriptions and performing visual reasoning, and 994 examples for OCR testing. Notably, for OCR, same as in the training set, only the user prompts are in Slovenian, while the model responses are the actual texts mostly extracted from the book covers, intentionally left in their original English to accurately assess the VLM’s ability to extract untranslated text within a Slovenian instructional context.

The construction of the test set included a manual quality control process by two native Slovenian speakers. The primary objective was to transform the raw machine translation output into a gold-standard, locally adapted resource ready for the VLM evaluation. Both reviewers were given an initial set of 4000 machine translated examples (1000 per tasks) and guidelines for manual review and cor-

---

<sup>4</sup>The proprietary Gemini 1.5 Pro model was utilized for the curation of these examples due to its strong ranking on several tasks on the most comprehensive Slovenian natural language processing benchmark, SloBench (<https://slobench.cjvt.si/>) at the time of the dataset creation.

<sup>5</sup>The test set is publicly available at <https://kt-cloud.ijs.si/index.php/s/DTTGFTaJ7xK4se3>.

<p>PROMPT:</p> <p>Generate a dialog between a human and a chat robot (chatbot) helper, where the human asks questions related to the input image and the following topic: "" + document[title] + ""\n""          "The robot helper answers these questions according to the information inside this text: "" + document[content] + ""\n\n"" + ""          ""The text was published on Wikipedia.\n The generated dialog needs to be in Slovenian and should at some point include the description of the image. You can use your background knowledge on the topic to produce longer and more thorough answers, and to generate relevant questions. If the image or input text is about the specific event, you should include the information about the date of the event in the dialog or warn the user that this information was accurate at a specific date. The dialog can be of arbitrary length.</p> <p>Output dialog should be in JSON format.</p> <p>Use this JSON schema:</p> <pre>output = { "dialog": list[{"from": "human", "value": str}, {"from": "chatbot", "value": str}...] }</pre> <p>Return: output""</p>	<p>ORIGINAL ARTICLE:</p>  <p>"title": "»Velikih pet« afriških divjih živali"          "content": "V Afriki predstavlja pet velikih divjih živali lev, leopard, črni nosorog, afriški savanski slon in kafrski bivol.[1] Izraz so skovali lovci na veliko divjad in se nanaša na pet živali, ki jih je v Afriki najtežje loviti peš [2][3], zdaj pa ga pogosto uporabljajo tudi organizatorji safari potovanj.[4] Na izdajah južnoafriških bankovcev randov iz leta 1990 in pozneje, je na vsakem apoenu prikazana druga velika žival. Vsaka od petih velikih je primer karizmatične megafavne, ki ima pomembno vlogo v popularni kulturi in je med najbolj znanimi velikimi živalmi v Afriki. Države, kjer je vse moške najti, so Angola, Bocvana, Zambija, Uganda, Namibija, Republika Južna Afrika, Kenija, Tanzanija, Zimbabve, Demokratična republika Kongo, Ruanda in Malavi."</p> <p>CURATED TRAINING EXAMPLE:</p> <p><b>Q: Kaj so »Velikih pet« afriških divjih živali?</b>          A: V Afriki izraz »Velikih pet« divjih živali označuje leva, leoparda, črnega nosoroga, afriškega savanskega slona in kafrskega bivola. Izraz so prvotno skovali lovci na veliko divjad, saj so to živali, ki jih je bilo najtežje loviti peš. Sedaj pa izraz pogosto uporabljajo tudi organizatorji safari potovanj.</p> <p><b>Q: Mi lahko poveš več o izvoru tega izraza?</b>          A: Seveda. Izraz »Velikih pet« nima nobene zveze z velikostjo živali. Izhaja iz časa lova na veliko divjad in se nanaša na pet afriških živali, ki so veljale za najnevarnejše in najtežje za lovcem. Torej, ne gre za največje ali najtežje, ampak za tiste, ki so predstavljale največji izziv in nevarnost lovcem.</p> <p><b>Q: Katere države v Afriki so najboljše za opazovanje »Velikih pet«?</b>          A: Po podatkih Wikipedije, so države kjer lahko najdete vseh »Velikih pet« Angola, Bocvana, Zambija, Uganda, Namibija, Republika Južna Afrika, Kenija, Tanzanija, Zimbabve, Demokratična republika Kongo, Ruanda in Malavi.</p>
---	---

Figure 2: Example of Wikipedia article data curation together with prompt used.

Sources	Examples	With Image	No Image	Tokens	Avg. Tokens
<b>SLO-VLM-IT-Dataset 1.0</b>					
24ur	100000	100000	0	29864709	298.65
Siol	100000	100000	0	32014376	320.14
Rtv	100000	100000	0	27199709	272.00
Wiki	139433	139433	0	44477416	318.99
LLaVA-665k translated	634197	597719	36478	71381058	112.55
<b>Additional training resources</b>					
LLaVA-665k English	182445	145972	36473	27812198	152.44
GaMS FT dataset	26844	0	26844	2403050	89.52
NEMOTRON FT dataset	97917	0	97917	38360856	391.77
<b>— TOTAL —</b>	<b>1380836</b>	<b>1183124</b>	<b>197712</b>	<b>273513372</b>	<b>198.08</b>

Table 1: Composition of the VLM training dataset, detailing the number of examples, image presence, token counts for each source, and average number of tokens per example. The subsection SLO-VLM-IT-Dataset 1.0 presents the structure of the newly created instruction-tuning dataset for Slovenian, while additional training resources describe the (English and text-only) resources additionally added to the VLM training set to prevent catastrophic forgetting and overfitting of the model.

rection<sup>6</sup>. More specifically, they were instructed to systematically correct all linguistic and grammatical errors specific to Slovenian, resolve complex case, gender, and number agreements, ensure natural word order, and replace awkward literal translations with correct idiomatic expressions. Furthermore, they performed essential content localization, such as converting all imperial units to metric (e.g., feet to meters), adjusting currency to Euros, and adapting cultural references to ensure the instructions were fluent and culturally natural.

The reviewers were also given strict criteria for

content rejection, flagging examples for removal if they were fundamentally untranslatable, such as those relying on word games, puns, or untranslatable metaphors, or if they pertained to specific cultural concepts that could not be reasonably localized without changing the core intent of the query. A specific methodological constraint was maintained for the 1000 OCR examples: to accurately assess the VLM’s ability to extract and ground untranslated text within a Slovenian instructional context, only the Slovenian user prompts were manually corrected and localized. The model responses, which constitute the actual extracted text from the

<sup>6</sup>See Appendix A.

images (e.g., book covers), were strictly preserved in their original English, serving as the intended cross-lingual ground truth.

Dataset File	Examples	With Image	No Image	Tokens	Avg Tokens
General VQA	817	817	0	31001	37.94
Multi-choice VQA	974	974	0	24286	24.93
OCR	994	994	0	26529	26.69
Visual grounding	995	995	0	19720	19.82
— TOTAL —	3780	3780	0	101536	26.86

Table 2: Composition of the test dataset, detailing the number of examples, image presence, and token counts for each source file.

## 4. VLM Training

We employed Supervised Fine-Tuning (SFT) to adapt two open-source VLMs, gemma-3-4b-it and gemma-3-12b-it, for instruction following in the Slovenian language. We selected the Gemma-3 architecture due to its multimodal capabilities and demonstrated strong performance on Slovenian benchmarks<sup>7</sup>. Due to computational constraints, we did not explore larger open-source multimodal models which rank even better on Slovenian benchmarks, namely gemma-3-27b-it and Llama-4-Maverick-17B-128E-Instruct<sup>8</sup>.

As explained above, the fine-tuning process, utilizing the high-quality SLO-VLM-IT-Dataset 1.0, can be considered a form of knowledge distillation. By training smaller Gemma-3 models on the complex instruction-following examples curated by the much larger, high-performing Gemini 1.5 Pro, we enable the student Gemma-3 models to mimic the teacher’s sophisticated behavior. This approach in theory allows us to achieve a high level of performance and fluency in Slovenian while significantly reducing the computational resources required compared to training a large-scale VLM from scratch.

### 4.1. Fine-Tuning Strategy and Data Composition

The SFT procedure involved full-weight training of the language model (LLM) component and the projection layer, while the weights of the vision encoder were kept frozen, aligning with the training methodology established in the original Gemma VLM papers (Kamath et al., 2025).

The final training dataset used was composed to achieve proficiency in Slovenian while mitigating

<sup>7</sup>At the time of writing, gemma-3-12b-it was the best performing multimodal model on the Slovene LLM Arena Benchmark in its size category, i.e., up to 15B parameters.

<sup>8</sup><https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct>

risks such as catastrophic forgetting and overfitting. The core of the dataset is the SLO-VLM-IT-Dataset 1.0, which includes a machine-translated corpus of instruction-following tasks alongside a significant body of natively generated, culturally-specific Slovenian VLM data. The complete training corpus totaled 1,380,836 examples and besides the SLO-VLM-IT-Dataset 1.0 included additional resources to enhance model robustness, detailed in Table 2:

- **LLaVA-665k English:** A portion of the original English instruction-tuning data was included to prevent catastrophic forgetting of general VLM capabilities. The examples were selected randomly from the original dataset.
- **GaMS FT and NEMOTRON FT:** These text-only instruction-tuning datasets were added to prevent the models from overfitting on the multimodal text-image input, a problem we noticed in initial experiments. GaMS FT is a manually curated Slovenian text dataset used for training GaMS-Instruct models, while NEMOTRON FT is a larger corpus composed of approximately 80% machine-translated Slovenian and 20% English text-only instructions used for training the novel GaMS-Instruct-Nemotron series<sup>9</sup>.

### 4.2. Computational Setup and Hyperparameters

All training was conducted on the EuroHPC Leonardo Booster<sup>10</sup>. To manage the substantial memory requirements associated with full-weight fine-tuning of large models, we utilized the DeepSpeed ZeRO-2 optimization technique.

We utilized 8 nodes, each equipped with 4 NVIDIA A100 GPUs, each featuring 64GB of VRAM. The GPUs on a single node are connected through a 600 GB/s NVLINK while the nodes are connected through 2x200 Gb/s Infiniband. The key hyperparameters were set consistently across both models, i.e. we set the learning rate of 1e-5, warmup ratio to 0.05, weight decay to 0.1 and maximum sequence length to 3000 tokens to prevent memory spikes. We train the models for one epoch using the Adam optimizer. The specific training details and computational costs for each model are summarized in Table 3.

## 5. Evaluation

We assess the fine-tuned VLMs on the manually crafted test set described in Section 3. The fine-tuned models are compared to the baseline non-fine-tuned Gemma-3 4b and 12b models, which

<sup>9</sup><https://huggingface.co/cjvt>

<sup>10</sup><https://www.hpc.cineca.it/systems/hardware/leonardo/>

Model	Steps	Nodes	GPUs/Node	Batch Size/Device	Global Batch Size	Duration	Total GPU-Hours
SVILA-4b	5388	8	4	8	256	24 hours	24×4×8 (768)
SVILA-12b	10775	8	4	4	128	48 hours	48×4×8 (1,536)

Table 3: Training details and computational resources for the fine-tuning process.

allows us to directly measure the performance improvement of fine-tuning, and to the Gemini-2.5-pro model, which is at the time of writing the best performing model on the Slovene LLM Arena Benchmark<sup>11</sup>. The test set is organized into four distinct evaluation tasks, and the models’ performance was assessed using task-specific metrics as described below.

### 5.1. Evaluation Criteria

The **Visual Grounding task** requires the VLM to extract bounding box coordinates for a specific object mentioned in the Slovenian prompt. We measure the **Intersection over Union (IoU)** between the predicted bounding box coordinates and the ground-truth coordinates. The model’s raw string output is parsed to extract the coordinates for calculation. Due to a lack of specific formatting instructions in the visual grounding test prompts, the baseline models frequently outputted bounding boxes in the wrong format. To address this, for this task only, the baseline models were evaluated in a one-shot setting. This means they were provided with one example of the correct output format (i.e., a list of coordinates as  $[x1, y1, x2, y2]$  with values normalized between 0 and 1).

The **Multi-Choice VQA** evaluates the model’s ability to select the correct answer from a set of options (A, B, C, or D) based on the image content. Besides **Accuracy** we also report the **F<sub>1</sub> score** (macro weighted) to account for a very slight class imbalance.

The **Optical Character Recognition (OCR)** requires the model to extract text, primarily from book covers, following a Slovenian instruction. Consistent with the training strategy, the user prompts are in Slovenian, while the model responses (the actual extracted text) are intentionally left in their original English. This setup evaluates the VLM’s ability to ground and extract untranslated content from an image based on a Slovenian instruction. We measure the models’ extraction fidelity using the **Character Error Rate (CER)** and **Word Error Rate (WER)** between the predicted and ground-truth text.

The **General VQA** includes complex visual reasoning, description generation, and general VQA tasks. Given the subjective nature of long-form generative responses, we employed an automated evaluation approach. We utilize the **LLM-as-a-**

**Judge** methodology, as proposed in Chiang and Lee (2023), to assess the quality of the generated answers. As a judge, we use Gemini-2.5-pro model, the best ranked multimodal model for Slovenian. We adopted the four criteria from Li et al. (2024) to score the VLM’s responses on a 1-5 scale. The criteria are described in Table 4 and the judge model is prompted to evaluate the model-generated responses according to these criteria based on a given (text and image) prompt and a gold standard response<sup>12</sup>.

### 5.2. Results

The results of the evaluation are presented in Table 5. Overall, they demonstrate the value of the newly curated Slovenian instruction-tuning dataset and the efficacy of our fine-tuning approach. A direct comparison between our fine-tuned SVILA models and their respective Gemma-3 baselines reveals substantial improvements across a range of tasks. In visual grounding benchmarks, both SVILA-4b and SVILA-12b markedly outperform their pre-trained counterparts according to the average IoU. The substantial gains are also observed in the OCR tasks, where the Word Error Rate (WER) for the 12B model plummeted from 0.5501 to 0.4029. This is an indication that the fine-tuning has enabled the model to better follow Slovenian instructions. While the much larger proprietary Gemini-2.5-pro model still outperforms other models in multi-choice Accuracy and F1-score by a large margin, our SVILA-12b model manages to reduce the gap slightly by improving over the baseline Gemma-3-12b-it by 1.5 percentage points.

In the LLM-as-a-judge evaluation, a clear and consistent hierarchy emerges, with SVILA-12b and SVILA-4b consistently receiving higher scores than the baseline Gemma models across all criteria, including visual faithfulness, relevance, and grammatical fluency. This superiority is powerfully summarized in the final win rate, where SVILA-12b was deemed the best model in about 50% of the evaluations, roughly doubling the preference rate of its closest competitor. Gemini-2.5-pro model was excluded from this part of the evaluation to prevent the risk of inherent self-preference bias, since the same model acted as a judge in this setting.

<sup>11</sup><https://arena.cjvt.si/en/leaderboard>

<sup>12</sup>The prompt given to the Judge model is in Appendix B.

Criterion	Description	Score 5 (Excellent)	Score 1 (Very Poor)
<b>Visual Faithfulness (Accuracy)</b>	How factually correct is the response when compared to the visual information and gold standard? Rewards truthfulness and penalizes invented/hallucinated details.	The response is perfectly accurate and contains no factual errors.	The response contains significant factual errors or hallucinations.
<b>Relevance &amp; Completeness</b>	Does the response directly and fully answer the specific question asked, including all critical information from the gold standard without omissions?	Perfectly relevant and addresses all parts of the question completely.	The response is off-topic or fails to answer the core question.
<b>Grammaticality &amp; Fluency</b>	Is the response grammatically correct, well-formed, and does it read like natural, fluent language?	Flawless grammar and natural flow.	Difficult to read due to major grammatical errors.
<b>Conciseness &amp; Clarity</b>	Is the response presented clearly, without unnecessary verbosity, and easy to comprehend?	Clear, to the point, and easy to understand.	Overly wordy, confusing, or poorly structured.

Table 4: LLM-as-a-Judge evaluation criteria for general VQA.

	Gemini-2.5-pro	Gemma-3-4b-it	Gemma-3-12b-it	SVILA-4b	SVILA-12b
<i>Visual Grounding</i>					
<b>Avg. IoU</b>	0.2343	0.0979	0.1252	0.2379	<b>0.3506</b>
<i>Multi-choice VQA</i>					
<b>Accuracy</b>	<b>0.9148</b>	0.6971	0.7906	0.7279	0.8049
<b>F1-Score (Macro)</b>	<b>0.9150</b>	0.6956	0.7904	0.7278	0.8052
<i>OCR</i>					
<b>Character Error Rate (CER)</b>	0.4534	0.4747	0.4561	0.3724	<b>0.3320</b>
<b>Word Error Rate (WER)</b>	0.5183	0.5664	0.5501	0.4672	<b>0.4024</b>
<i>General VQA - LLM as a Judge Evaluation</i>					
<b>Visual Faithfulness</b>	/	2.60	3.60	3.48	<b>3.74</b>
<b>Relevance And Completeness</b>	/	2.72	3.54	3.67	<b>3.95</b>
<b>Grammaticality And Fluency</b>	/	3.41	4.44	4.73	<b>4.82</b>
<b>Conciseness And Clarity</b>	/	3.08	4.00	4.52	<b>4.74</b>
<b>Win Rate</b>	/	5.26%	26.19%	20.20%	<b>50.31%</b>

Table 5: Model performance on several tasks. Best result per measure is bolded. For Avg. Euclidean Distance, CER and WER measures, lower is better, for other measures higher is better. Note that Win Rate percentages sum up to more than 100% due to allowed win ties.

## 6. Conclusion

In this work, we have addressed the critical gap in vision-language resources for the Slovenian language by introducing the first large-scale, general instruction-tuning dataset, SLO-VLM-IT-Dataset 1.0. Through a hybrid strategy of machine translation and native content curation, we successfully compiled over one million image-text pairs, providing a rich and diverse resource for multimodal research. By fine-tuning the open-source Gemma-3 4B and 12B models on this dataset, we created the SVILA series. Our evaluation on a novel, manually curated test set (the first of its kind for Slovenian VLMs) reveals that SVILA models achieve significant performance gains in tasks like OCR and vi-

sual grounding.

This study does not only provides the Slovenian language community with new training and testing language resources, and a high-performing, open-source VLMs, but also validates our methodology as a viable and effective blueprint for bootstrapping VLM capabilities in other less-resourced languages. By strategically combining automated translation, native data scraping, and knowledge distillation from powerful proprietary models, we have presented a resource-efficient pathway to bridge the digital divide and foster more equitable AI development.

As for future work, a critical next step is the development of additional diverse benchmarks that will allow a more comprehensive and generaliz-

able evaluation. More specifically, the next step will include the development of a native Slovenian benchmark which will allow us to quantify how well the SVILA models have learned to process culturally-specific information, and the development of a Slovenian OCR benchmark that will allow us to test the models' OCR capabilities on Slovenian text extraction. Future work will also explore novel training regimes and loss functions for specific tasks, such as visual grounding.

## 7. Limitations

Our study has several limitations. First, the dataset on which the models were tested was translated by Gemini 1.5 Pro and the dataset on which the SVILA models were trained was curated by Gemini 1.5 Pro. On the other hand, in our LLM-as-a-judge evaluation, the judging was performed by Gemini 2.5 Pro. Although this avoids direct self-preference, a potential for implicit architectural bias between models from the same family cannot be entirely dismissed. Furthermore, the LLM-as-a-judge evaluation was conducted without calibrating against human evaluators, as inter-annotator agreement between the judge LLM and human annotators was not measured, which means the results may not fully capture human preferences and evaluation criteria.

Second, our current evaluation dataset is a manually-checked translation from English, which does not measure the models' ability to analyze uniquely Slovenian visual and textual content. Since the fine-tuning data included a large corpus of native Slovenian news media, we cannot yet quantify how well the SVILA models have learned to process this culturally-specific information. Additionally, since the train and test sets are derived from the same data distribution, the evaluation may not fully reflect the model's generalization capabilities to unseen data from different distributions, potentially leading to an overestimation of its performance.

Our OCR benchmark also has limitations. The instructions are in Slovenian, but the text extracted from the images is English. Furthermore, the source OCR examples from the LLaVA-665k dataset primarily consist of book covers. Therefore, the generalizability of our models' OCR capabilities needs to be validated on a broader range of documents, such as text-heavy images and PDF files.

Finally, there is a possibility that our SVILA models have overfitted on stylistic artifacts of the Gemini 1.5 Pro teacher model used for training dataset curation. This could mean that the model's performance gains are partly due to superficially mimicking the teacher's output patterns rather than a

genuine improvement in underlying reasoning and capabilities.

## 8. Ethical Consideration

The development of vision-language models raises some ethical considerations, which could be roughly categorized into three categories, namely the potential for biased outputs, breaches of privacy, and the risk of misuse. In order to address these concerns, we have adopted transparent development practices and responsible handling of data. To avoid potential biased outputs and breaches of privacy, a lot of effort has been put into ethical data generation and preprocessing. We made sure that the training data is diverse and representative. We also plan to proactively detect and mitigate bias in VLMs outputs, contributing to responsible AI development. To address the privacy concerns, we made sure to not collect any personal data protected by the General Data Protection Regulation (GDPR). We minimize the risk of misuse by offering full transparency of algorithmic processes, data sources, decision-making mechanisms, and by making the models and datasets freely available to a larger scientific community for inspection. This will help the potential users to make informed choices, understand how AI interactions affect them, and increase trust and accountability.

## 9. Acknowledgments

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101214398 (ELLIOT) and from the Digital Europe Programme under grant agreement No. 101198470 (LLMs4EU). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors also acknowledge the financial support from the Slovenian Research Agency for research core funding (No. P2-0103) and projects Large Language Models for Digital Humanistics (LLM4DH, No. GC-0002) and Artificial Intelligence for Science (AI4sci, No. GC-0001).

## 10. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florenzia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al.

2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Špela Arhar Holdt, Špela Antloga, Tina Munda, Eva Pori, and Simon Krek. 2025. [From words to action: A national initiative to overcome data scarcity for the Slovene LLM](#). In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 130–136, Tallinn, Estonia. University of Tartu Library, Estonia.
- Ashhadul Islam, Md Rafiul Biswas, Wajdi Zaghouni, Samir Brahim Belhaouari, and Zubair Shah. 2023. Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–5. IEEE.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Vaclav Kosar, Antonín Hoskovec, Milan Šulc, and Radek Bartyzal. 2022. GLAMI-1M: A multilingual image-text fashion dataset. *arXiv preprint arXiv:2211.14451*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Qiaomu Li, Ying Xie, Nina Grundlingh, Varsha Rani Chawan, and Cody Wang. 2024. [Assessing image-captioning models: A novel framework integrating statistical analysis and metric patterns](#). In *Proceedings of the Seventh Workshop on e-Commerce and NLP @ LREC-COLING 2024*, pages 79–87, Torino, Italia. ELRA and ICCL.
- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Marian Lupascu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2025. Large multimodal models for low-resource languages: a survey. *arXiv preprint arXiv:2502.05568*.
- Nitesh Patnaik, Navdeep Nayak, Himani Bansal Agrawal, Moinak Chinmoy Khamaru, Gourav Bal, Saishree Smaranika Panda, Rishi Raj, Vishal Meena, and Kartheek Vadlamani. 2025. Small vision-language models: A survey on compact architectures and techniques. *arXiv preprint arXiv:2503.10665*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. [WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. In *Proceedings of the Language technologies and digital humanities conference 2024*, Ljubljana, Slovenia.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

## 11. Language Resource References

Kosar, Vaclav and Hoskvec, Antonín and Šulc, Milan and Bartyzal, Radek. 2022. *GLAMI-1M: A Multilingual Image-Text Fashion Dataset*. Zenodo. PID <https://doi.org/10.5281/zenodo.73387922>.

Izidor Mlakar, Simona Majhenič, Matej Rojc, and Darinka Verdonik. 2020. [Multimodal corpus EVA 1.0](#). Slovenian language resource repository CLARIN.SI.

## 12. Appendices

### **Appendix A. Guidelines for Manual Review and Correction of the Slovenian VLM Test Set**

#### **I. Introduction and Objective**

The primary goal of this review is to ensure the absolute fluency, grammatical correctness, and cultural naturalness of the machine-translated (MT) Slovenian test set. This process will elevate the dataset quality from a raw MT output to a gold-standard, localized resource for evaluating Vision-Language Models (VLMs).

Your tasks are to:

- Fix all grammatical and syntactic errors.
- Localize content, concepts, and phrasing to be natural for a Slovenian audience.
- Identify and flag examples that are fundamentally untranslatable or unlocalizable, requiring removal.

Please remember that the VLM is being tested on its ability to follow Slovenian instructions; therefore, the quality of the Slovenian text is paramount.

#### **II. General Review Principles (Applies to all 4000 Examples)**

*A. Grammatical and Syntactic Correction:* Review and correct the Slovenian text for the following common MT errors:

- Case, Gender, and Number Agreement (Sklanjanje in Spol/Števílo): Ensure all nouns, adjectives, pronouns, and verbs agree correctly in the Slovenian sentence structure. This is the most critical fix.
- Word Order (Besedni Red): Adjust the sentence structure to reflect natural Slovenian word order, which often differs significantly from English.
- Preposition and Conjunction Usage (Predlogi in Vezniki): Correct inappropriate direct translations of English prepositions (e.g., ‘on’, ‘at’, ‘in’) that do not hold the same meaning or case requirement in Slovenian.
- Idiomatic Expressions (Idiomi): Replace awkward, literal translations of English idioms with their natural Slovenian equivalents.
- Orthography and Punctuation (Pravopis in Ločila): Correct all spelling errors, including consistent use of diacritics (č, š, ž), and ensure correct Slovenian punctuation (e.g., use of commas, quotation marks).

*B. Localization and Cultural Adaptation:* Adapt the text to make it sound like it originated in a Slovenian context:

- **Measurement Units:** Convert all non-metric units (feet, miles, pounds, Fahrenheit, gallons) to their metric equivalents (meters, kilometers, kilograms, Celsius, liters). Example: “5 feet tall” should be localized to approx. “1.5 metra visok”.
- **Currency and Time:** Adjust foreign currencies (USD, GBP) to Euros (€) where contextually appropriate, or leave the original amount and specify “USD” if the original context is crucial. Use the 24-hour clock format where applicable.
- **Cultural References:** If an example refers to an obscure US/UK holiday, celebrity, or historical event that has no common equivalent or recognition in Slovenia, attempt to localize it to a known Slovenian equivalent or a more general concept. If localization is impossible, flag it for removal (see Section C).
- **Tone and Style:** Ensure the instructional and descriptive language is natural, clear, and idiomatic in Slovenian. Avoid overly formal or overly informal language unless specifically required by the prompt’s context.

*C. Rejection Criteria (When to Remove an Example):* You must flag an entire example (Prompt + Response/Answer) for removal if the core meaning is lost, or if a high-quality translation is impossible. Mark these examples clearly with [REMOVE].

- **Untranslatable Linguistic Phenomena:** The example relies heavily on a word game, pun, specific rhyming scheme, or untranslatable metaphor that completely loses its cognitive value or humor when translated.
- **Unlocalizable Content:** The example queries a very specific, culturally embedded concept (e.g., “Describe the mascot for the Milwaukee Bucks in 2018”) that has zero context in the Slovenian language sphere and cannot be reasonably adapted to a Slovenian equivalent without fundamentally changing the query’s intent.
- **Ambiguity:** The original English concept was vague or ambiguous, and the machine translation has amplified this to the point where the VLM’s intended task is unclear, even after grammatical correction.

### III. Task-Specific Instructions

The test set is divided into four distinct files. Apply the General Principles (Section II) while adhering to these task-specific rules:

#### 1. Visual Grounding (1000 Examples)

- **Task Structure:** User Prompt (Slovenian) → Model Response (Bounding Box Coordinates: e.g., ).
- **Focus:** The clarity and specificity of the instruction. The Slovenian prompt must precisely and unambiguously describe the object the VLM is meant to locate.
- **Critical Rule:** DO NOT modify the numerical bounding box coordinates. You are only allowed to modify the Slovenian User Prompt.

#### 2. Multi-Choice Question Answering (1000 Examples)

- **Task Structure:** User Prompt (Slovenian), Choices A-D (Slovenian), Correct Answer (Slovenian).
- **Focus:** Consistency and plausibility.
- **Critical Checks:** Ensure the User Prompt, all answer Choices (A, B, C, D), and the Correct Answer text are grammatically and contextually consistent after correction; Verify that the translation of the “distractor” choices is plausible. If a machine translation makes a distractor option obviously nonsensical (compared to the original English intent), fix it so it remains a viable, albeit incorrect, option.

#### 3. General VQA (1000 Examples):

- **Task Structure:** User Prompt (Slovenian) → Model Response (Slovenian Description/Reasoning).
- **Focus:** Fluency and completeness of both prompt and expected response.
- **Critical Checks:** Ensure the Slovenian Prompt (the question/instruction) is clear; Ensure the Slovenian Model Response (the expected answer/description) is a full, fluent, and comprehensive response to the prompt.

#### 4. Optical Character Recognition (OCR) Testing (1000 Examples)

- **Task Structure:** User Prompt (Slovenian) → Model Response (Extracted Text, Intentionally in English/Original Language).
- **Focus:** Maintaining the cross-lingual test setup.
- **CRITICAL RULES:** You must only fix the Slovenian User Prompt; DO NOT TRANSLATE THE MODEL RESPONSE: The model response, which is the actual extracted text (e.g., text from

a book cover), must remain in its original English (or other source language) to accurately assess the VLM's ability to extract and ground untranslated text within a Slovenian instructional context. Fix only grammatical errors in the English response if they were introduced by the machine translation process (e.g., accidental truncation), but do not translate the content.

#### IV. Documentation and Reporting

You must track your changes and rejections meticulously.

- **Tracking Method:** Use a shared spreadsheet or collaborative environment (e.g., Google Sheets, Excel with change tracking) to record your work.
- **Marking Corrections:** Highlight or use a different font color for every text segment you correct so that changes can be easily reviewed.
- **Marking Rejections:** If an example must be removed (Section II.C), clearly mark the entire row/entry with the tag [REMOVE] and add a brief note explaining the reason (e.g., "Untranslatable pun", "Cannot be localized").

#### Appendix B. The prompt given to the Judge model

```
"""You are an expert evaluator of vision
↳ language models. Your task is to act
↳ as a judge and evaluate the
↳ following model-generated responses
↳ based on a given prompt and a gold
↳ standard (true) response.
```

```
**Prompt:**
{original_prompt}
```

```
**Gold Standard Response:**
{gold_standard}
```

```
**Model-Generated Responses:**
"""
    for model_id, answer in
↳ model_responses.items():
        prompt += f"- {model_id} Answer:
↳ {answer}\n"

    prompt += """
```

Your evaluation must be based on the  
↳ following four criteria, which are  
↳ designed to assess the quality of  
↳ answers about visual content.

Visual Faithfulness (Accuracy): How  
↳ factually correct is the response  
↳ when compared to the gold standard?  
↳ The primary goal is to reward  
↳ answers that are true to the visual  
↳ information and penalize any details  
↳ that are invented, incorrect, or  
↳ "hallucinated." (Score 1-5)  
5 (Excellent): The response is perfectly  
↳ accurate and contains no factual  
↳ errors.  
1 (Very Poor): The response contains  
↳ significant factual errors or  
↳ hallucinations.

Relevance & Completeness: Does the  
↳ response directly and fully answer  
↳ the specific question asked in the  
↳ prompt? It should include all  
↳ critical information found in the  
↳ gold standard without omitting key  
↳ details or adding irrelevant  
↳ information. (Score 1-5)  
5 (Excellent): Perfectly relevant and  
↳ addresses all parts of the question  
↳ completely.  
1 (Very Poor): The response is off-topic  
↳ or fails to answer the core  
↳ question.

Grammaticality & Fluency: Is the  
↳ response grammatically correct,  
↳ well-formed, and easy to understand?  
↳ Does it read like natural, fluent  
↳ language? (Score 1-5)  
5 (Excellent): Flawless grammar and  
↳ natural flow.  
1 (Very Poor): Difficult to read due to  
↳ major grammatical errors.

Conciseness & Clarity: Is the response  
↳ presented clearly and without  
↳ unnecessary verbosity? A good answer  
↳ is straightforward and easy to  
↳ comprehend. (Score 1-5)  
5 (Excellent): Clear, to the point, and  
↳ easy to understand.  
1 (Very Poor): Overly wordy, confusing,  
↳ or poorly structured.

```
**Your Task:**
For each model-generated response,
↳ provide a score from 1 to 5 (where 1
↳ is the worst and 5 is the best) for
↳ each of the four criteria. After
↳ evaluating all the responses,
↳ declare which model's response is
↳ the best overall and provide a brief
↳ justification for your choice.
```

Please provide your evaluation in a JSON  
↳ format like this:

```
{
```

```
"evaluations": [  
  {  
    "model": "Model 1",  
    "scores": {  
      "visual_faithfulness": <score>,  
      "relevance_and_completeness":  
        ↪ <score>,  
      "grammaticality_and_fluency":  
        ↪ <score>,  
      "conciseness_and_clarity":  
        ↪ <score>  
    },  
    "justification": "<brief  
    ↪ justification>"  
  }  
],  
"best_model": "<Model X>",  
"overall_justification": "<brief  
    ↪ overall justification>"  
}  
"""
```