



Molmo & PixMo

Camille Lançon, Arthur Leene,
Elia Mangin, Errel Coz, Hugo Khlaut



Summary

1. Context

2. Architecture

3. Data

4. Training

5. Evaluation

6. Ablation

7. Related Works

8. Conclusion

Context

Vision language model (VLM): a model capable of understanding and generating visual data and textual data simultaneously.



Gives keys on how to build a performant Vision Language Model from scratch

- There is a ***distillation*** of VLMs : their data rely on other VLMs
- Early works like LLaVA are open weight but now behind the state of the art

Context

Molmo is a family of VLMs that relies on the dataset **PixMo** (pixels for Molmo)

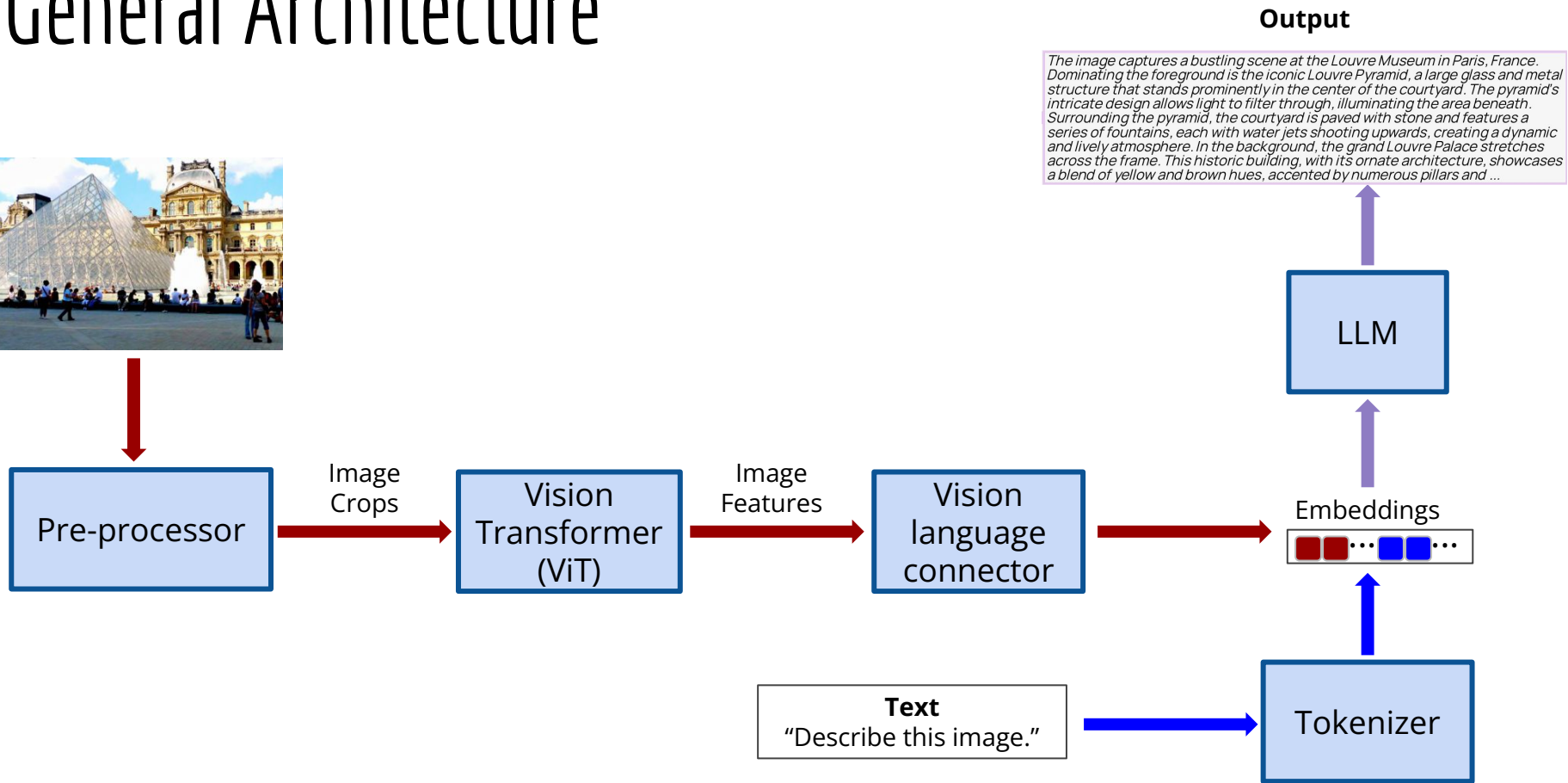
It provides great results thanks to :

- Quality of data collected
- Careful modeling choices
- Well tuned pipeline

Summary

1. Context
- 2. Architecture**
3. Data
4. Training
5. Evaluation
6. Ablation
7. Related Works
8. Conclusion

General Architecture



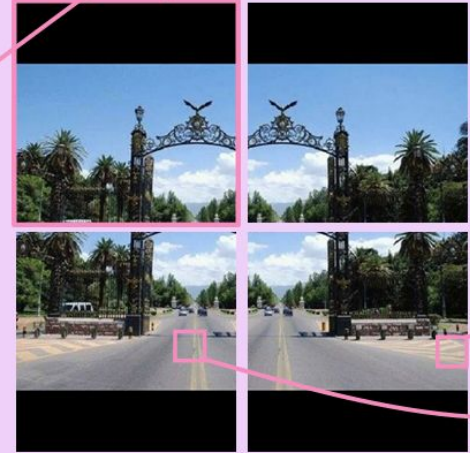
Pre-processor

- Fixed ViT input size: too small for high resolution (prevents recognition of details)
- Full image in low resolution + Image divided into crops
- Allow crops to overlap to bring in the surrounding context

Image

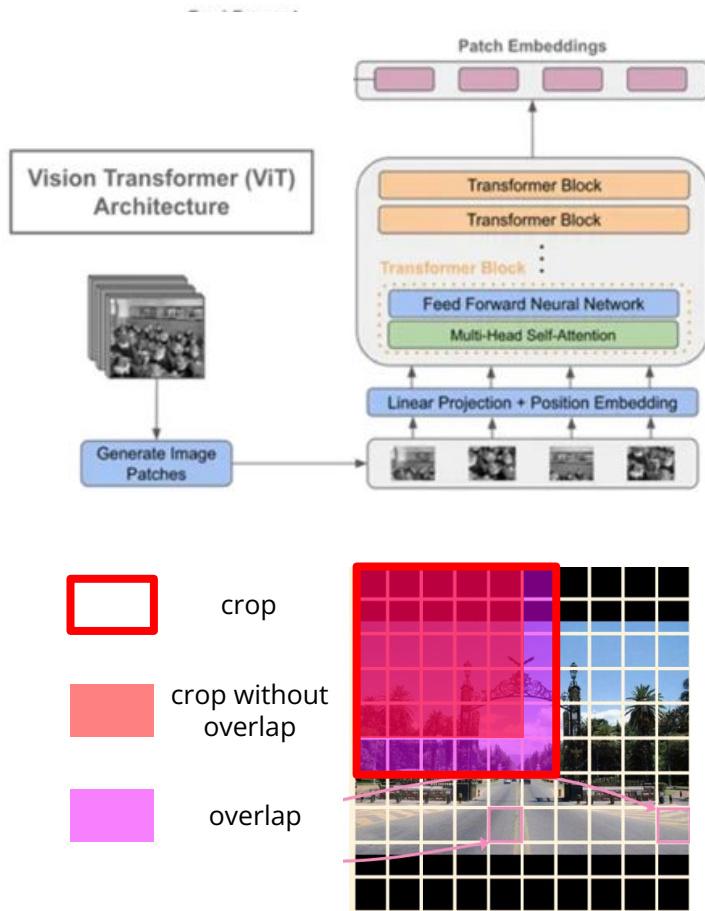


Crops



Vision Transformer (ViT) Encoder

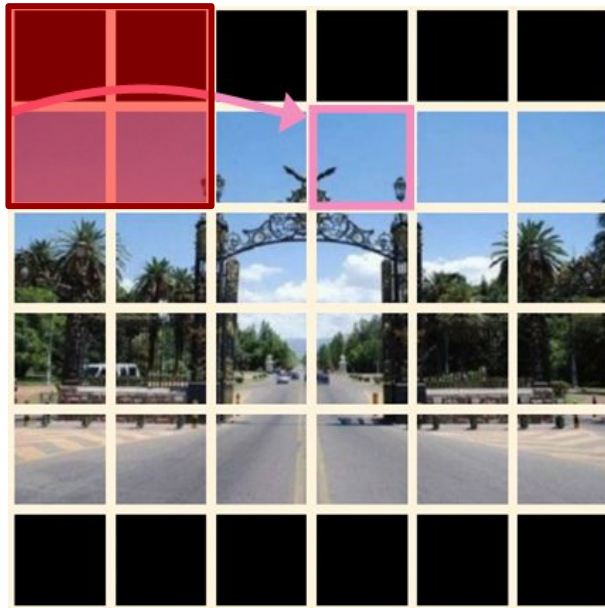
- Each crop is encoded independently by the ViT.
- Each crop is divided into patches to form the sequence of input tokens.
- The ViT produces one feature per patch.
- Features from overlapping patches are not transmitted to the Vision-language connector.
- Final features : concatenation of features from the third-to-last and 10-from-last layers.



Vision-language connector

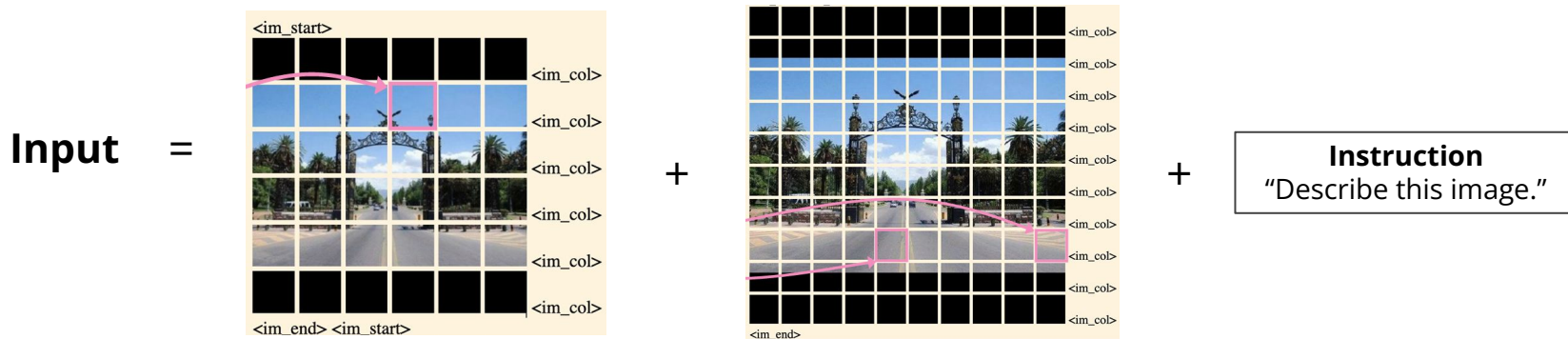
- Patch features from all crops are brought together and pooled on a 2x2 window
- Multi-headed attention pooling:
 - query = average of the 4 patch features
 - keys and values = patch features.The attention scores are calculated to determine a weighted sum of the 4 patch features.
- Pooled features are mapped to the LLM's embedding space via an MLP.

2x2 window



LLM

- Inputs: tokens from image processing and from an instruction or question about the image.
- Order:
 - Patches from the low resolution image
 - Patches from left to right, top to bottom
 - Special tokens inserted to indicate the beginning and end of patches.

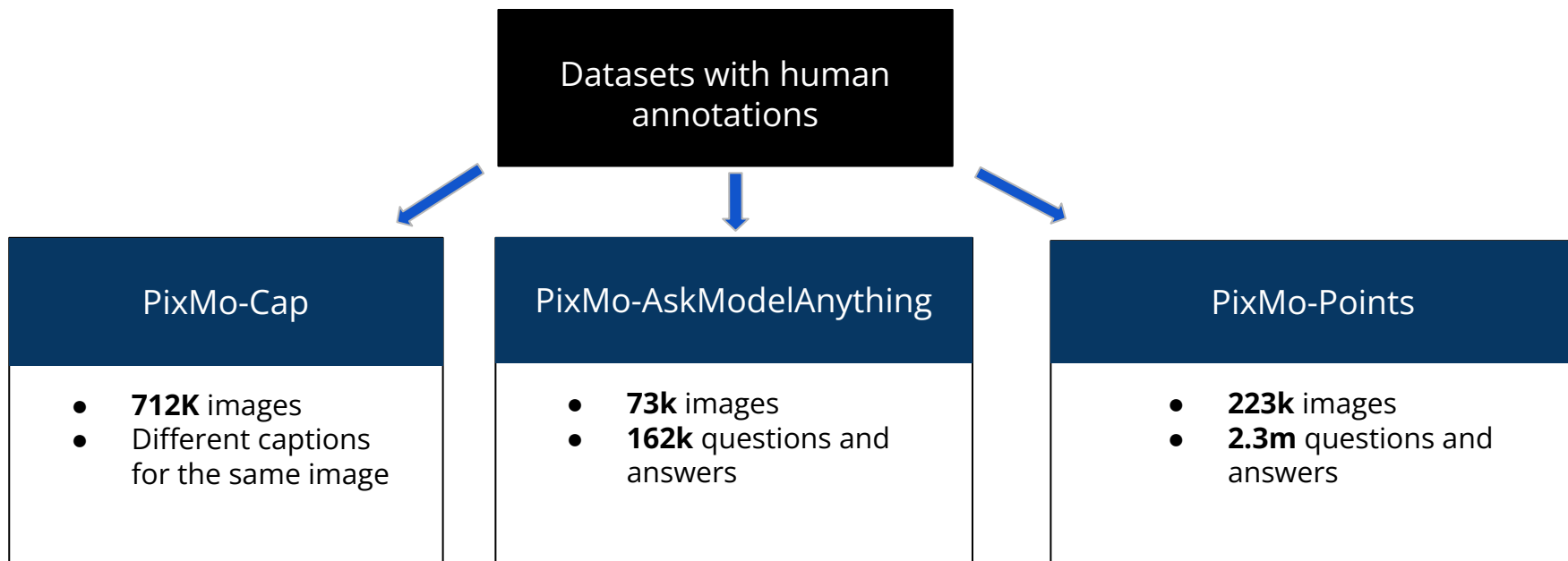


Summary

1. Context
2. Architecture
- 3. Data**
4. Training
5. Evaluation
6. Ablation
7. Related Works
8. Conclusion

PixMo Datasets

💡 High quality data, because of new methods of data collection

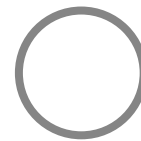


PixMo-Cap

💡 Use voice-to-speech to be more specific and to prevent them from using other VLM

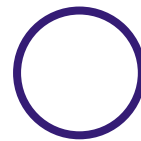


Voice description



Text transcript

LLM

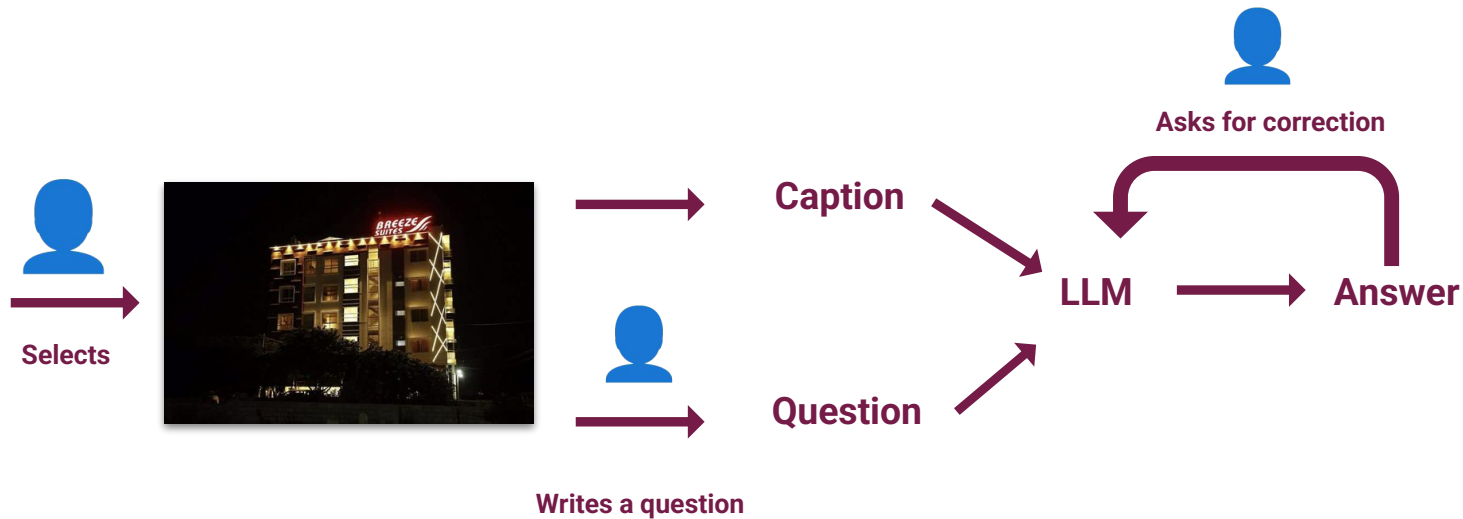


Re-written caption

This photograph, taken at night, captures the striking image of a hotel called Breeze Suites. The sky is pitch black, with no stars or a moon visible...

PixMo-AskModelAnything

Fine-tuning dataset for a model pre-trained on PixMo-Cap



PixMo-Pointing

3 goals

- Point items described by text
- enable the model to count by pointing
- use pointing as a form of visual explanation when answering questions



Way data collected than other well-known datasets
223k images vs **20k** (COCO) or **141k** (RefCOCO+)



Find the plates

```
<points x1="33.0"
y1="63.7" x2="34.2"
y2="67.1" x3="34.5"
y3="62.5" x4="36.1"
y4="65.9" x5="36.4"
y5="68.4" x6="37.6"
y6="71.6"
alt="plates">plates</points>
```



Can you show me where the top of the logs are?

There are none.

PixMo Synthetic datasets



Data generated by other language only LLMs

- **PixMo-CapQA** : questions & answers based on a caption
- **PixMo-Clocks**: synthetic clocks matched with a time-telling question-answer pair
- **PixMo-Docs**: Code generation thanks to generated graphs, tables, diagrams...
- **PixMo-Count**: non-VLM object detector provides the dominating class of object in a picture. Generated pairs of questions & answers about objects of this class.



What time is being shown?

The time shown is 1:30:59

Example of
PixMo-Clocks



What is the title of the figure?

Lukinhas ba Fan Meetup Participation in Salvador

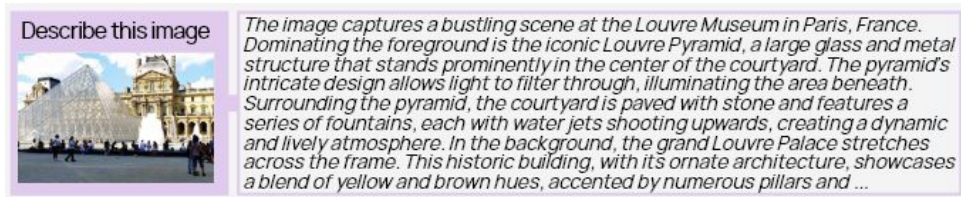
Example of
PixMo-Docs

Summary

1. Context
2. Architecture
3. Data
- 4. Training**
5. Evaluation
6. Ablation
7. Related Works
8. Conclusion

Pre training

Objective: Pre-train all model parameters on PixMo-Cap to generate captions or audio transcripts for images.



Key Features:

- Prompt Guidance: 90% of prompts include a length hint to improve caption quality.
- No Separate Connector Training: Higher learning rate and shorter warmup for connector parameters.

Optimization:

- Optimizer: AdamW with cosine learning rate decay.
- Learning Rates: Connector ($2e-4$), ViT ($6e-6$), LM ($2e-5$).
- Warmup Steps: 200 (connector), 2000 (ViT and LM).
- Gradient Clipping: Applied separately to LM, image encoder, and connector

Fine Tuning

Datasets: Mix of PixMo datasets and open-source datasets (e.g., VQA v2, TextVQA, DocVQA, ChartQA).

Sampling Strategy:

- Proportional to the square root of dataset size.
- Down-weight large synthetic datasets (e.g., PlotQA, FigureQA).
- Up-weight pointing tasks (learn slower than QA tasks).

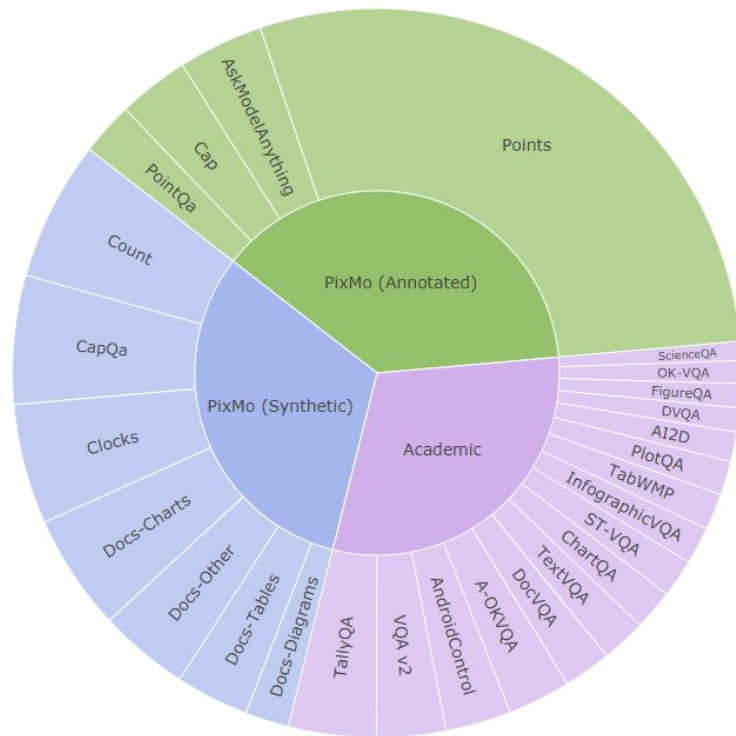


Figure 4. Datasets used for fine-tuning, shown in proportion to their sampling rates. Green denotes human-annotated data we collected, blue denotes synthetic data we generated, and purple represents pre-existing academic datasets. PixMo-Docs has been subdivided into charts, tables, diagrams, and other.

Fine Tuning

Style Tags:


- Prevent dataset-specific answer styles.
 - a. Example: Prefix VQA v2.0 questions with "vqa2:".
 - b. Exceptions: PixMo-AskModelAnything, -CapQA, -Points, -Count, -Cap.

Prompt Design:

- Caption Generation: ~30 prompts for PixMo-Cap.
- Pointing Tasks: ~100 question templates for location/count queries.
- Pointing Output: Normalized coordinates (0-100), ordered top-down, left-to-right.


User Interaction

Where is this bus going?



The bus is heading to Highland Park. This is clearly visible on the destination sign at the front of the bus, which displays "71B HIGHLAND PARK"


Its 2:00pm on November 15th, can I park here?



No, you cannot park here at 2:00 PM on November 15th. The parking sign clearly shows that from 1 PM to 11 PM on November 15th, there is no parking allowed in this area...


Pointing and Counting

Count the tables




15 tables

What things can be used to write with?



Things that can be used to write with


Where is the Odometer?



Odometer


Visual Skills

What were the total sales of Cassettes and Vinyl in 1980?




10 billion

According to the data sheet, what is the bounce rate of Udacity?



34%

What time is it in New York?



2:53

Notable features

Pointing as Chain-of-Thought:

- Model counts/localizes by sequentially pointing to each object.
- Improves accuracy, interpretability, and generalization.



Evaluation:

- Focused on academic benchmarks and user-facing performance.
- Ensures the model adapts to diverse tasks (QA, pointing, captioning) while maintaining high-quality outputs.

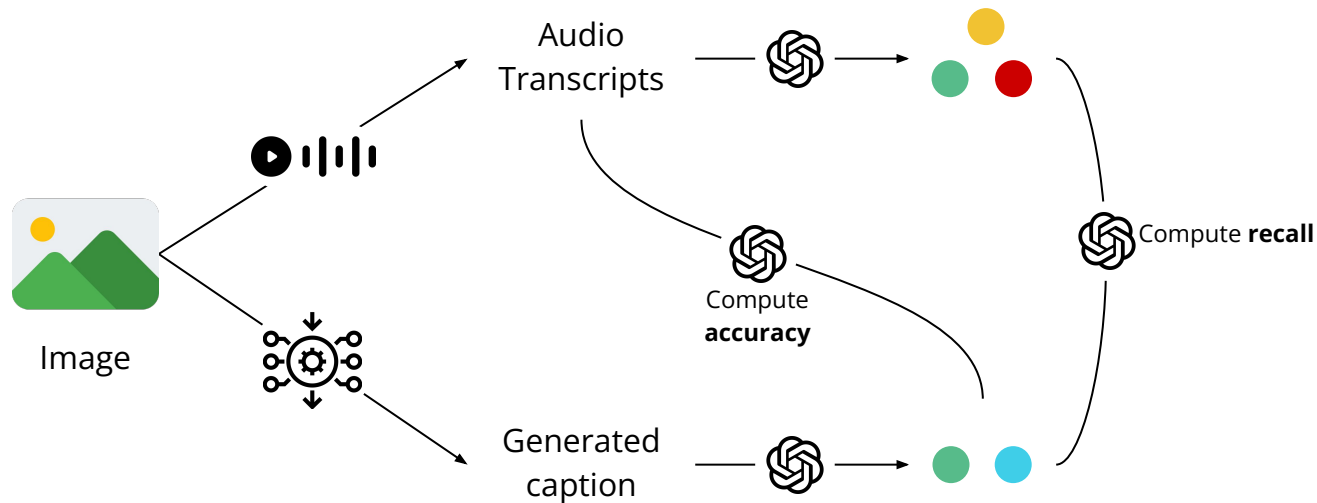
Key Training Features:

- Simplified training pipeline by skipping separate connector training.
- Balanced dataset sampling and style tags ensure robust generalization.

Summary

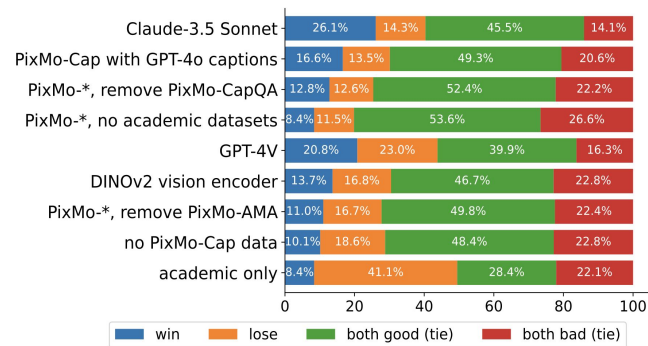
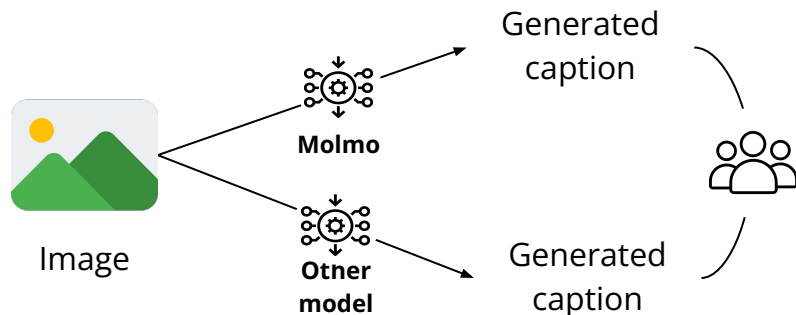
1. Context
2. Architecture
3. Data
4. Training
- 5. Evaluation**
6. Ablation
7. Related Works
8. Conclusion

Cap F1



- 1500 images selected manually with up to 6 audio transcripts per image
- Compute mean accuracy and mean recall to get final F1 score
- Influenced model design and pre-training data decision

Human Evaluation



Human evaluation outcomes for matches between various models vs. Molmo-7B-D

- Necessity to overcome the specificity of traditional academic benchmarks
- 15k image-text-prompt instances
- 870 humans with a total of 325k ratings
- Elo ranking using the Bradley-Terry model

Benchmark

model	Average	Elo score	Elo rank
API call only			
GPT-4V	71.1	1041	10
GPT-4o-0513	78.5	1079	1
Gemini 1.5 Flash	75.1	1054	7
Gemini 1.5 Pro	78.3	1074	3
Claude-3 Haiku	65.3	999	18
Claude-3 Opus	66.7	971	21
Claude-3.5 Sonnet	76.7	1069	4
Open weights only			
PaliGemma-mix-3B	50.0	937	27
Phi3.5-Vision-4B	59.7	982	19
Qwen2-VL-7B	73.7	1025	14
Qwen2-VL-72B	79.4	1037	12
InternVL2-8B	69.4	953	23
InternVL2-Llama-3-76B	77.1	1018	16
Pixtral-12B	69.5	1016	17
Llama-3.2V-11B-Instruct	69.8	1040	11
Llama-3.2V-90B-Instruct	74.5	1063	5
Open weights			
LLaVA-1.5-7B	40.7	951	26
LLaVA-1.5-13B	43.9	960	22
xGen-MM-interleave-4B	59.5	979	20
Cambrian-1-8B	63.4	952	25
Cambrian-1-34B	66.8	953	24
LLaVA OneVision-7B	72.0	1024	15
LLaVA OneVision-72E	76.6	1051	8
The Molmo family			
MolmoE-1B	68.6	1032	13
Molmo-7B-O	74.6	1051	9
Molmo-7B-D	77.3	1056	6
Molmo-72B	81.2	1077	2

11 Academic benchmarks :

Nature

RealWordQA, VQA v2.0

→ match or outperform all models

OCR

AI2D, ChartQA, DocQA, InfoQA, TextVQA

→ surpass open models

Reasoning

MMMU, Math Vista

→ lag behind other models

Counting

CountBenchQA, PixMo-Count

→ lead all models

Key points :

- Added the PixMo-Count
- Elo Rank ≠ from Chatbot Arena's one

Results :

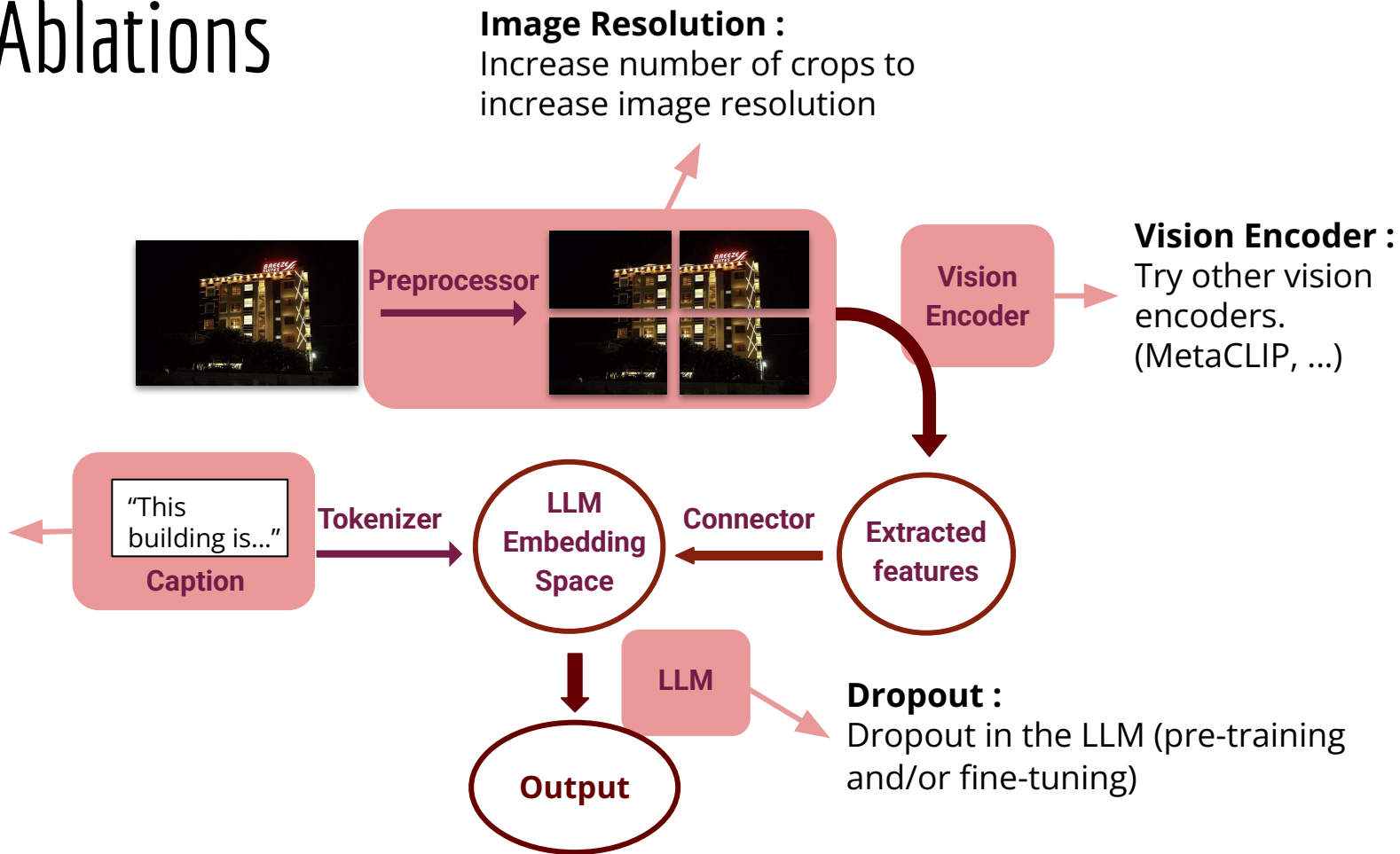
- MolmoE-1B** nearly matches GPT-4V
- Molmo-72B** achieves the **highest** academic benchmark and ranks **2nd**

Summary

1. Context
2. Architecture
3. Data
4. Training
5. Evaluation
- 6. Ablation**
7. Related Works
8. Conclusion

Model Ablations

Length Conditioning :
Remove the length hint for captioning task



Data Ablations and Counting

PixMo Cap



"This building is..."

Caption

VS

Web Data



ShareGPT4v/o

- **PixMo Cap Scaling** : amount of data used in pre-training and fine-tuning => significant impact on the performances
- **Pre-training data** : Adding noisy web data does not improve the pre-training, PixMo Cap is of better quality.
- **Supervised fine-tuning data** : PixMo supervised fine-tuning datasets improve the performance on complex tasks.

- **Counting** : Way of encoding points, point then count, count in order,...

Summary

1. Context
2. Architecture
3. Data
4. Training
5. Evaluation
6. Ablation
- 7. Related works**
8. Conclusion

Related works

Vision-Language Models

- CLIP & ALIGN [5]: Train on noisy web data for language-aligned image encoders.
 - Excel in classification and retrieval.
 - Limitations: Struggle with fine details.
- Open-Source Efforts: Post-CLIP, some works [1, 6] aim for fully open pipelines.

Multimodal LLMs

- Use CLIP-style encoders + LLMs via connectors or cross-attention.
- Some input raw pixels instead of embeddings.
- Efficient models rise due to compute constraints.
- Proprietary models dominate; open models lack training details.

Related works

Instruction Tuning:

- Generate QA pairs from annotations using LLMs. [2]
- Often relies on proprietary VLMs for annotations, creating closed pipelines.
- Limitations: Noisy annotations and incomplete descriptions.

Synthetic Datasets:

- Chart Generation: Focus on bar/line charts [4]
- Clock Data: Synthetic datasets lack real-world diversity (e.g., watch faces). [7]
- VLM Grounding: Use object detectors or referring expression datasets (e.g., GRES [3]).

Bootstrapping from LLMs:

- Use closed LLMs for dataset curation will be open LLMs when they are good enough.
- Goal: Build open VLMs without proprietary dependencies.

Summary

1. Context
2. Architecture
3. Data
4. Training
5. Evaluation
6. Ablation
7. Related works
- 8. Conclusion**

Conclusion

Molmo VLMs establish themselves as **state-of-the-art** solutions :

- High-quality dataset (no distillation)
- Open weights, open data, open training code, open evaluations
- Well-designed pipeline
- Impressive results on academic benchmark and human evaluation

Sources

- [1] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In CVPR, 2023.
- [2] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, C. Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425, 2023.
- [3] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In CVPR, 2023.
- [4] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In WACV, 2020.
- [5] 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. In ICML, 2021.
- [6] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In ICLR, 2024.
- [7] Charig Yang, Weidi Xie, and Andrew Zisserman. It’s about time: Analog clock reading in the wild. In CVPR, 2022.



Thanks for your attention



Appendix

model	A12D test [49]	ChartQA test [82]	VQA v2.0 testdev [36]	DocVQA test [83]	InfoQA test [84]	TextVQA val [100]	RealWorldQA [116]	MMMU val [129]	MathVista testmini [78]	CountBenchQA [10]	PixMo-Count test	Average	Elo score	Elo rank
<i>API call only</i>														
GPT-4V [88]	89.4	78.1	77.2	87.2	75.1	78.0	61.4	63.1	58.1	69.9	45.0	71.1	1041	10
GPT-4o-0513 [90]	94.2	85.7	78.7	92.8	79.2	77.4	75.4	69.1	63.8	87.9	59.6	78.5	1079	1
Gemini 1.5 Flash [103]	91.7	85.4	80.1	89.9	75.3	78.7	67.5	56.1	58.4	81.6	61.1	75.1	1054	7
Gemini 1.5 Pro [103]	94.4	87.2	80.2	93.1	81.0	78.7	70.4	62.2	63.9	85.8	64.3	78.3	1074	3
Claude-3 Haiku [7]	86.7	81.7	68.4	88.8	56.1	67.3	45.5	50.2	46.4	83.0	43.9	65.3	999	18
Claude-3 Opus [7]	88.1	80.8	66.3	89.3	55.6	67.5	49.8	59.4	50.5	83.6	43.3	66.7	971	21
Claude-3.5 Sonnet [7]	94.7	90.8	70.7	95.2	74.3	74.1	60.1	68.3	67.7	89.7	58.3	76.7	1069	4
<i>Open weights only</i>														
PaliGemma-mix-3B [10]	72.3	33.7	76.3	31.3	21.4	56.0	55.2	34.9	28.7	80.6	60.0	50.0	937	27
Phi3.5-Vision-4B [1]	78.1	81.8	75.7	69.3	36.6	72.0	53.6	43.0	43.9	64.6	38.3	59.7	982	19
Qwen2-VL-7B [111]	83.0	83.0	82.9	94.5	76.5	84.3	70.1	54.1	58.2	76.5	48.0	73.7	1025	14
Qwen2-VL-72B [111]	88.1	88.3	81.9	96.5	84.5	85.5	77.8	64.5	70.5	80.4	55.7	79.4	1037	12
InternVL2-8B [104]	83.8	83.3	76.7	91.6	74.8	77.4	64.2	51.2	58.3	57.8	43.9	69.4	953	23
InternVL2-Llama-3-76B [104]	87.6	88.4	85.6	94.1	82.0	84.4	72.7	58.2	65.5	74.7	54.6	77.1	1018	16
Pixtral-12B [3]	79.0	81.8	80.2	90.7	50.8	75.7	65.4	52.5	58.0	78.8	51.7	69.5	1016	17
Llama-3.2V-11B-Instruct [5]	91.1	83.4	75.2	88.4	63.6	79.7	64.1	50.7	51.5	73.1	47.4	69.8	1040	11
Llama-3.2V-90B-Instruct [5]	92.3	85.5	78.1	90.1	67.2	82.3	69.8	60.3	57.3	78.5	58.5	74.5	1063	5
<i>Open weights + data († distilled)</i>														
LLaVA-1.5-7B [69]	55.5	17.8	78.5	28.1	25.8	58.2	54.8	35.7	25.6	40.1	27.6	40.7	951	26
LLaVA-1.5-13B [69]	61.1	18.2	80.0	30.3	29.4	61.3	55.3	37.0	27.7	47.1	35.2	43.9	960	22
xGen-MM-interleave-4B† [119]	74.2	60.0	81.5	61.4	31.5	71.0	61.2	41.1	40.5	81.9	50.2	59.5	979	20
Cambrian-1-8B† [106]	73.0	73.3	81.2	77.8	41.6	71.7	64.2	42.7	49.0	76.4	46.6	63.4	952	25
Cambrian-1-34B† [106]	79.7	75.6	83.8	75.5	46.0	76.7	67.8	49.7	53.2	75.6	50.7	66.8	953	24
LLaVA OneVision-7B† [59]	81.4	80.0	84.0	87.5	68.8	78.3	66.3	48.8	63.2	78.8	54.4	72.0	1024	15
LLaVA OneVision-72B† [59]	85.6	83.7	85.2	91.3	74.9	80.5	71.9	56.8	67.5	84.3	60.7	76.6	1051	8
<i>The Molmo family: Open weights, Open data, Open training code, Open evaluations</i>														
MolmoE-1B	86.4	78.0	83.9	77.7	53.9	78.8	60.4	34.9	34.0	87.2	79.6	68.6	1032	13
Molmo-7B-O	90.7	80.4	85.3	90.8	70.0	80.4	67.5	39.3	44.5	89.0	83.3	74.6	1051	9
Molmo-7B-D	93.2	84.1	85.6	92.2	72.6	81.7	70.7	45.3	51.6	88.5	84.8	77.3	1056	6
Molmo-72B	96.3	87.3	86.5	93.5	81.9	83.1	75.2	54.1	58.6	91.2	85.2	81.2	1077	2