Assessing Urban Environments with Vision-Language Models

A **Comparative Analysis** of **AI-Generated Ratings** and **Human Evaluations**

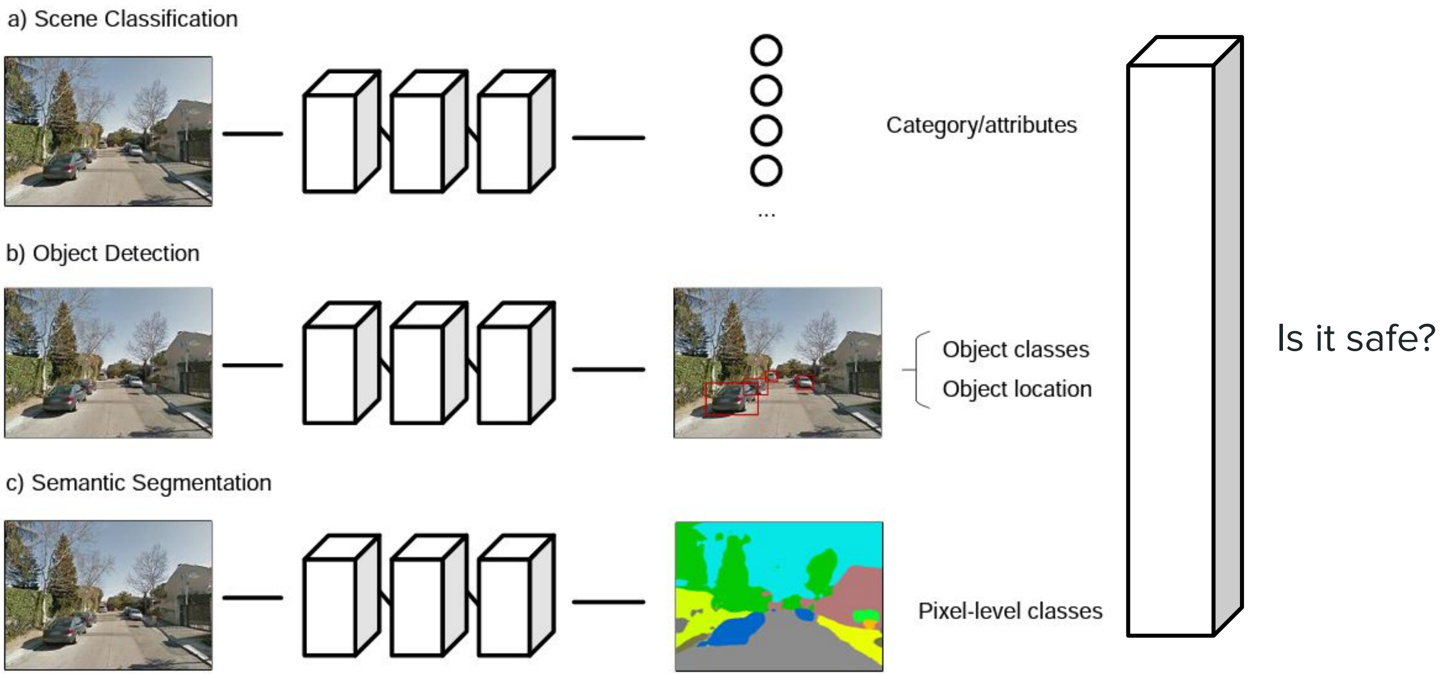
Felipe Moreno-Vera and Jorge Poco

Fundação Getulio Vargas / Visual Data Science Lab

www.visualdslab.com

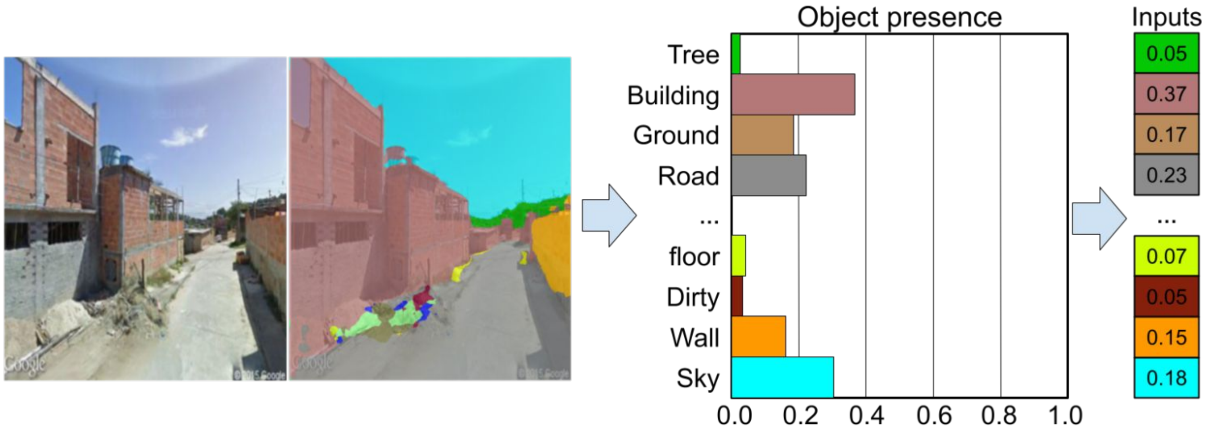
Context & Motivation

Techniques to analyze the urban perception



Hint: Street view images **contain rich, complex scenes** that are **difficult** to **interpret** purely through **pixels** or vector **embeddings**.

What information is obtained?



I have the pixel ratios of the objects, but what else?

- Traditional methods **identify objects** – e.g., they see "a building, sidewalks, roads"
- But they may **miss context** — e.g., they see “a **building**” but not "a **neglected building** with **graffiti** and **broken windows**."

How to obtain the rich information (context) from images?

The street's edges are **poorly defined**. On the right, there is a narrow, **rudimentary** concrete **sidewalk** and **minor cracks** visible in the **unfinished wall**. On the left, there is **no sidewalk** at all; the pavement gives way to dirt, sand, and piles of construction debris that spill onto the road. Besides, It is **littered** with **construction materials**, sand, and other **rubble**.

We have a description about the current status of the street

- Image descriptions adds **semantic depth** and **human-like perception**.

- Vision-Language models analyze the **context of the image** and provide further information.

Overview

- Place Pulse
- Image-to-Text descriptions
- UrbanVLM
- Ablation study
- Conclusions

Place Pulse

Place Pulse dataset

Which place looks livelier ?

For this question: **362,708** clicks collected

Goal: **500,000** clicks

SEE REAL-TIME RANKINGS

RANK	CITY	CLICKS	TREND
1	Washington DC	6296	
2	London	17982	
3	New York	22424	

RANK	CITY	CLICKS	TREND
54	Cape Town	16228	
55	Belo Horizonte	12728	
56	Gaborone	4717	

https://centerforcollectivelearning.org/urbanperception

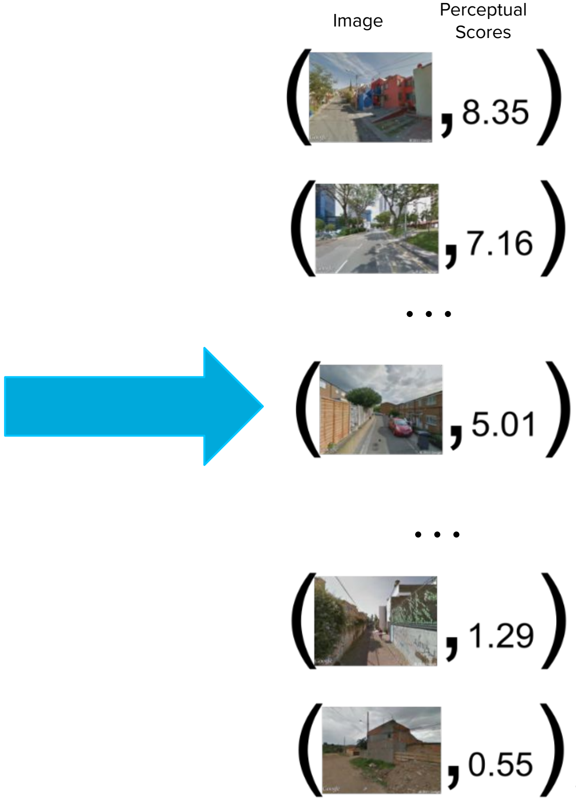
Cities included

- 1 223 649 Comparisons
- 111 390 images
- 32 countries
- 56 cities
- 6 categories:
  - Safety
  - Boring
  - Depressing
  - Wealthy
  - Lively
  - Beauty

Note: Same color means same country.

Strength of Schedule

left	right	winner
		draw
		left
		right
⋮	⋮	⋮
		right
		left



High safety scores images

Low safety scores images

Image-to-Text descriptions

Types of descriptions

- **General description:** Simple description of the image
- **Subjective description:** Based on the perception
    - **Positive:** Describe the image based on its corresponding perception.
    - **Negative:** Describe the image based on its opposite perception.

Image general descriptions

Image		
Model	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 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.	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 green residential area in the Philippines.

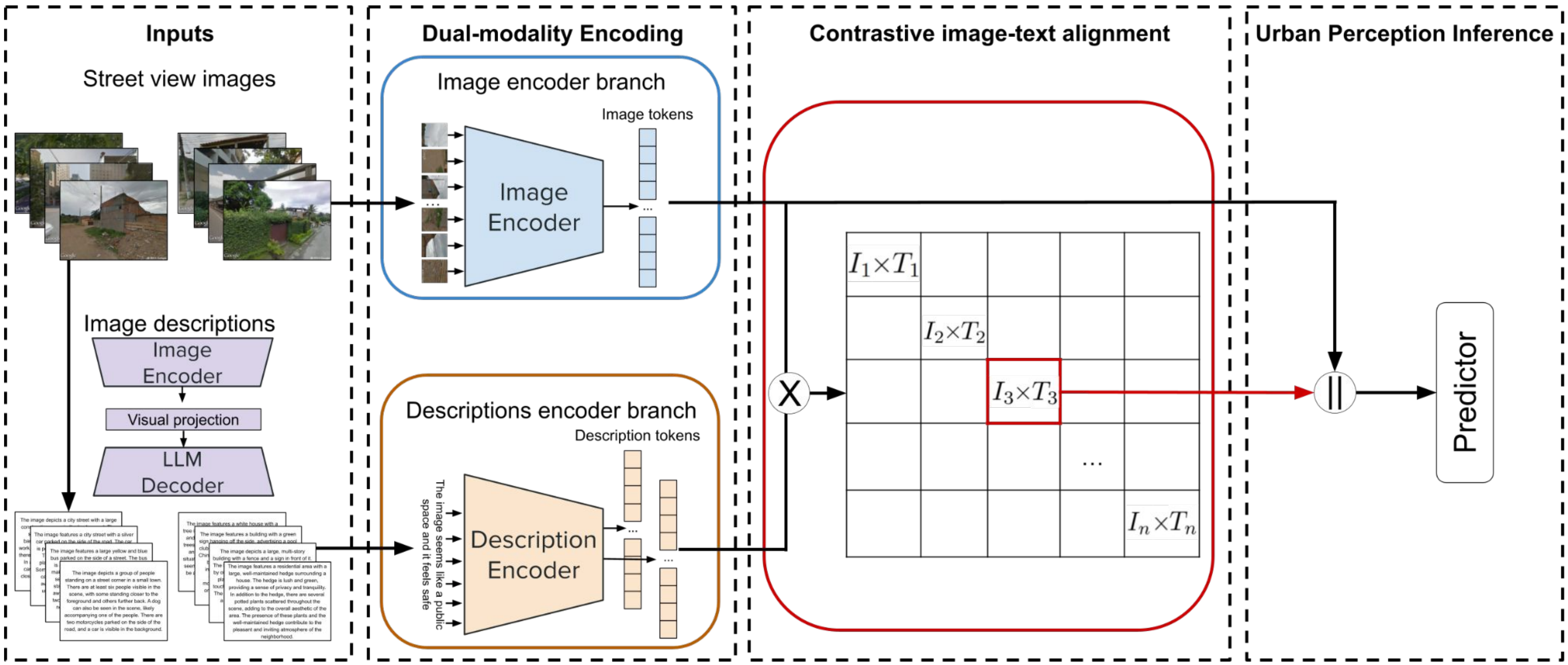
Image general descriptions

Image		
Model	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 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.	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 green residential area in the Philippines.~~

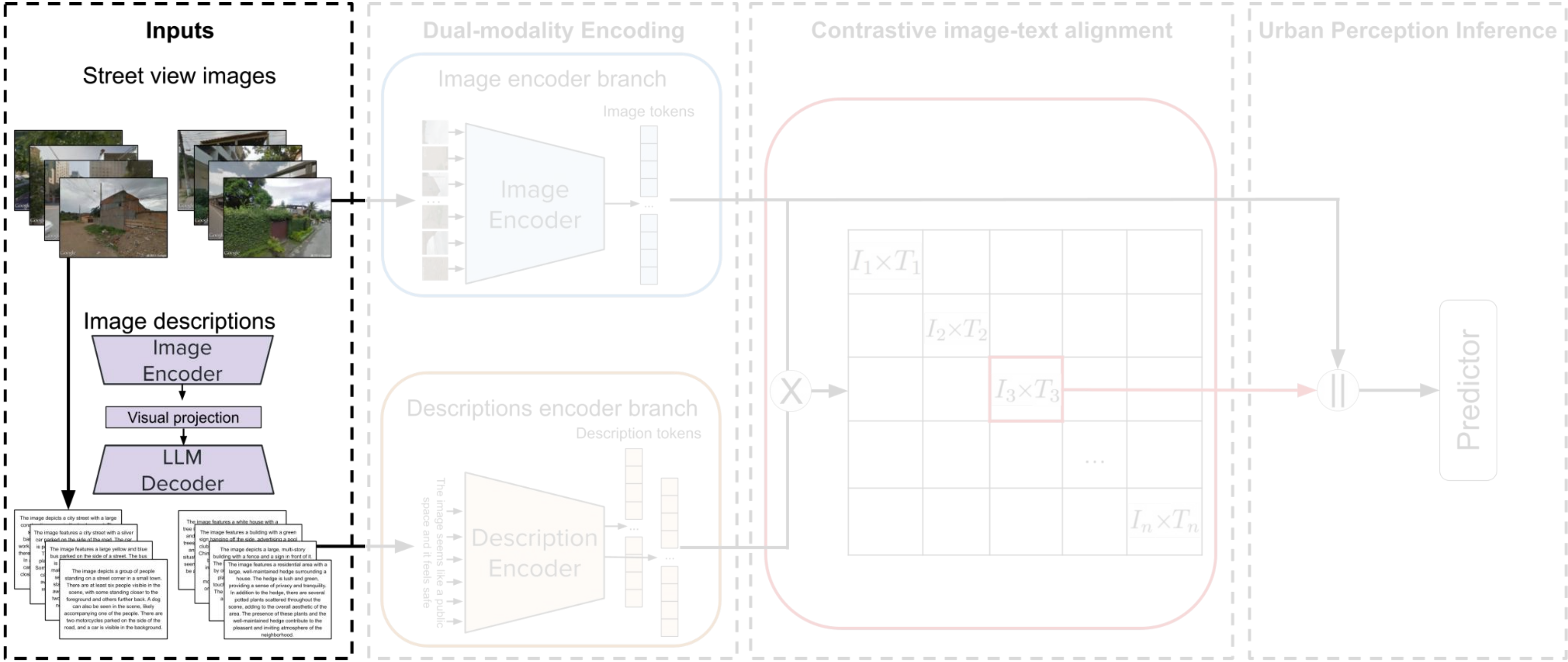
Randomly select 50 samples and compare the description results

Urban Vision-Language Model

Model architecture

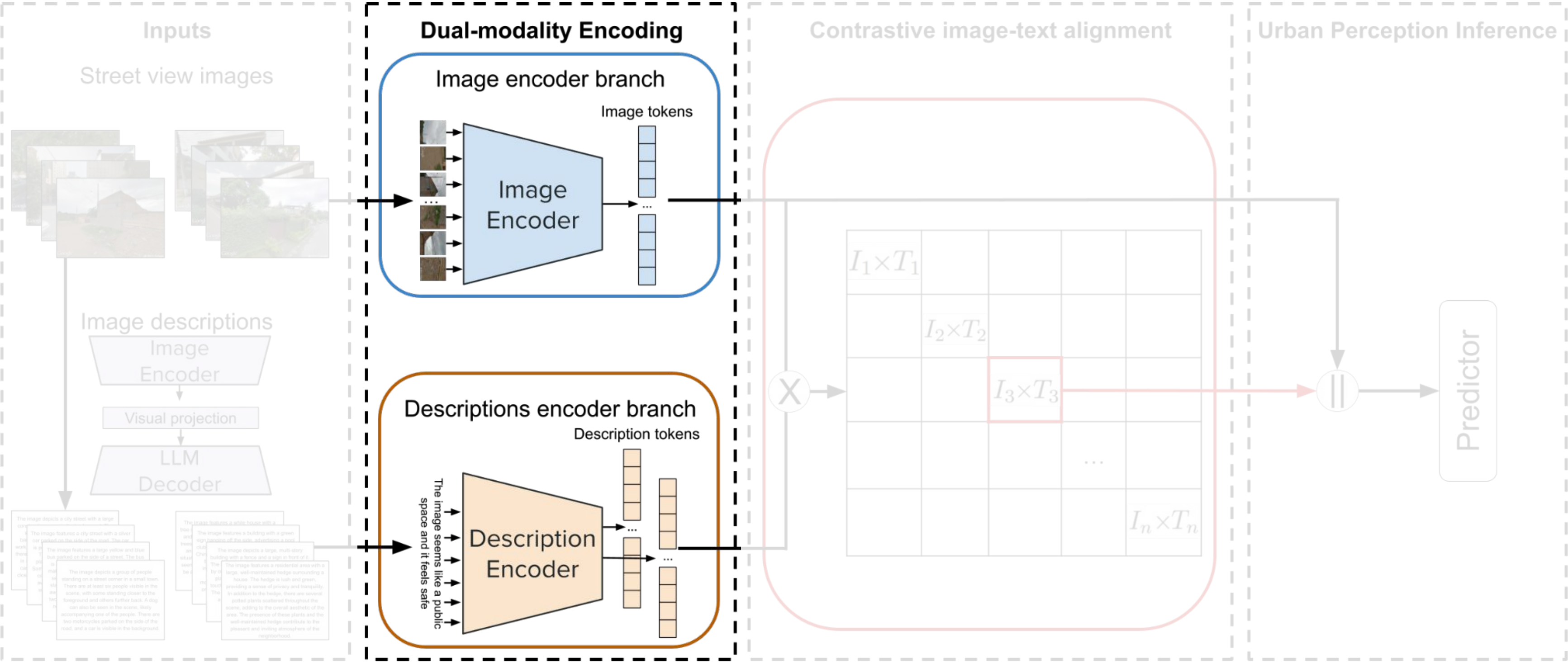


Model architecture - image descriptions



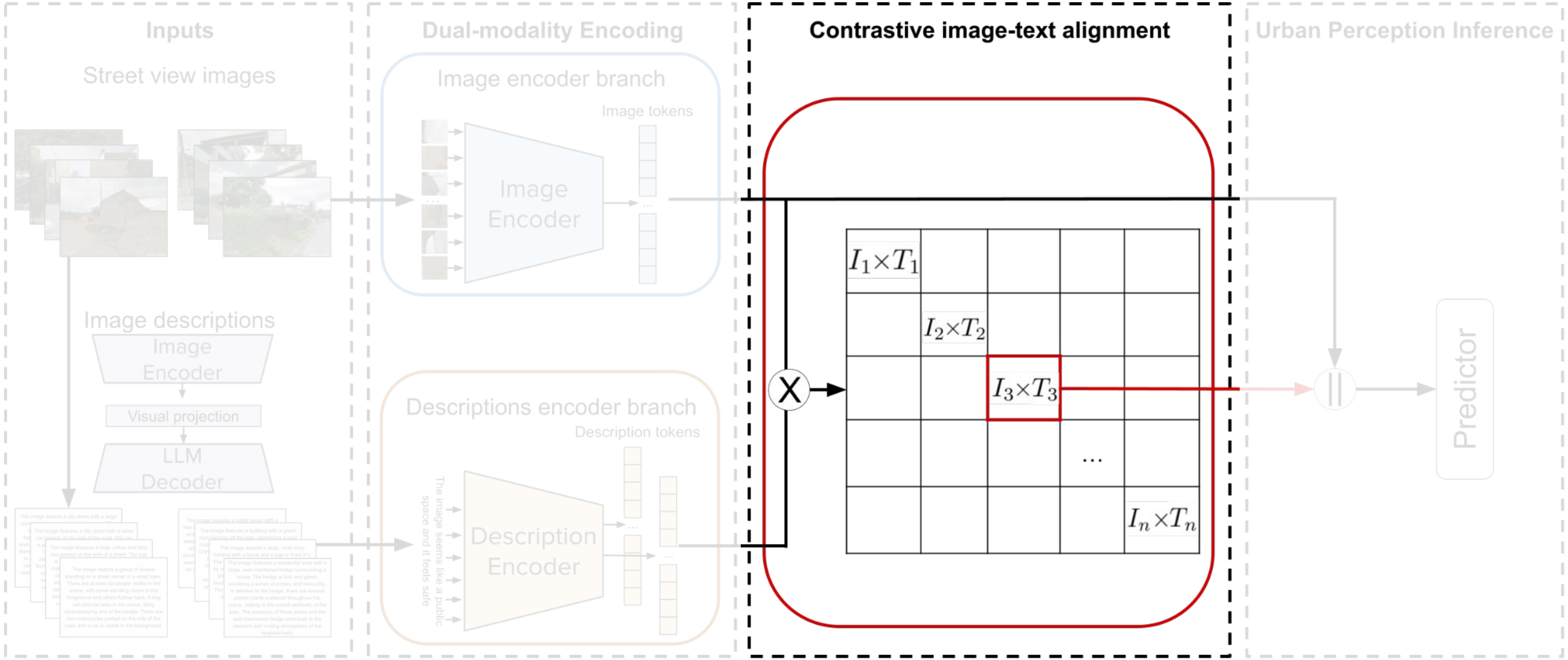
We use and compares **LlaVA** and **BLIP-2** performances

Model architecture - image-text encoders

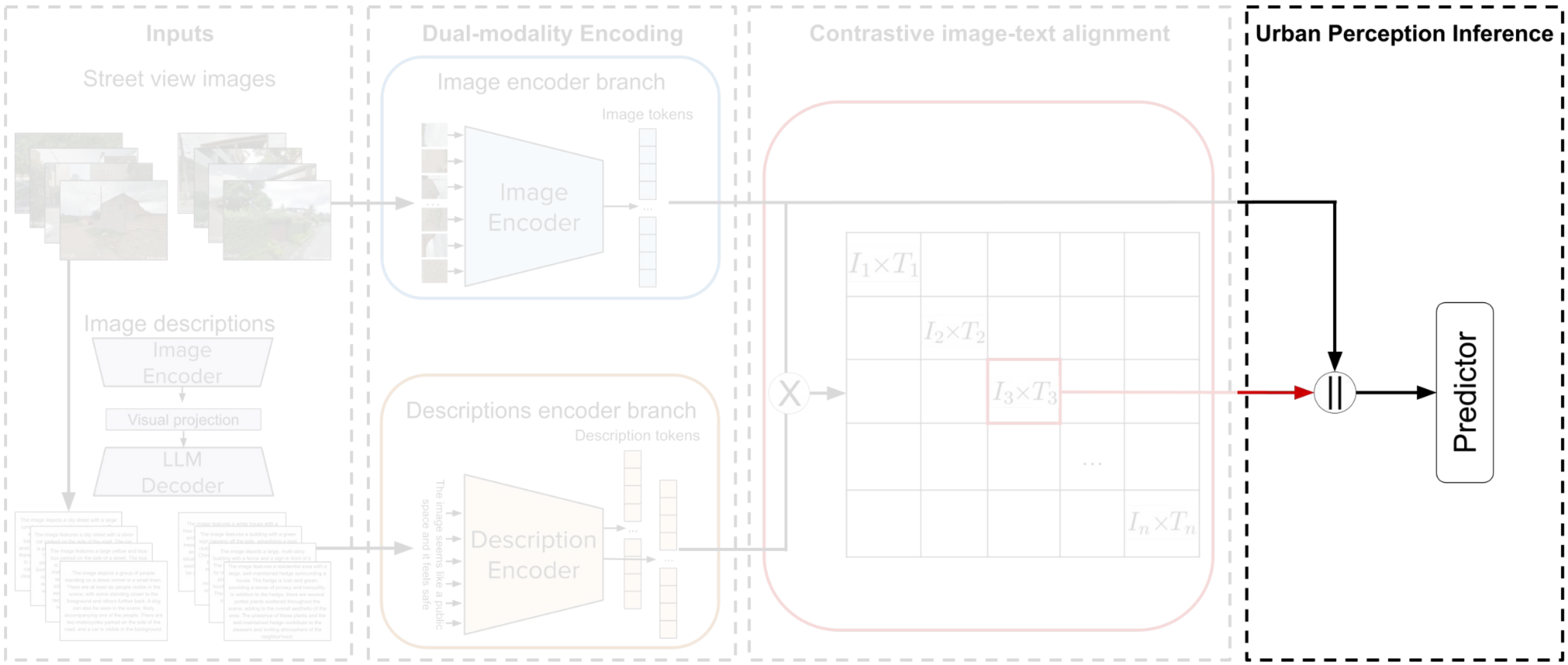


We use and compares **CLIP** and **SigLIP** performances

Model architecture - contrastive



Model architecture - heads



Classification and regression results

Classification

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

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

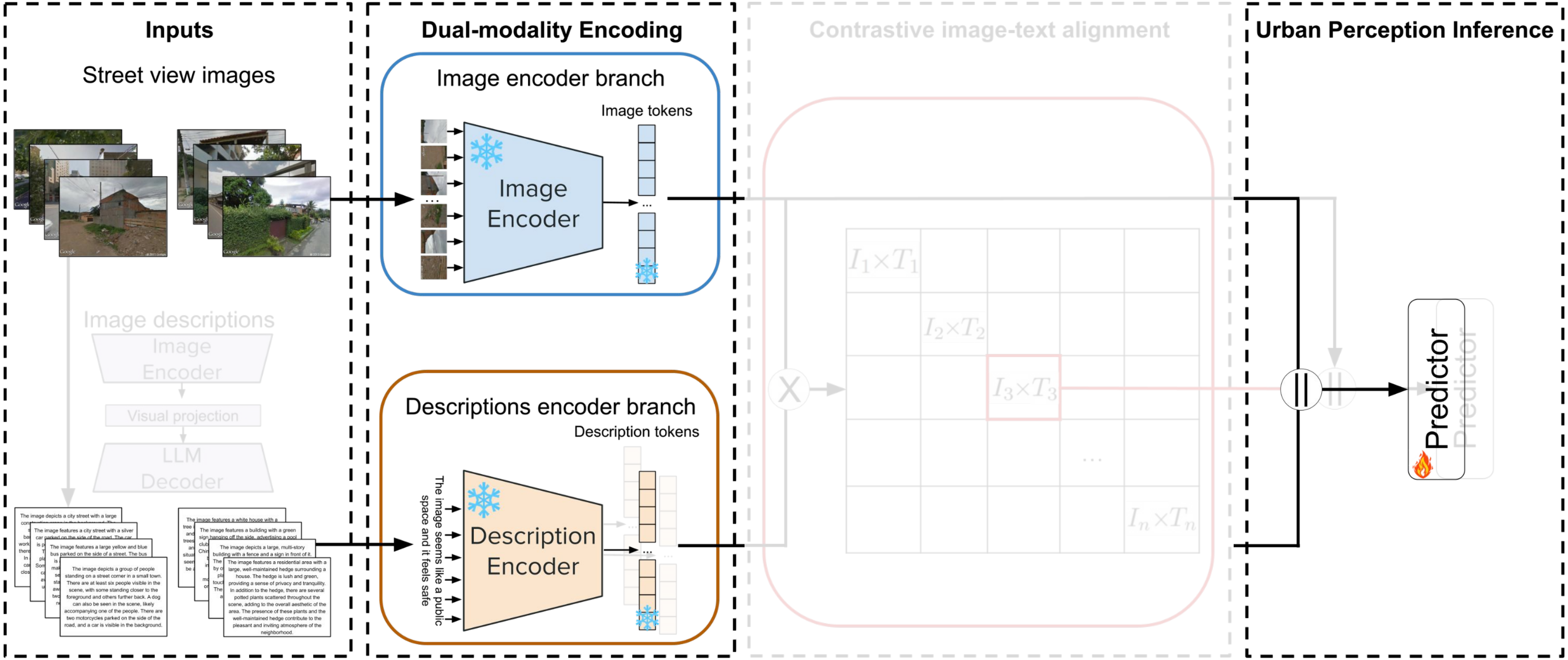
Ablation study

Components

We define 4 main components:

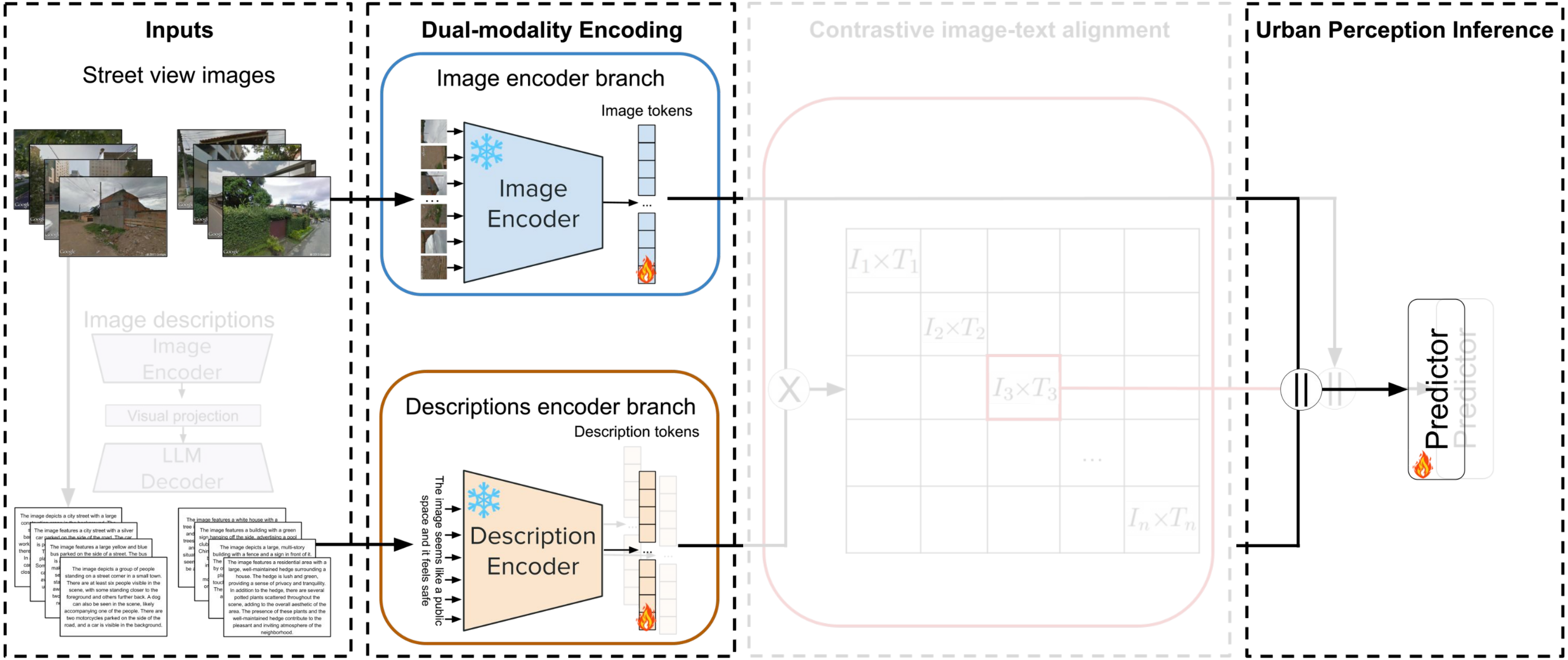
- **Heads:** Classification and regression MLP
- **Dual-modality:** Linear projections from Image and Text encoders
- **Image-to-Text:** Generates **positive** and **negative** descriptions
- **Contrastive Learning**: Image-text alignment

Only heads *(learns to classify and regress)*



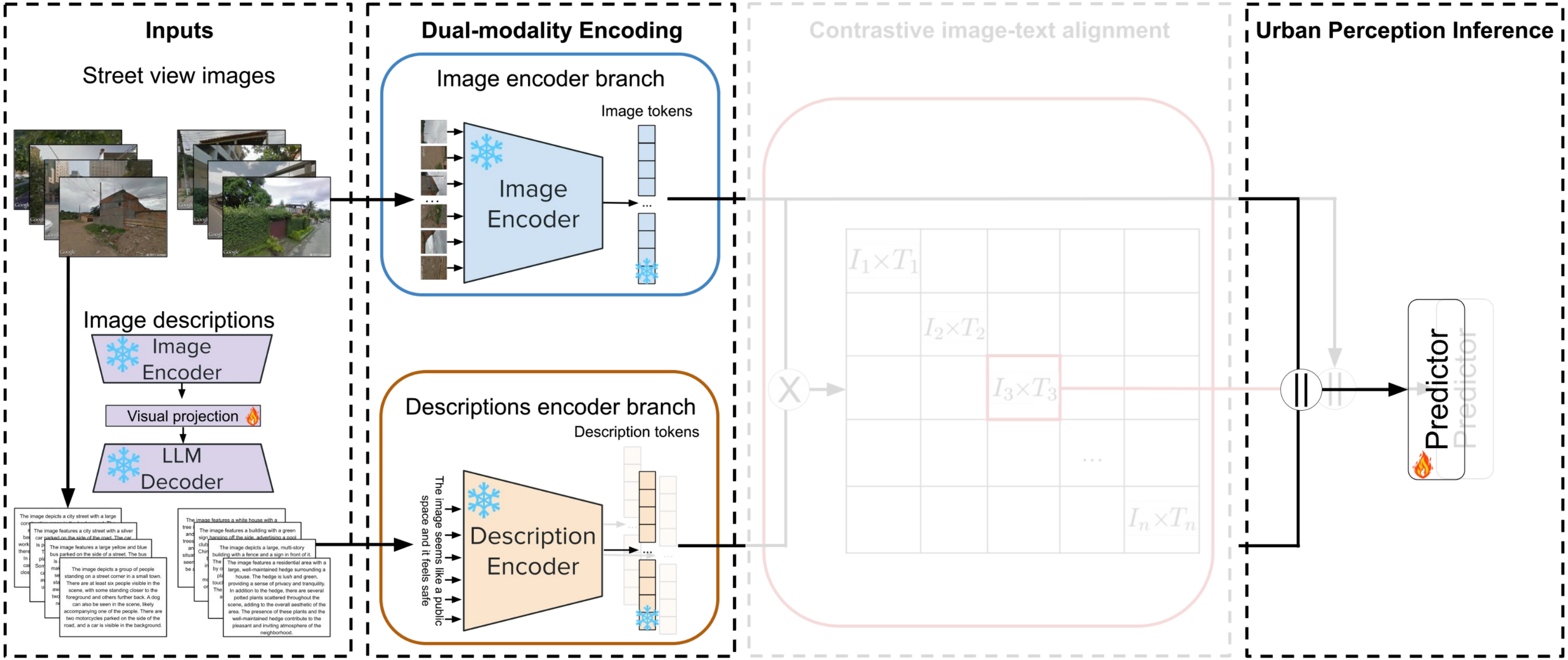
Use the corresponding **positive** description and concatenate.

Dual-modality *(learns to project image and text)*



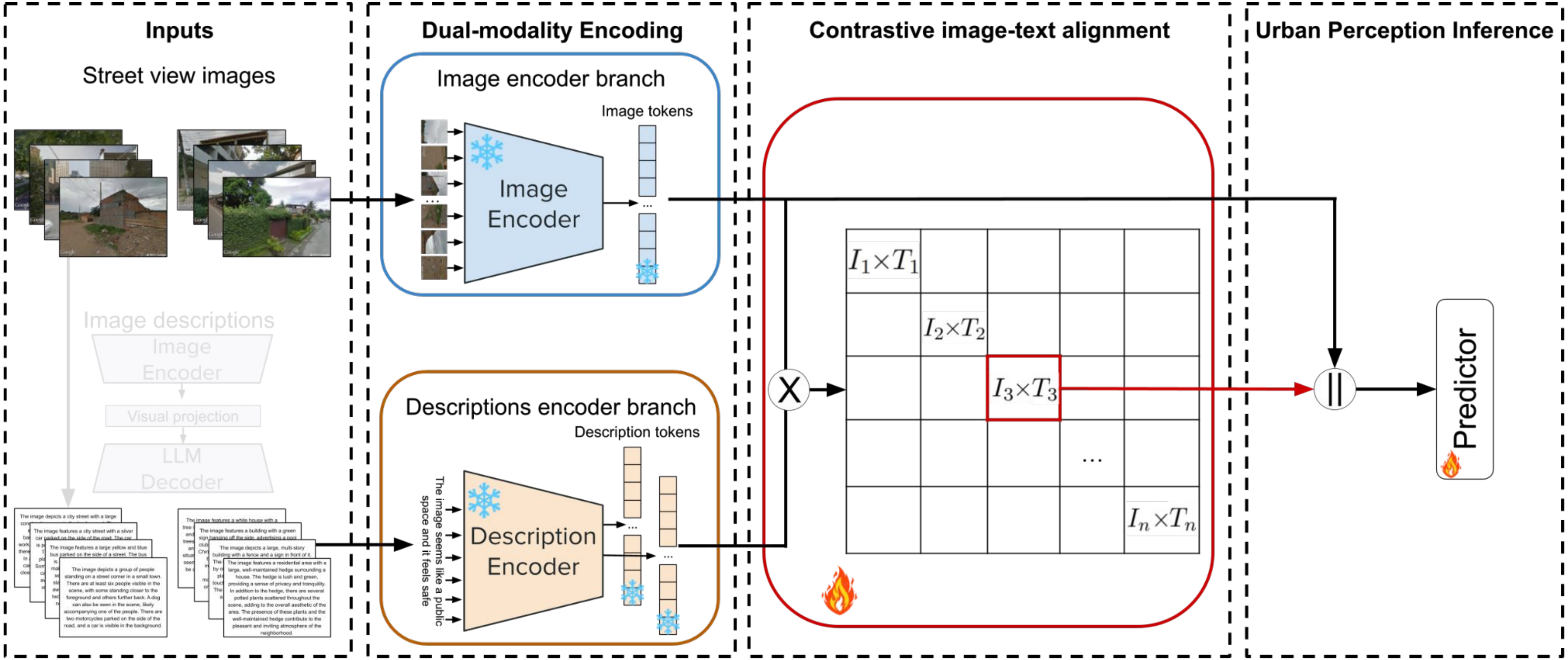
Use the corresponding **positive** description and concatenate.

Image-to-text *(learns to describe)*



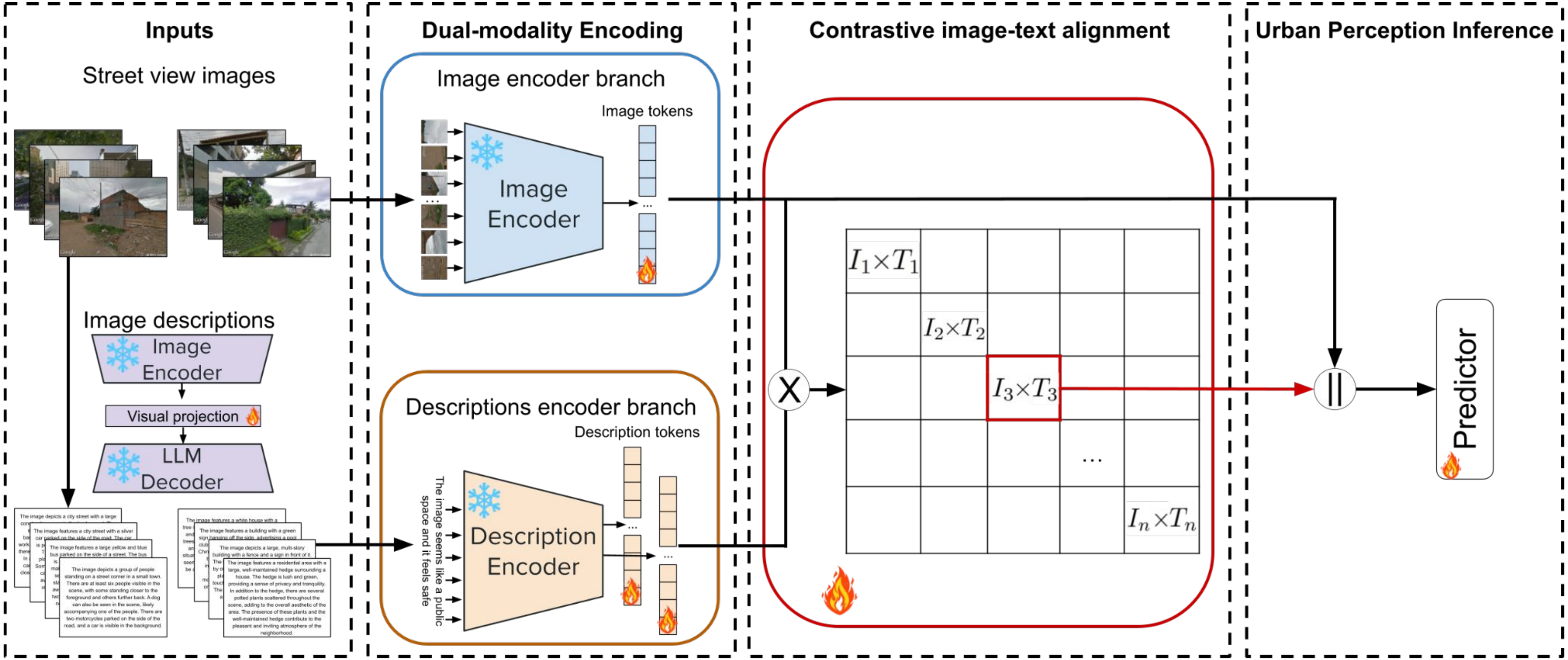
Improve the corresponding **positive** description and concatenate.

Contrastive *(learns to match image-text)*



Find the best match image- **positive** and **negative** descriptions, and concatenate.

UrbanVLM *(learns all together)*



Improve and match image-text descriptions and concatenate.

Conclusions

Conclusions

- **Ablation studies** allows to analyze and understand the relevance of each component in our model.
- **Adding robust descriptions** improve the urban perception inference of images giving a human-based perception descriptions.
- **UrbanVLM** successfully learns each component and improve the classification and regression tasks by using image descriptions.

THANKS!