Vision-to-Text:Benchmarking Multimodal LLMs on Extremely Low-Resource Languages

A visually grounded and metric-driven pipeline for building SilkRoad-VL

Shuoshuo Hou[1234], Mieradilijiang Maimaiti[1234]*, Zhexin Li[1234], Jiaxin Wang[1234],
Ahmad Hassan[1234], Kaishaer Jiapaer[1], Nilufar Abdurakhmonova[5], Roza Urinbayeva[5],
Madina Mansurova[6], Shormakova Assem[6], Gulnar Murat[7], Le Wu[8], Wushouer Silamu[1234]

[1] School of Computer Science and Technology, Xinjiang University;
[2] Xinjiang Laboratory of Multi-Language Information Technology;
[3] Xinjiang Multilingual Information Technology Research Center;
[4] Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5] National University of Uzbekistan;
[6] Al-Farabi Kazakh National University;
[7] Kapshagay Bidai Onimderi, Kapshagay, Kazakhstan
[8] Integrated Laboratory for Space, Air, and Ground Systems

* Corresponding author: miradeljan51@xju.edu.cn

CATALOG

Background and Motivation

Research gap

High-resource dominance | Limited multimodal coverage | Low-resource underrepresentation

Challenges

Translationese artifacts | Semantic drift | Weak visual grounding

Our goal

Build SilkRoad-VL for six extremely low-resource languages: ug, kk, ky, tg, uz, and ur.

The key idea is to use visual grounding as a quality constraint during corpus construction.

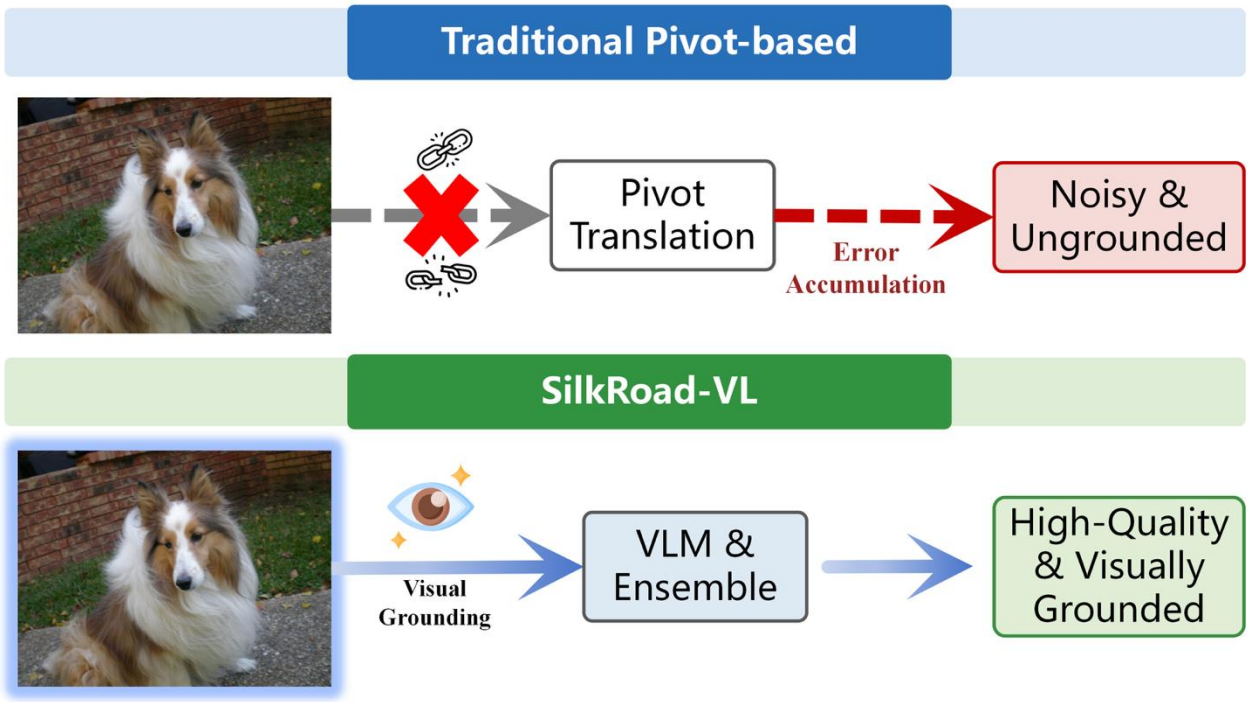


Fig. 1. Traditional pivot-based generation vs. SilkRoad-VL.

Research Goal and Contributions

Goal

Build a high-quality, visually grounded benchmark that improves multimodal translation for extremely low-resource languages.

01 Metric-driven captioning

Select grounded source captions with controllable short/long descriptions.

02 Hybrid ensemble generation

Use complementary MT and LLM systems to create diverse candidates.

03 Tri-QE filtering

Jointly enforce quality, semantic consistency, and visual grounding.

04 Benchmark validation

Verify gains on SilkRoad-VL, Multi30K, and Visual Genome.

Methodology: Overall Pipeline

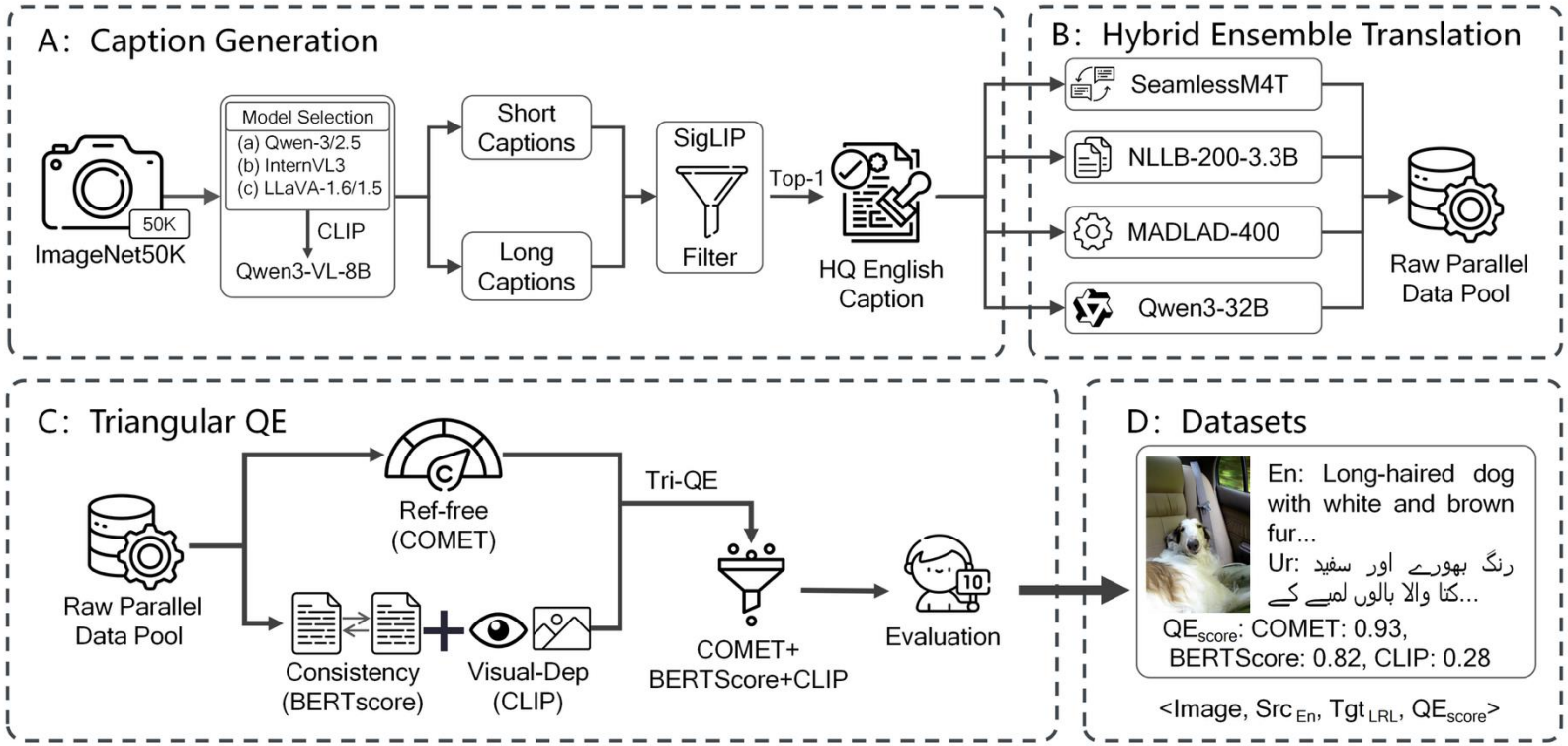


Fig. 2. Overview of the SilkRoad-VL construction pipeline.

Three-stage workflow

1 Caption generation

Select strong visual captions as English anchors.

2 Hybrid ensemble

Generate diverse translations from complementary models.

3 Tri-QE filtering

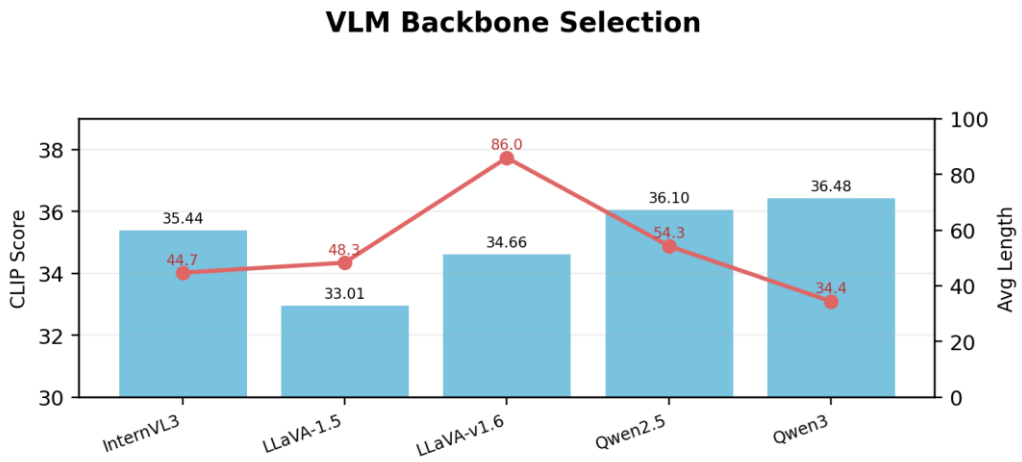
Retain candidates that are high-quality and visually grounded.

Methodology: Captioning and Ensemble Generation

Metric-driven caption generation

VLM Selection: compare candidate VLMs and select Qwen3-VL-8B as the optimal caption backbone.

SigLIP Filtering: generate short and long captions, then retain visually faithful descriptions.

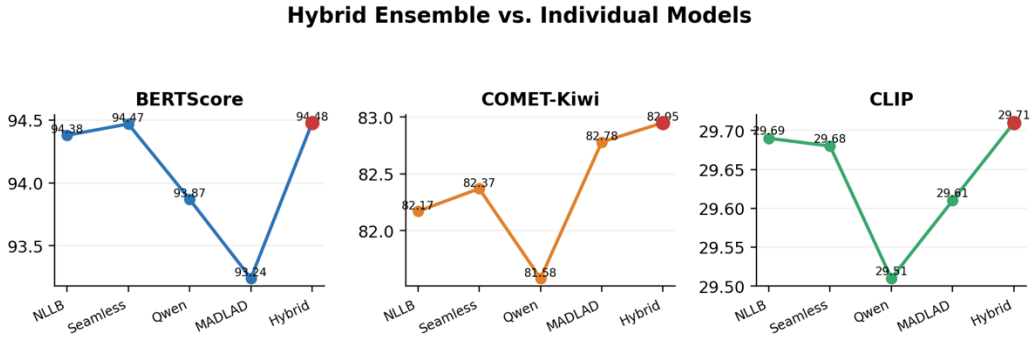


Qwen3-VL balances visual grounding and length controllability.

Hybrid ensemble translation

Diverse Models: NLLB-200, MADLAD-400, SeamlessM4T-v2, and Qwen3-32B generate candidates.

Complementary Strengths: combine lexical fidelity, fluency, and multimodal alignment while reducing single-model bias.



The ensemble-then-filter paradigm produces stronger supervision.

Methodology: Tri-QE Filtering

Triangular Quality Estimation

Quality: COMET-Kiwi

Semantic consistency: BERTScore

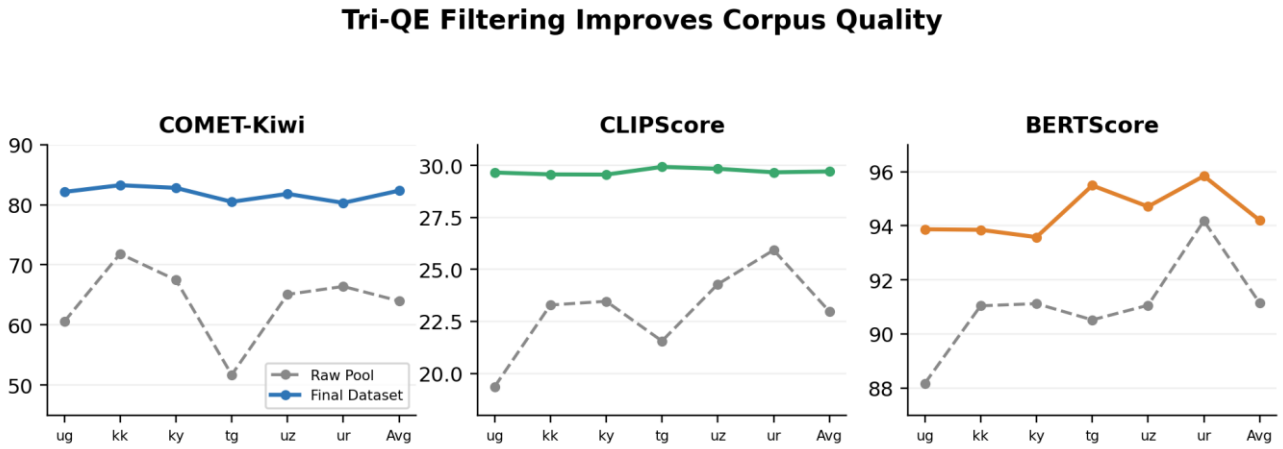
Visual grounding: CLIPScore

Strict gate

Only candidates passing all three constraints are retained.

Thresholds

78.0 / 0.90 / 0.27



Tri-QE removes noisy and hallucinated samples before downstream fine-tuning.

Dataset and Evaluation Setup

SilkRoad-VL corpus

84.9k pairs

82.7k train · 0.4k dev · 1.8k test

Target languages

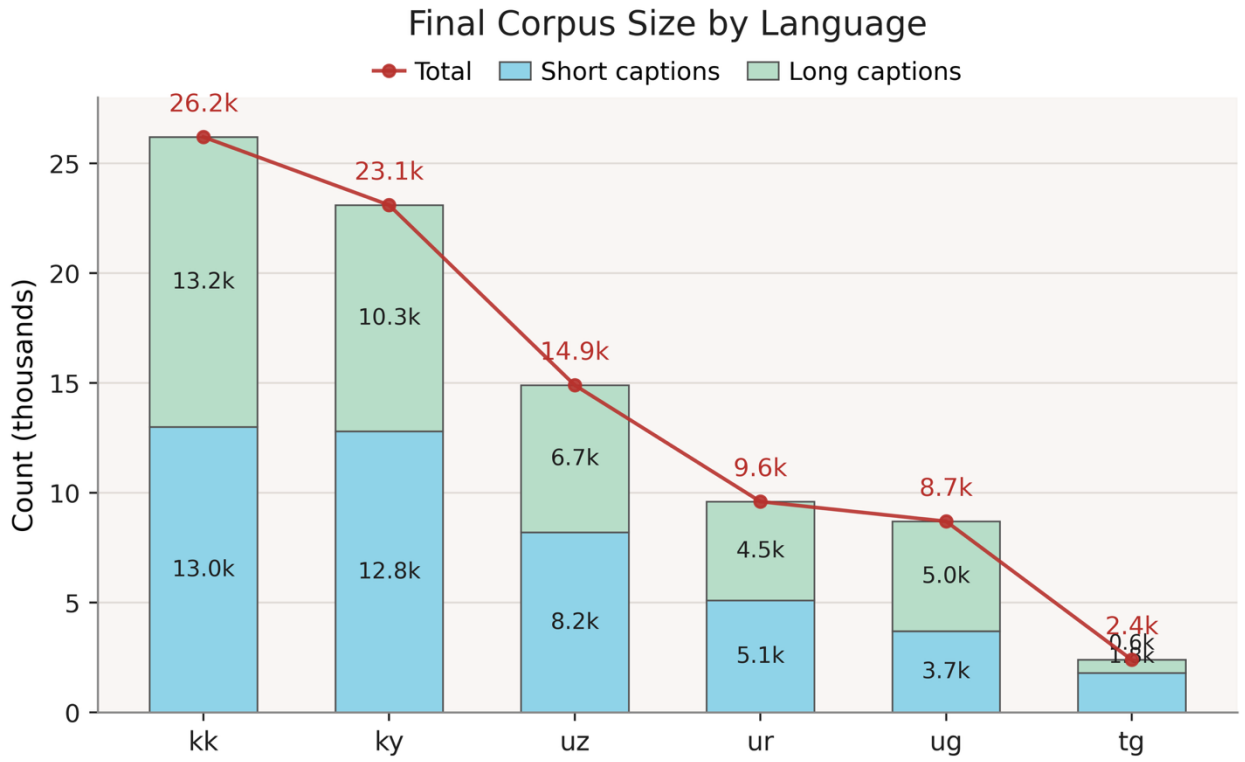
ug · kk · ky · tg · uz · ur

Metrics

COMET-Kiwi: adequacy

BERTScore: semantic consistency

CLIPScore: visual grounding



Language-wise distribution of the final SilkRoad-VL corpus.

Experiments: Main Results on SilkRoad-VL

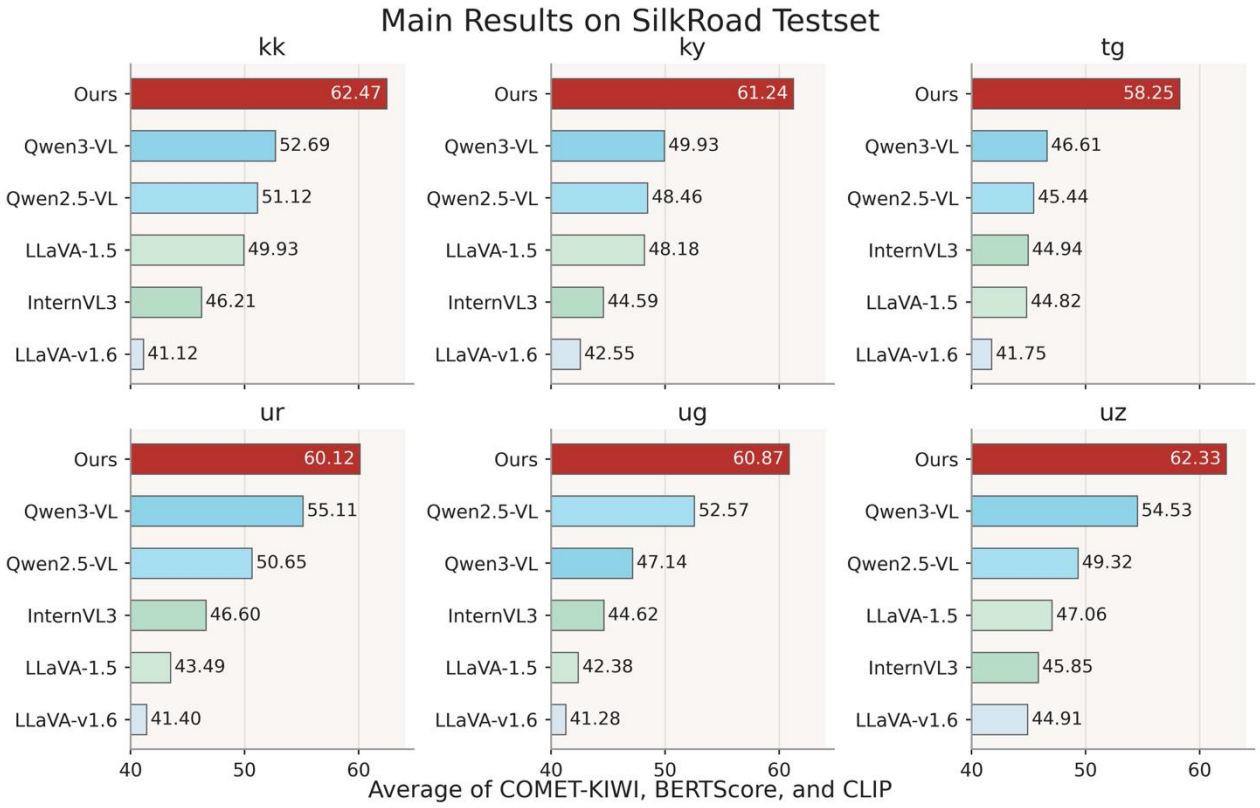


Fig. 3. Main results on the SilkRoad-VL test set across six languages.

Key findings

Ours outperforms all strong baselines on all six languages.

Highest score: 62.47 in Kazakh; 62.33 in Uzbek; 61.24 in Kyrgyz.

Four languages exceed 60 points.

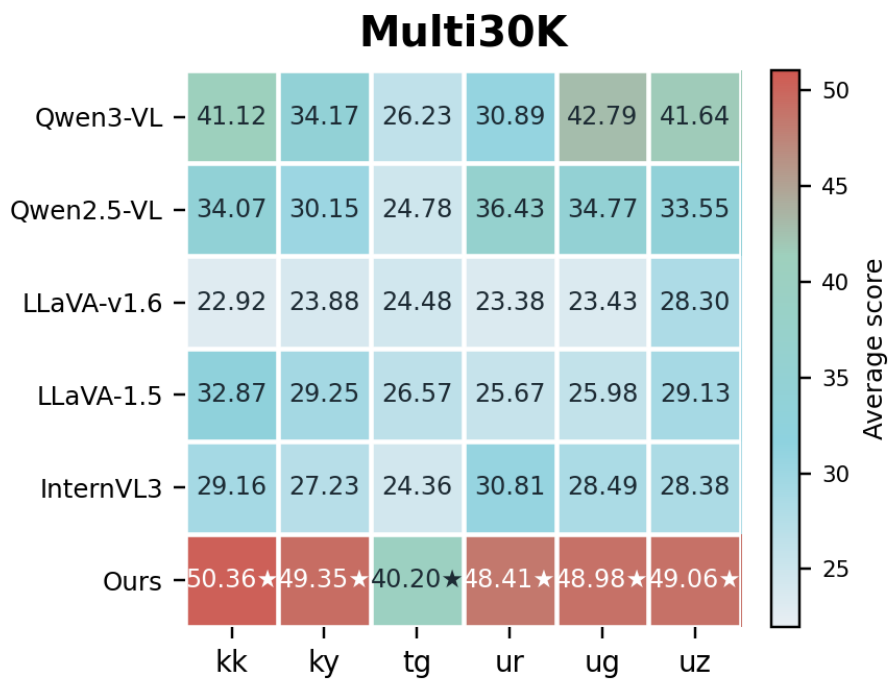
Tajik rises from 46.61 to 58.25 (+11.64).

Takeaway

SilkRoad-VL provides strong and consistent gains in extremely low-resource settings.

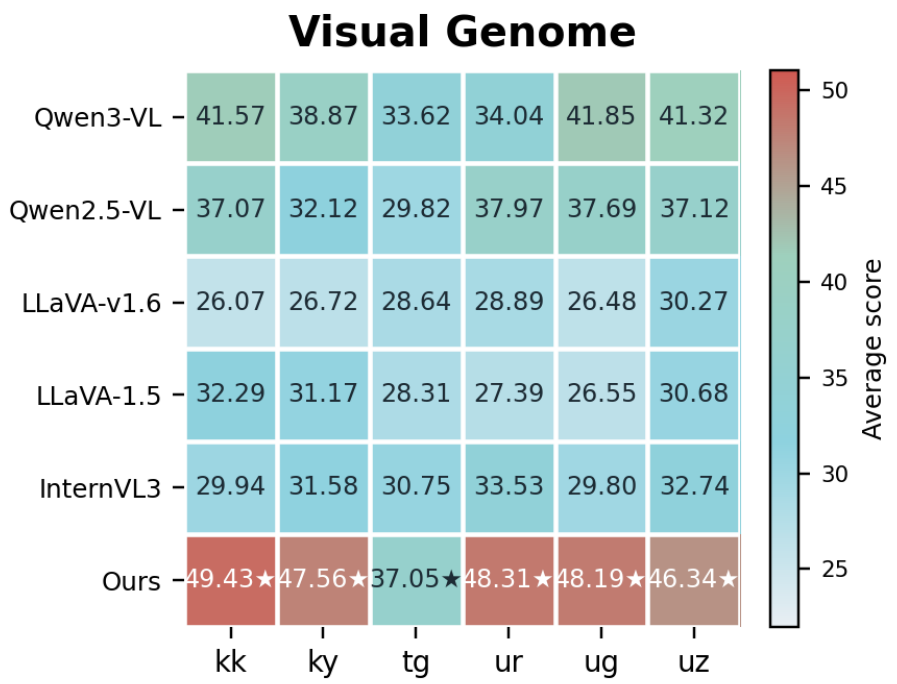
Experiments: Zero-shot Transfer

Multi30K



Tajik: 26.57 → 40 .20; Kazakh reaches 50.36.

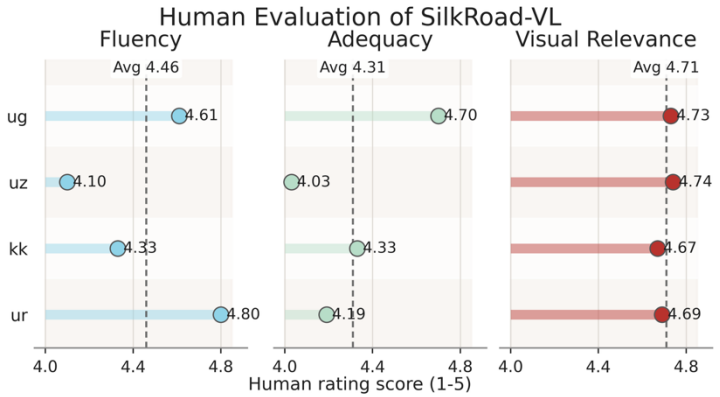
Visual Genome



Urdu: 37.97 → 48.31; Tajik reaches 37.05.

Analysis and Evidence

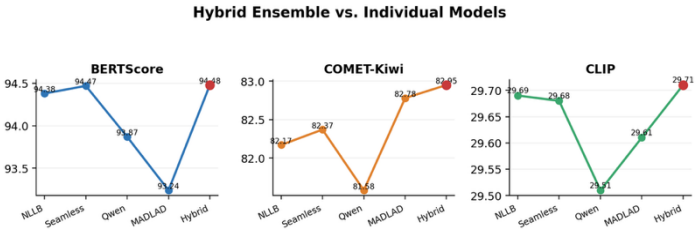
Human evaluation



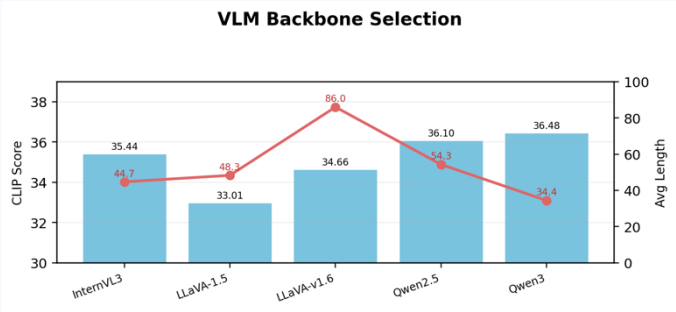
4.46 Fluency · 4.31 Adequacy · 4.71 Visual Relevance

Native-speaker ratings confirm linguistic quality and visual grounding.

Ablation evidence



Hybrid ensemble outperforms single generators across text and visual metrics.



Qwen3-VL gives the best grounding/length trade-off.

Case study

Fine-tuning corrects hallucinations in Kazakh and repetition loops in Urdu.

Conclusion and Q&A

Takeaways

We construct SilkRoad-VL, a high-quality multimodal benchmark for six extremely low-resource languages.

Metric-driven captioning, hybrid ensemble generation, and Tri-QE improve corpus quality.

Fine-tuned models generalize well to Multi30K and Visual Genome.

Future work will extend the pipeline to more low-resource languages and modalities.

Thank you!

Questions are welcome.

Code & Dataset

Corresponding author:
miradeljan51@xju.edu.cn

Supported by NSFC (Grant Nos. 62406316, 201704041014, and 201404041254) and the Xinjiang “Tianchi Talent” Recruitment and Introduction Program.