# Toward Data-driven Skill Identification for General-purpose Vision-language Models

Paper OLIVE

Anthony Meng Huat Tiong*, Junqi Zhao*, Boyang Li,
Junnan Li, Steven C.H. Hoi, Caiming Xiong

{anthonym001, junqi.zhao, boyang.li}@ntu.edu.sg

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · SMU SINGAPORE MANAGEMENT UNIVERSITY · salesforce

## Introduction

Vision-language (VL) models have gained broad competencies that made evaluation difficult. Most existing benchmarks rely on human intuition to categorize evaluation tasks. We propose a data-driven approach that leverages transfer performance and Factor Analysis (FA) to identify latent skills essential for VL tasks. Further, we discover patterns and biases from 2,784 experimental results.

## Key Findings

- Generation tasks exhibit a length bias, where the output length significantly influences transfer performance.
- Factor analysis effectively identifies unexpected yet reasonable factors that explain model performance.
- Datasets requiring reasoning on top of knowledge retrieval improve transfer performance.
- The newly introduced OLIVE dataset exhibits behaviors markedly different from those of other datasets we experimented with.

## Experiments

We finetune four VLMs – BLIP-2, Mini-GPT4, LLaVA, and mPLUG-Owl – across 23 source tasks and evaluate them on 29 target tasks. Together with the model performance before any finetuning (zero-shot), we obtain 2,784 measurements.

### Source and Target Tasks

| Intuitive category | Task | Source | Target |
|---|---|---|---|
| Image captioning | COCO Caption | ✓ | ✓ |
| | Flickr30k | ✓ | ✓ |
| | Web CapFilt | ✓ | ✗ |
| | TextCaps | ✓ | ✓ |
| Generic VQA | VQAv2 | G | G, MC |
| Knowledge-based VQA | OK-VQA | G | G, MC |
| | A-OKVQA | G, MC | G, MC |
| | ScienceQA | MC | MC |
| OCR VQA | TextVQA | G | G, MC |
| | OCR-VQA | G | G, MC |
| Visual reasoning | GQA | G | G, MC |
| | VSR | MC | MC |
| | IconQA | MC | MC |
| | CLEVR | ✗ | G, MC |
| | RAVEN-FAIR | ✗ | MC |

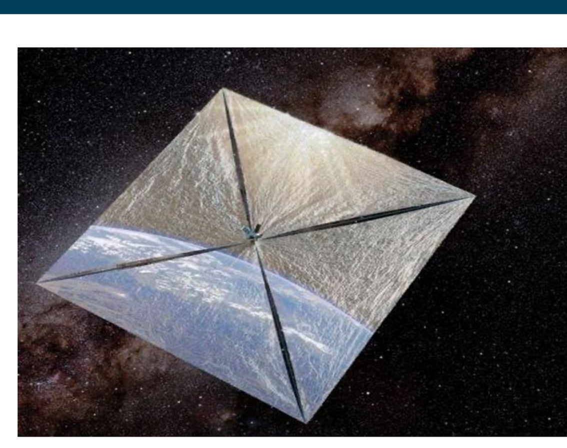
| Intuitive category | Task | Source | Target |
|---|---|---|---|
| Classification | Hateful Memes | MC | MC |
| Humor & sarcasm | New Yorker Ranking | ✗ | ✓ |
| | New Yorker Explanation | ✗ | ✓ |
| | MORE | ✗ | ✓ |
| Chart reading | OpenCQA | G | G |
| | ChartQA | ✗ | G, MC |
| Open-ended generation | OLIVE (Ours) | ✓ | ✓ |
| | LLaVA Conversation | ✓ | ✗ |
| | LLaVA Reasoning | ✓ | ✗ |
| | LLaVA Description | ✓ | ✗ |
| Question generation (QG) | VQAv2 QG | ✓ | ✗ |
| | OK-VQA QG | ✓ | ✗ |
| | A-OKVQA QG | ✓ | ✗ |

Table 1: Tasks overview. Each VQA dataset features two types of tasks: Generative (G), requiring exact matches with ground-truth answers, and Multiple-Choice (MC) with five answer options.

## The OLIVE Dataset

We introduce a diverse multimodal dataset, containing 9,450 images, 30,120 unique instructions, and 47,250 human-edited gold responses.

OLIVE exhibits drastically different behaviors from the other datasets we experimented with as shown by SVD and factor analysis.



**Instruction**: What is the item in the image?

**Output**: The item in the image is a solar sail, which is a device that is designed to harness the energy from sunlight to propel a spacecraft through space without the use of fuel. It is a square shaped piece of cloth that acts like a sail and captures the radiation pressure from the sun to propel the spacecraft forward.

## Results and Discussion

### Top 10 Source Tasks by Performance

| Source task | Harmonic mean |
|---|---|
| A-OKVQA (MC) | 1.3 |
| VQAv2 (G) | 1.3 |
| ScienceQA (MC) | 3.8 |
| A-OKVQA (G) | 4.6 |
| OCR-VQA (G) | 6.0 |
| GQA (G) | 6.2 |
| Flickr30k (G) | 7.2 |
| OK-VQA(G) | 7.8 |
| WebCapFilt (G) | 7.9 |
| IconQA (MC) | 8.4 |

Table 2: Harmonic mean of ranking scores of source tasks across models

### Similarity of Target Tasks

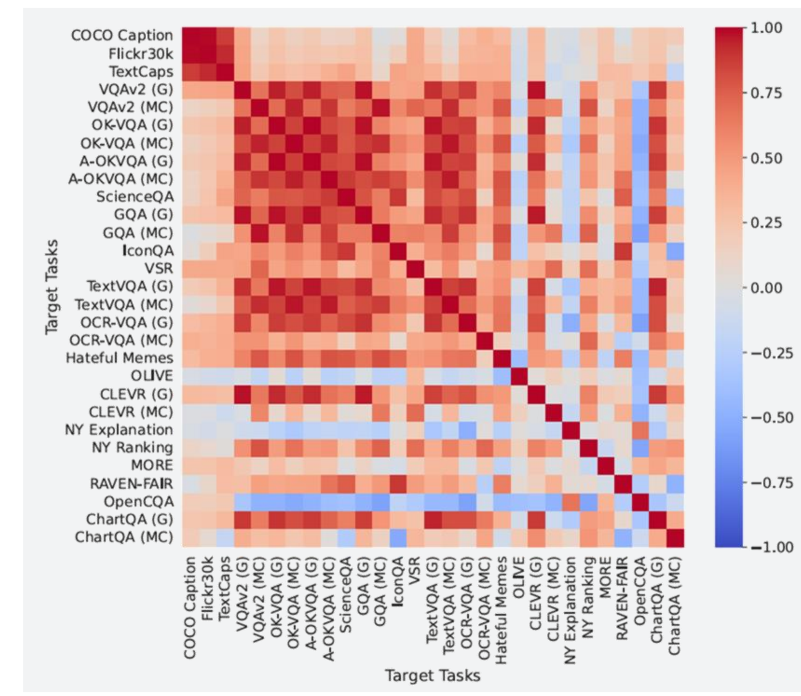


Figure 1: Cosine similarity of target tasks computed using SVD features. OLIVE, with an average similarity of -0.06, ranks as the third least similar to other tasks.

### Output Length Bias

| Source task output length | Target task output length | | |
|---|---|---|---|
| | 1-3 | 6-12 | >40 |
| 1-3 | **-0.03 / 1.00** | -0.78 / 0.79 | -0.85 / 0.44 |
| 6-12 | -0.49 / 0.64 | **-0.43 / 0.75** | **-0.43** / 0.48 |
| >40 | -0.90 / 0.43 | -0.87 / 0.28 | **-0.26 / 0.55** |

Table 3: Mean normalized transfer performance by mean output lengths of source and target tasks. Left (right) values consider all (top 5) source tasks in a group. In-domain source tasks are excluded. A mismatch between output lengths results in significant performance drops.

### Exploratory Factor Analysis

We assume that each source task imparts specific latent skills to a model. These skills, while not directly observable, are reflected in the model's performance on related target tasks. When target tasks tap on similar skills, they tend to exhibit similar performance patterns. To identify these latent factors, we apply Exploratory Factor Analysis (EFA) to the performance data.

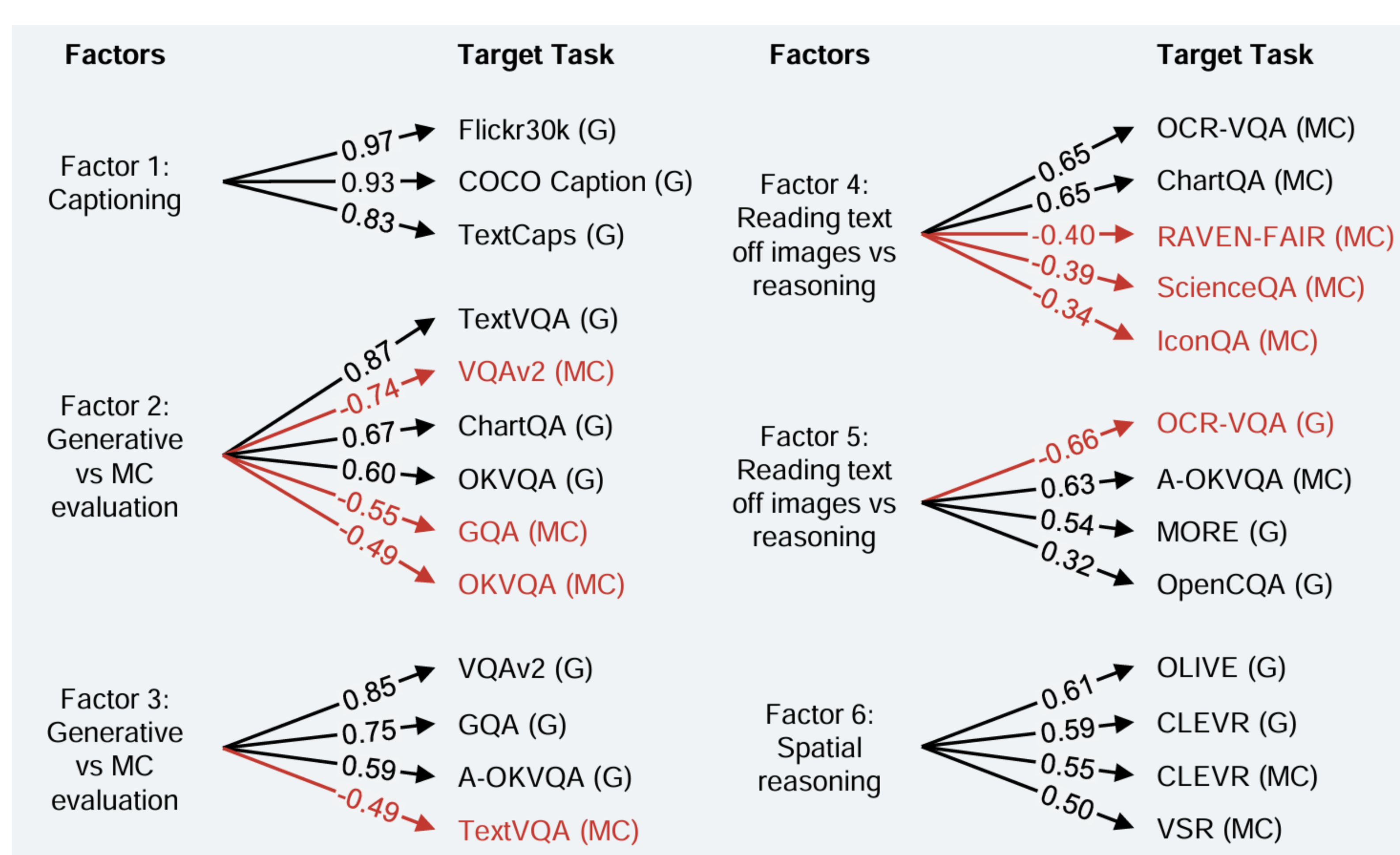



Figure 2: Results of EFA on the residuals $\bar{A}$ after isolating the dominating factor influencing classification and most VQA tasks. Black (red) arrows indicate positive (negative) loadings. Cut-off for factor loadings=0.3. Notably, New Yorker Explanation and Ranking, and Hateful Memes, do not have loadings above 0.3 on any discovered factor.

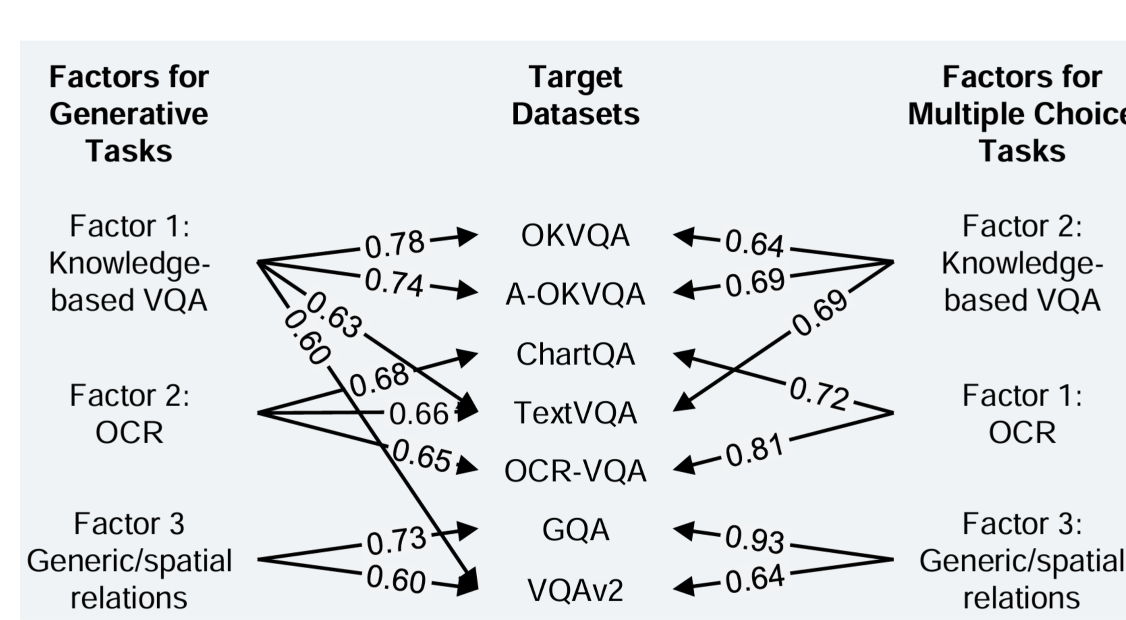



Figure 3: Results of separate EFA on generative and MC VQA tasks. Cut-off for factor loadings= 0.6. Similar structures observed highlight EFA's efficacy in capturing underlying structures with suitable data.