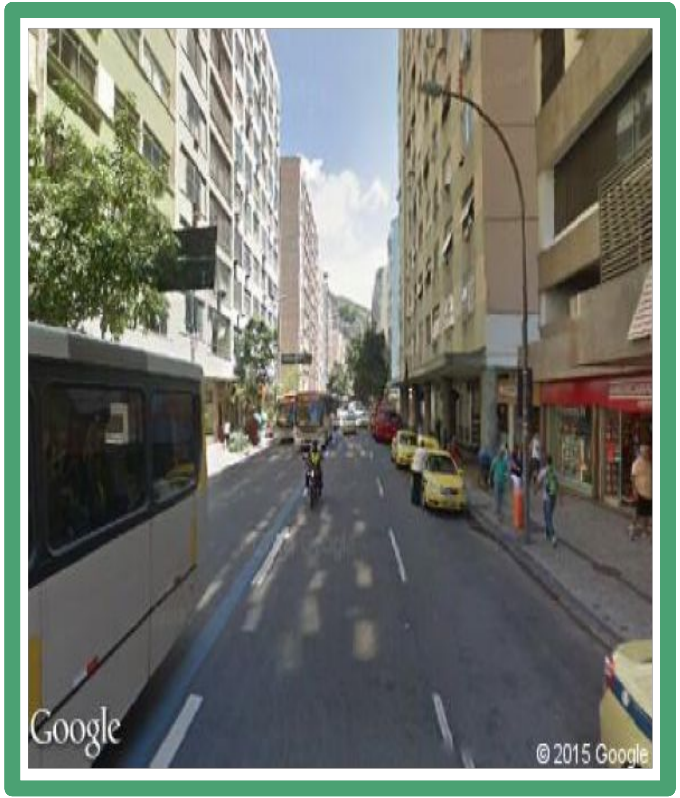
Assessing Urban Environments with Vision-Language Models: A Comparative Analysis of AI-Generated Ratings and Human Volunteer Evaluations

Felipe A. Moreno-Vera and Jorge Poco

Context & Motivation

Which one looks safer?

Bangú (RJ)



City Center (RJ)

Context

By understanding how people perceive and experience cities, we can create more complex models to analyze the perception and obtain insights from inferences.

Motivation

Using vision-language models we add more subjective information from Street View images, aiming to understanding urban perception based on dual-modality studies.

Overview

- **Main goal:** Analyze the impact of adding textual descriptions to images for evaluating urban perception.

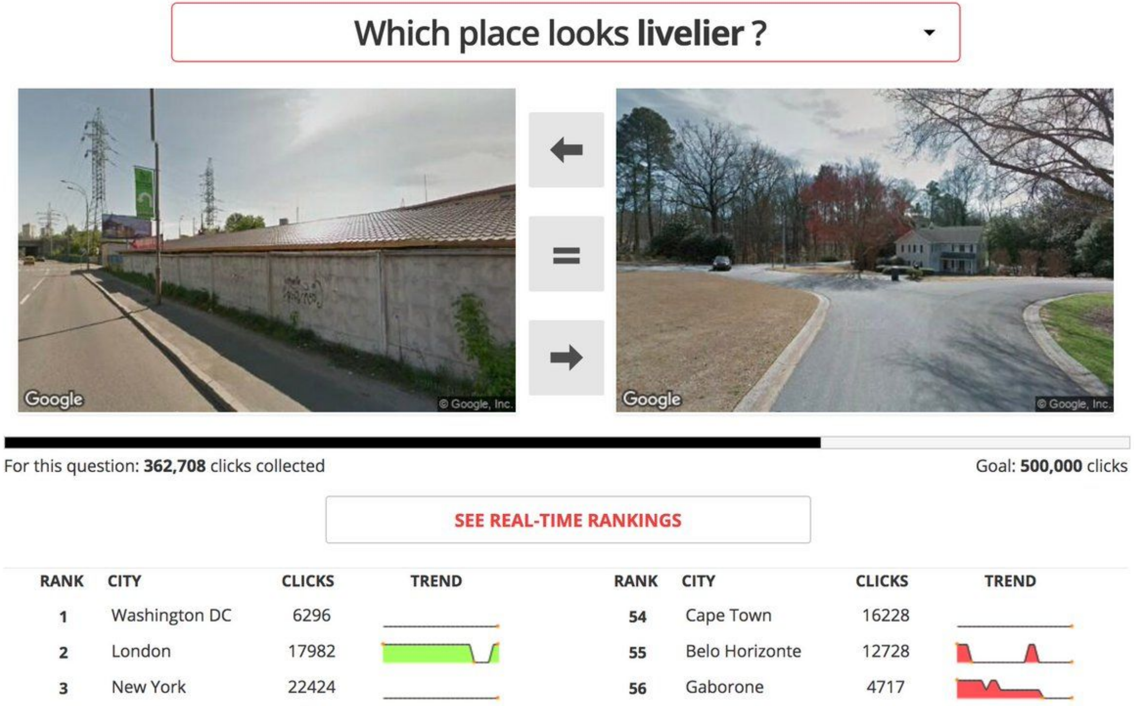
- **Image-to-Text model comparison:** We compare LlaVA (1.5-7b-hf), BLIP-2 (opt-2.7b), and BLIP (ic-large) for image descriptions.

- **Model training:** Binary classification and regression tasks.

- **Ablation studies:** We freeze and unfreeze certain components and layers to study their impact and understand the contribution of each part of the model to the overall performance.
    - **Image-to-Text models:** Generates “positive” and “negative” descriptions
    - **Dual-modality:** The projections from Image and Text encoders
    - **Contrastive Image-text alignment**: Contrastive learning
    - **Heads:** Only classification and regression heads

Place Pulse

Place Pulse dataset



http://pulse.media.mit.edu/

* Comparisons were made using two random images from random cities.

Place Pulse dataset

left-id	right-id	winner	left-lat	left-long	right-lat	right-long	category
513d7e23fdc9f	513d7ac3fdc9f	equal	40.744156	-73.93557	-33.52638	-70.591309	depressing
513f320cfdc9f	513cc3acfdc9f	left	52.551685	13.416548	29.76381	-95.394621	safety
513e5dc3fdc9f	5140d960fdc9f	right	48.878382	2.403116	53.32932	-6.231007	lively

- 1 223 649 Comparisons
- 111 390 images
- 32 countries , 56 cities
- 6 categories: safety, lively, beauty, wealthy, boring, and depressing

Strength of Schedule*

$$Award_i^k = \frac{1}{w_i^k} \sum_{j=1}^{n_1} \frac{w_i^k}{w_i^k + d_i^k + l_i^k}$$

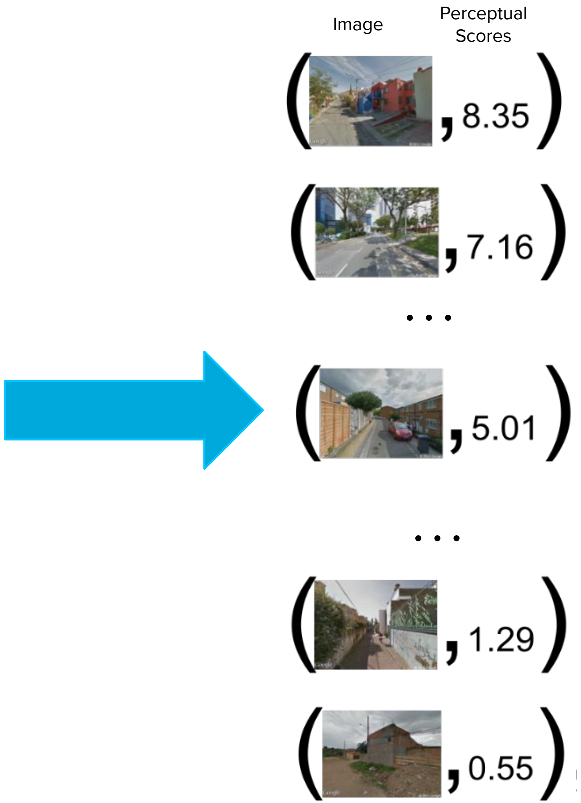
$$Penalty_i^k = \frac{1}{l_i^k} \sum_{j=1}^{n_2} \frac{l_i^k}{w_i^k + d_i^k + l_i^k}$$

$$Q_i^k = \frac{10}{3}(\frac{w_i^k}{w_i^k + d_i^k + l_i^k} + Award_i^k - Penalty_i^k + 1)$$

* Park et. al., A network-based ranking system for us college football

Strength of Schedule

left	right	winner
		draw
		left
		right
⋮	⋮	⋮
		right
		left



Processed samples

Image	ID	Safety	Lively	Wealthy	Beauty	Boring	Depressive
	513d7e23fdc9f	7.42	8.58	6.5	7.3	2.64	1.23
	513f320cfdc9f	6.07	4.97	7.13	8.61	1.67	0.86

Statistics

Place Pulse 2.0			
Continent	#countries	#cities	#images
Europe	19	22	38,747
North America	3	17	37504
South America	2	5	12,524
Asia	5	7	11,417
Oceania	1	2	6,097
Africa	2	3	5,101
Total	32	56	111,390

Place Pulse 2.0			
Category	# comparisons	# images	*mean*
Safety	368,926	111,389	5.188
Lively	267,292	111,348	5.085
Beautiful	175,361	110,766	4.920
Wealthy	152,241	107,795	4.890
Depressing	132,467	105,495	4.816
Boring	127,362	106,363	4.810
Total	1,223,649		

High safety scores images

Low safety scores images

Experiments and Results

Experiment settings

- **Place Pulse 2.0**
  - Dataset split into 75% for training and 25% for validation/testing.
  - Binary labeling:

$$y_{i,k} = \begin{cases} 1 & \text{if } Q_i^k > \mu^k + \delta\sigma^k\% \\ 0 & \text{if } Q_i^k < \mu^k - \delta\sigma^k\% \end{cases}$$

  - 5 Cross-Validation

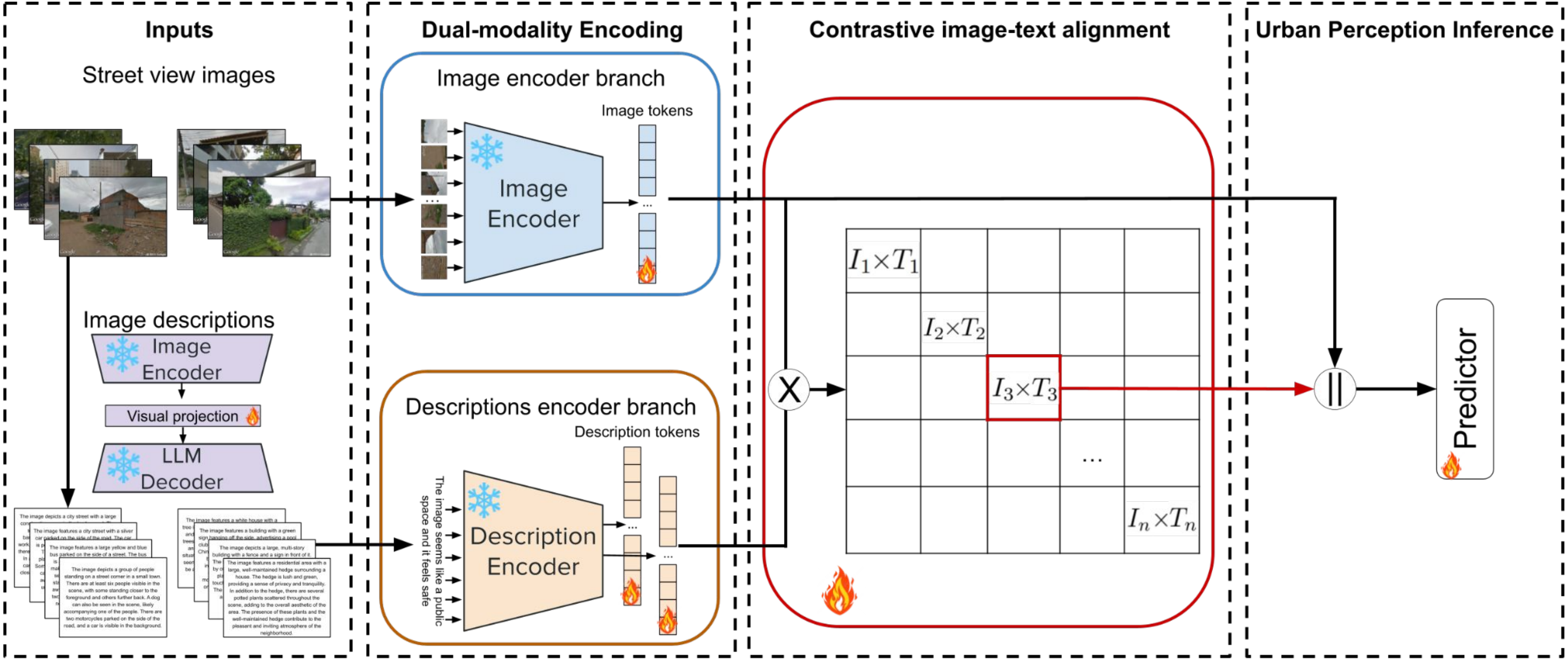
- **Environment:**
  - Python 3.10 + Pytorch
  - NVIDIA RTX 3090 GPU with 16 VRAM.
  - Intel Core i9-12900K CPU with 125Gb RAM.
  - We use float16 precision for memory save.

Image descriptions

ID	50f5ec33fdc9f065f00087c3	50f5ec42fdc9f065f00088e0	50f5eb65fdc9f065f000826f
Image			
Model	**Description**	**Description**	**Description**
LlaVA	The image depicts a narrow alleyway between two buildings, with one of the buildings being a brick structure. The alleyway is surrounded by a dirt road, and there are a few cars parked along the road.	The image depicts a narrow alleyway between two buildings, with one of the buildings being a brick structure. The alleyway is surrounded by a dirt road. There are a few cars parked along the road.	The image shows a residential area with a well-maintained hedge around a house and several potted plants, creating a pleasant, aesthetic, and inviting atmosphere.
BLIP-2	This image shows a narrow street in a residential area under development or construction. The buildings are primarily made of exposed and unfinished red bricks and concrete.	This image depicts a wide, well-maintained urban street in a more affluent area. The environment appears clean, organized, and secure, with good infrastructure and urban planning.	The image shows a residential street scene. Additionally, a tall hedge covers a gate and wall, possibly concealing a private residence.
BLIP	This is a Google Street View image of a building under construction.	This is a Google Street View image of a building in a city.	This is a Google Street View image of a green residential area in the Philippines.

Randomly select 50 samples and compare the description results

UrbanVLM

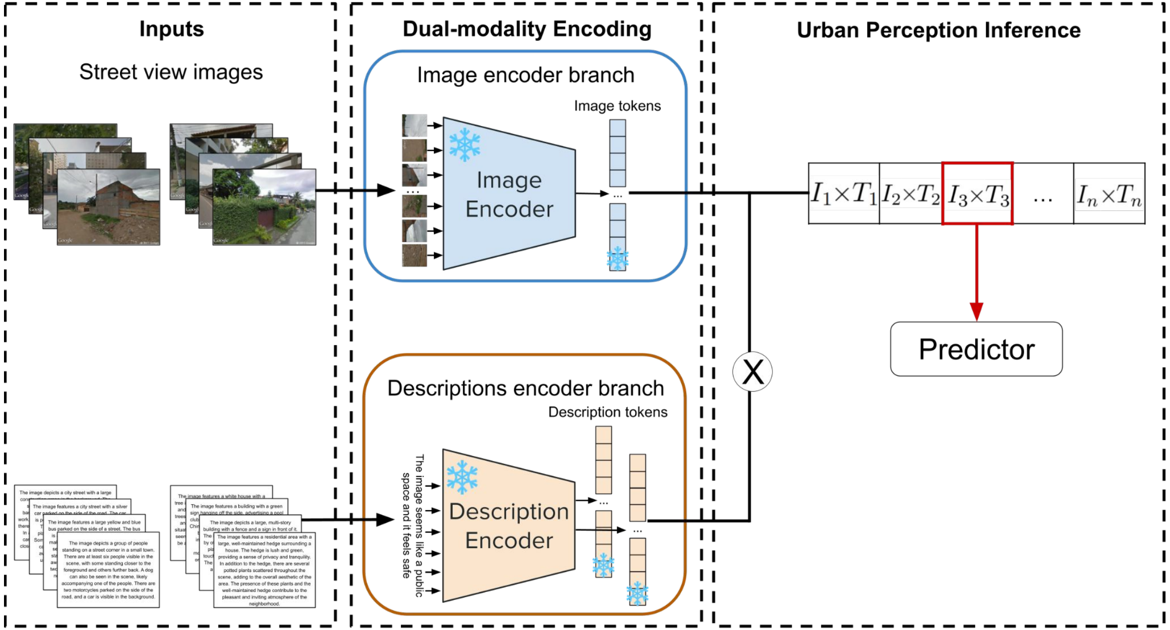

Inputs
Street view images
Image descriptions
Image Encoder
Visual projection
LLM Decoder
Dual-modality Encoding
Image encoder branch
Image tokens
Image Encoder
Descriptions encoder branch
Description tokens
The image seems like a public space and it feels safe
Description Encoder
Contrastive image-text alignment
$I_1 \times T_1$
$I_2 \times T_2$
$I_3 \times T_3$
$I_n \times T_n$
Urban Perception Inference
Predictor

Ablation study

Ablation Study	Model Tested	Classification				Regression		
		Acc	Precision	Recall	F-1	R^2	RMSE	MAE
Zero-shot	CLIP	0.39	0.41	0.39	0.24	-14.05	4.53	4.89
	SigLIP	0.57	0.43	0.57	0.45	-14.17	4.61	4.77
Only heads W/o description, contrastive & dual-modality	LlaVA+CLIP	**0.67**	**0.67**	0.66	**0.66**	**0.57**	**2.43**	**2.56**
	LlaVA+SigLIP	0.66	0.66	**0.67**	**0.66**	0.56	**2.43**	2.68
	BLIP-2+CLIP	0.63	0.61	0.62	0.61	0.53	3.4	3.21
	BLIP-2+SigLIP	0.64	0.63	0.63	0.63	0.53	3.38	3.35
Contrastive W/o description & dual-modality	LlaVA+CLIP	0.7	0.69	0.68	0.68	**0.62**	**1.81**	1.95
	LlaVA+SigLIP	**0.71**	**0.71**	**0.7**	**0.7**	0.61	1.98	**1.84**
	BLIP-2+CLIP	0.68	0.67	0.68	0.67	0.56	2.75	2.35
	BLIP-2+SigLIP	0.69	0.68	0.69	0.68	0.55	2.68	2.2
Visual projections W/o contrastive & dual-modality	LlaVA+CLIP	**0.73**	**0.72**	**0.71**	**0.71**	**0.67**	1.69	1.73
	LlaVA+SigLIP	0.72	**0.72**	**0.71**	**0.71**	0.65	**1.68**	**1.71**
	BLIP-2+CLIP	0.7	0.7	0.69	0.69	0.59	1.95	2.06
	BLIP-2+SigLIP	0.71	0.71	0.7	0.7	0.59	1.88	1.94
Dual-modality W/o description & contrastive	LlaVA+CLIP	**0.76**	**0.76**	**0.75**	**0.75**	**0.78**	1.33	**1.42**
	LlaVA+SigLIP	0.75	0.75	0.74	0.74	0.75	**1.29**	1.51
	BLIP-2+CLIP	0.72	0.72	0.73	0.72	0.69	1.6	1.34
	BLIP-2+SigLIP	0.73	0.73	0.72	0.72	0.68	1.4	1.21
UrbanVLM	LlaVA+CLIP	0.82	0.78	**0.79**	**0.78**	**0.84**	**1.04**	**0.78**
	LlaVA+SigLIP	**0.83**	**0.79**	0.78	**0.78**	0.83	1.08	0.79
	BLIP-2+CLIP	0.78	0.77	0.78	0.77	0.76	1.32	1.15
	BLIP-2+SigLIP	0.79	0.78	0.79	0.78	0.75	1.26	1.01

Ablation study: Zero-shot

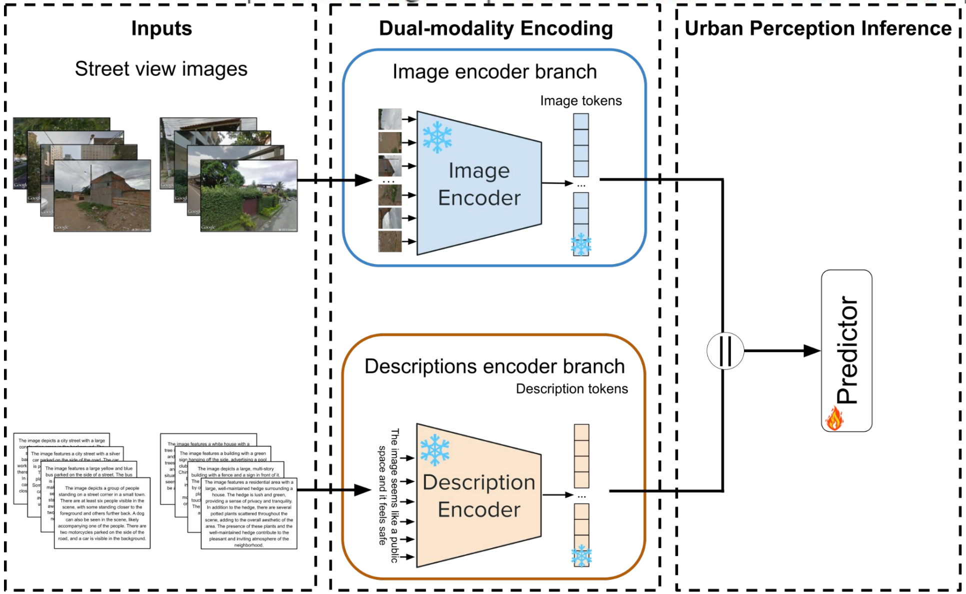
Ablation Study	Model Tested	Classification				Regression		
		Acc	Precision	Recall	F-1	R^2	RMSE	MAE
Zero-shot	CLIP	0.39	0.41	0.39	0.24	-14.05	4.53	4.89
	SigLIP	0.57	0.43	0.57	0.45	-14.17	4.61	4.77



We use the “positive” and “negative” description for each image.

Ablation study: Only heads

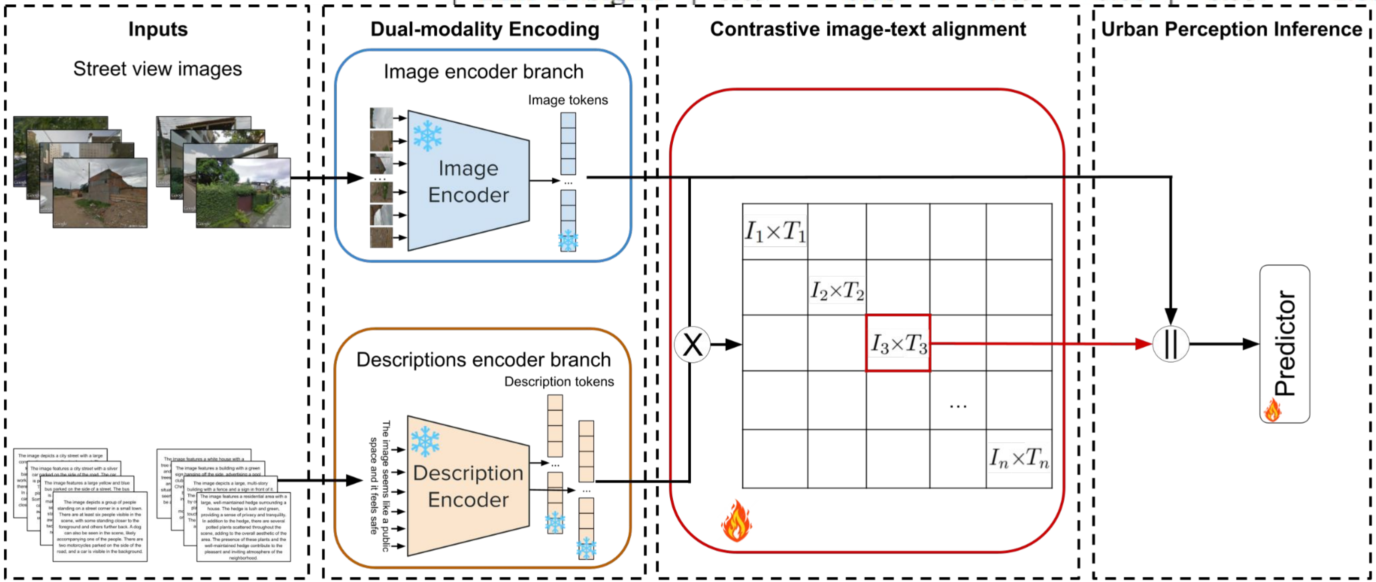
Ablation Study	Model Tested	Classification Acc	Precision	Recall	F-1	Regression R^2	RMSE	MAE
Only heads	LlaVA+CLIP	**0.67**	**0.67**	0.66	**0.66**	**0.57**	**2.43**	**2.56**
W/o description, contrastive & dual-modality	LlaVA+SigLIP	0.66	0.66	**0.67**	**0.66**	0.56	**2.43**	2.68
	BLIP-2+CLIP	0.63	0.61	0.62	0.61	0.53	3.4	3.21
	BLIP-2+SigLIP	0.64	0.63	0.63	0.63	0.53	3.38	3.35



We use the corresponding "positive" description (learns heads).

Ablation study: Contrastive

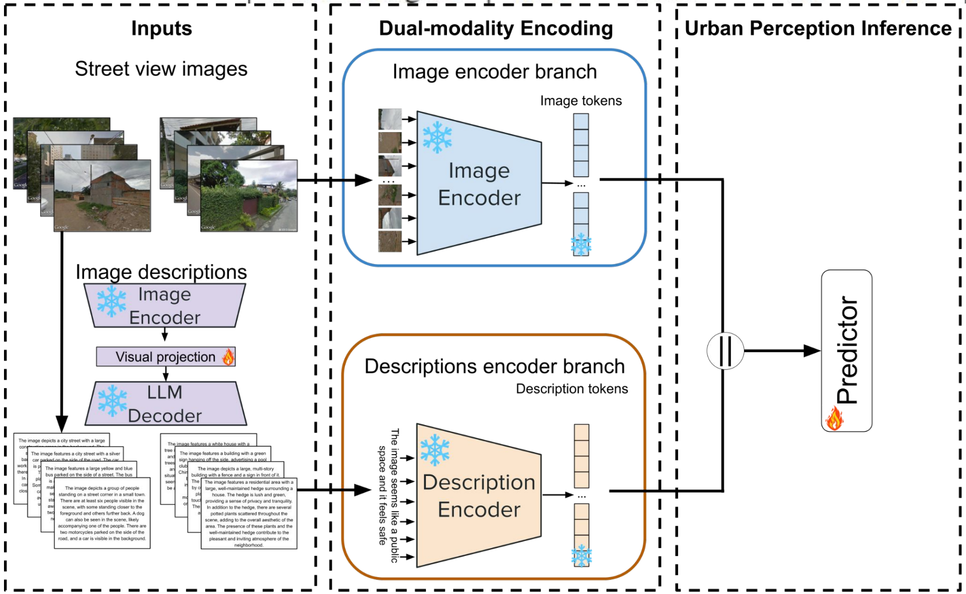
Ablation Study	Model Tested	Classification				Regression		
		Acc	Precision	Recall	F-1	R^2	RMSE	MAE
Contrastive W/o description & dual-modality	LlaVA+CLIP	0.7	0.69	0.68	0.68	**0.62**	**1.81**	1.95
	LlaVA+SigLIP	**0.71**	**0.71**	**0.7**	**0.7**	0.61	1.98	**1.84**
	BLIP-2+CLIP	0.68	0.67	0.68	0.67	0.56	2.75	2.35
	BLIP-2+SigLIP	0.69	0.68	0.69	0.68	0.55	2.68	2.2



We use the “positive” and “negative” descriptions (learns to match).

Ablation study: Visual projections

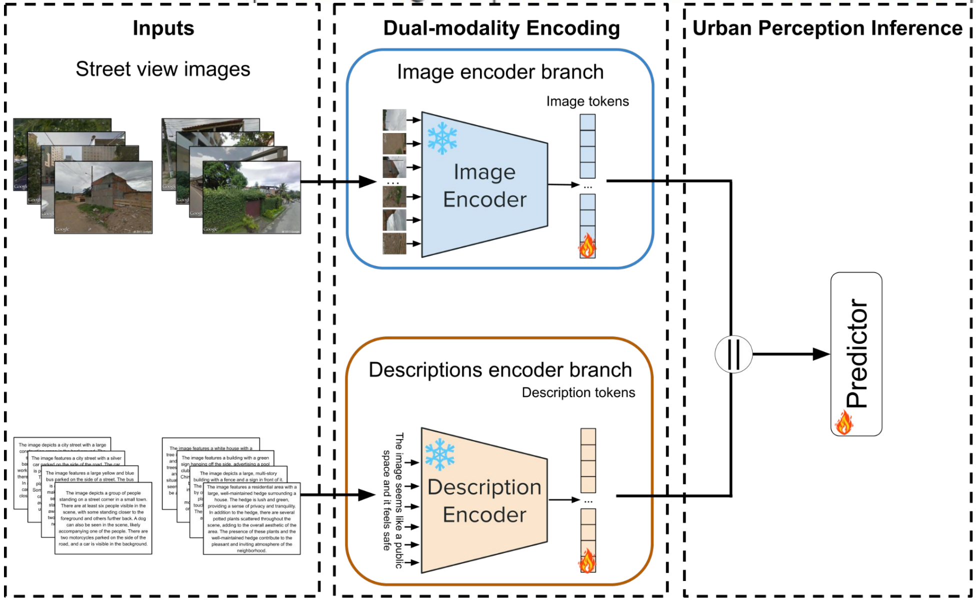
Ablation Study	Model Tested	Classification				Regression		
		Acc	Precision	Recall	F-1	R^2	RMSE	MAE
Visual projections W/o contrastive & dual-modality	LlaVA+CLIP	**0.73**	**0.72**	**0.71**	**0.71**	**0.67**	1.69	1.73
	LlaVA+SigLIP	0.72	**0.72**	**0.71**	**0.71**	0.65	**1.68**	**1.71**
	BLIP-2+CLIP	0.7	0.7	0.69	0.69	0.59	1.95	2.06
	BLIP-2+SigLIP	0.71	0.71	0.7	0.7	0.59	1.88	1.94



We refine the corresponding “positive” description (learns to describe).

Ablation study: Dual-modality

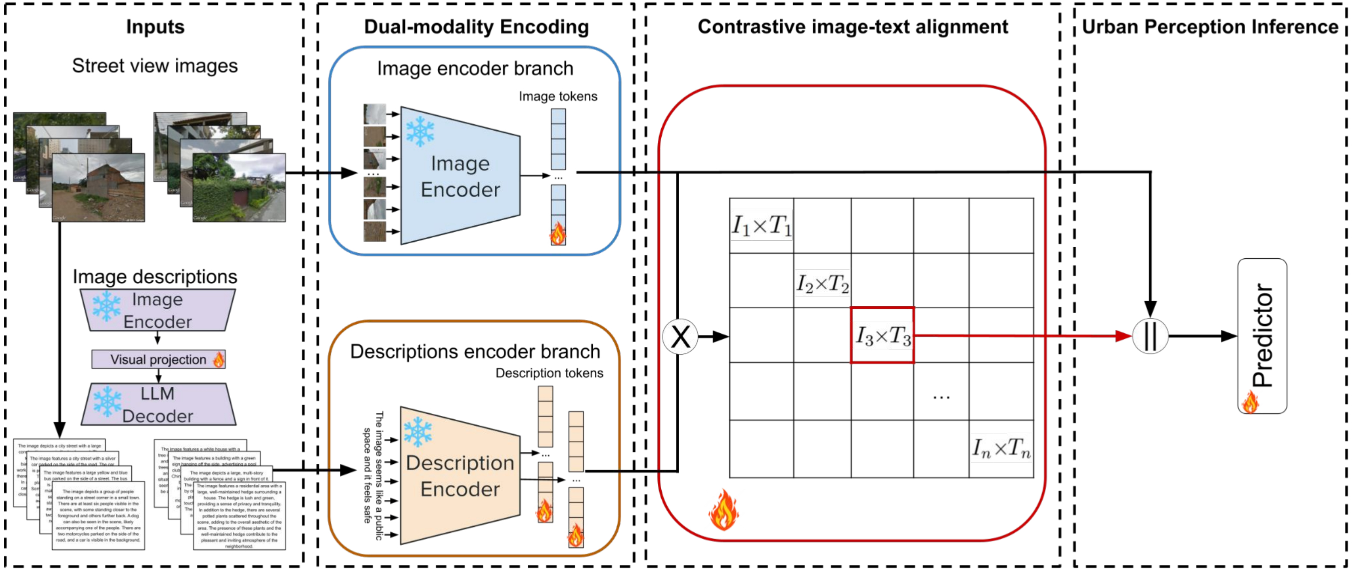
Ablation Study	Model Tested	Classification				Regression		
		Acc	Precision	Recall	F-1	R^2	RMSE	MAE
Dual-modality	LlaVA+CLIP	**0.76**	**0.76**	**0.75**	**0.75**	**0.78**	1.33	**1.42**
W/o description & contrastive	LlaVA+SigLIP	0.75	0.75	0.74	0.74	0.75	**1.29**	1.51
	BLIP-2+CLIP	0.72	0.72	0.73	0.72	0.69	1.6	1.34
	BLIP-2+SigLIP	0.73	0.73	0.72	0.72	0.68	1.4	1.21



We use the corresponding “positive” description (learns to encode).

Ablation study: UrbanVLM

Ablation Study	Model Tested	Classification				Regression		
		Acc	Precision	Recall	F-1	R^2	RMSE	MAE
UrbanVLM	LlaVA+CLIP	0.82	0.78	**0.79**	**0.78**	**0.84**	**1.04**	**0.78**
	LlaVA+SigLIP	**0.83**	**0.79**	0.78	**0.78**	0.83	1.08	0.79
	BLIP-2+CLIP	0.78	0.77	0.78	0.77	0.76	1.32	1.15
	BLIP-2+SigLIP	0.79	0.78	0.79	0.78	0.75	1.26	1.01



We use the “positive” and “negative” descriptions (learns all together).

Classification and regression results

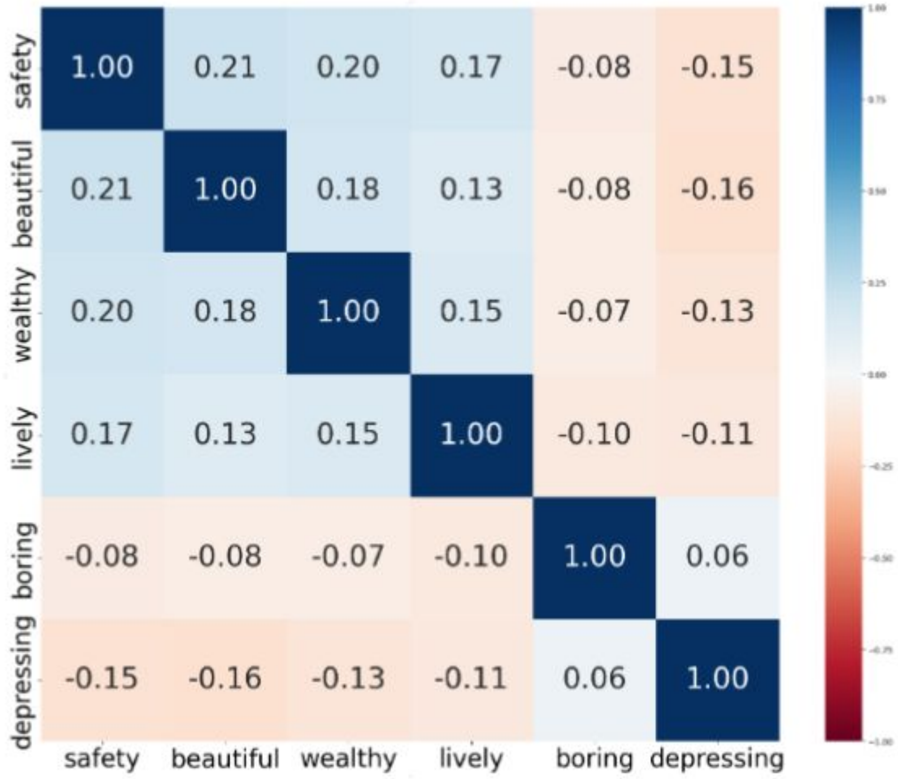
ACCURACY REPORT USING BINARY CLASSIFICATION IN SAFE CATEGORY

Model	Acc
PspNet+VGG [29]	48.38
DeepLabV3+VGG [29]	51.93
DSAPN+ResNet [54]	64.87
MTDRALN-LC [25]	65.07
MTDRALN-TC [25]	65.82
VGG+ImageNet [28]	65.72
VGG-GAP+ImageNet [28]	66.09
VGG+Places365 [28]	66.46
VGG-GAP+Places365 [28]	66.96
VGG19+ImageNet [4]	67.01
PSPNet+SVR [55]	70.63
DeiT+ResNet50 [40]	71.16
ViT-nn [27]	71.29
ViT-nn+OneFormer [27]	75.68
UrbanVLM (LlaVA+SigLIP)	82.55

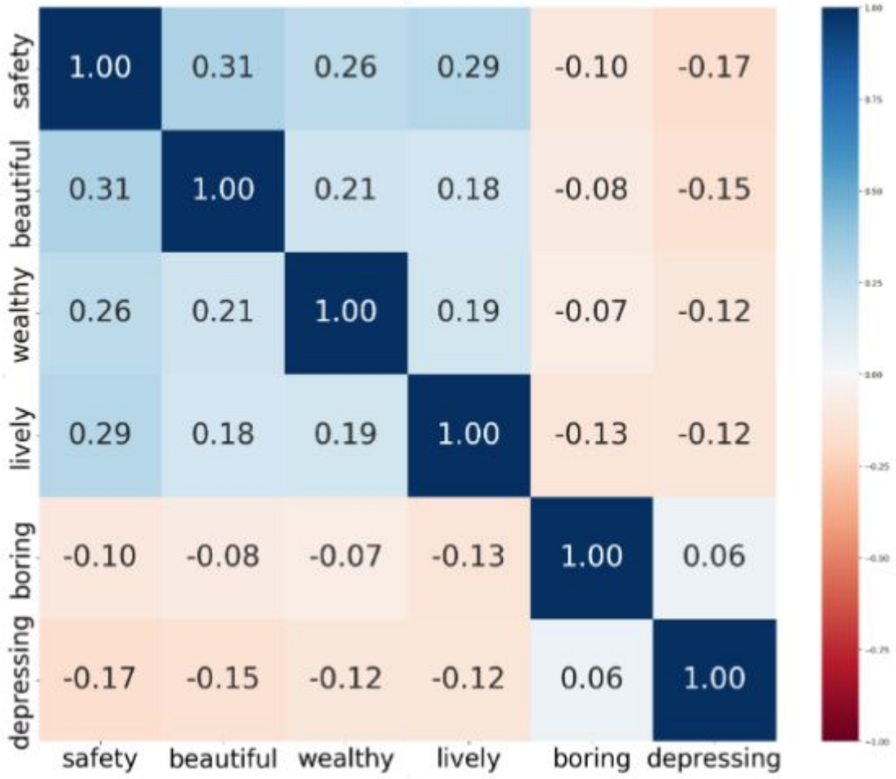
REGRESSION RESULTS IN SAFE CATEGORY

Model	R^2	RMSE
PSPNet-Regressor [55]	0.25	–
Fine-Tuned BERT [22]	0.42	–
FPN-based regressor [20]	0.52	–
DeepLabV3+ regressor [20]	–	2.16
DeepLabV3+ regressor [52]	–	2.91
SFB5+ConvNeXt-B+RF [60]	0.67	1.29
VIT+SegFormer+RF [11]	0.76	1.75
UrbanVLM (LlaVA+CLIP)	0.84	1.04

Scores comparison



(a) "Strength of schedule" algorithm



(b) UrbanVLM predictions

Conclusions

Conclusions

- We develop a VL-based model called **UrbanVLM**, aiming to improve binary classification and regression tasks.

- **Ablation studies:** The ablation results highlighted that fine-tuning image and text projection layers had the highest impact, while encoder layers contributed less to performance gains.
  - **Image-to-Text models:** Learns to refine descriptions.
  - **Dual-modality:** Learns to encode image and descriptions
  - **Contrastive Image-text alignment**: Learns to match image-text
  - **Heads:** Learns heads for each tasks.

- We **evaluate** the importance of **adding textual description** of images, by using MLLM models such as LlaVA and BLIP-2, we provide deeper context to our contrastive model.

THANKS!

Any Questions?