Motivation:
Most vision-language models and V-L tasks are English-centered. It is challenging yet rewarding to generalize V-L models to other 7000 languages.

Our solution:
Zero-shot cross-lingual transfer that utilizes one model to rule all the languages. We propose to learn multilingual multimodal representations with:
- Multilingual Multimodal Transformers
- Multilingual Multimodal Pre-training

Tasks of our focus in this study:
Multilingual text-video search.

Zero-Shot Cross-Lingual Transfer

Task-agnostic Task-specific

Step 1
- Task-agnostic multilingual multimodal pre-training

Step 2
- Task-specific English fine-tuning

Step 3
- Zero-shot transfer to non-English tasks

Multi-HowTo100M dataset
- 1.2 million instructional videos. 138 million clips.
- Transcriptions in 9 languages: English, German, French, Russian, Spanish, Czech, Swahili, Chinese, Vietnamese.

<table>
<thead>
<tr>
<th>NLP Tasks (e.g., EXTREME)</th>
<th>V-L Tasks (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Multilingual Transformer</td>
</tr>
<tr>
<td>Step 1</td>
<td>Task-agnostic multilingual pre-training</td>
</tr>
<tr>
<td>Step 2</td>
<td>Task-specific English fine-tuning</td>
</tr>
<tr>
<td>Step 3</td>
<td>Zero-shot transfer to non-English tasks</td>
</tr>
</tbody>
</table>

Qualitative Results
- a visual step by step walking out on the field
- ...