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

1st Felipe Moreno-Vera
Getulio Vargas Foundation (FGV)
School of Applied Mathematics (EMAp)
Rio de Janeiro, Brazil
felipe.moreno@fgv.br

2nd Jorge Poco
Getulio Vargas Foundation (FGV)
School of Applied Mathematics (EMAp)
Rio de Janeiro, Brazil
jorge.poco@fgv.br

***Abstract*—This research investigates the application of vision-language models to automatically assess and rate street view images based on the Place Pulse 2.0 dataset, with a focus on comparing AI-generated ratings with human evaluations. The study introduces a context-sensitive rating system that assigns a 0-10 scale to six key urban perception categories: safety, liveliness, wealth, beauty, boredom, and depression. By comparing these AI-generated ratings with those of human volunteers, the research explores how effectively vision-language models can replicate human judgment in assessing urban environments. The findings provide valuable insights into the potential of vision-language models to scale urban perception analysis, offering an objective methodology that complements and enhances human evaluation. This approach not only contributes to urban planning by enabling more efficient, data-driven decision-making but also enriches the Place Pulse 2.0 dataset by integrating machine-generated ratings, paving the way for future advancements in urban perception studies.**

***Index Terms*—Vision-Language Models, Urban Perception, AI Rating System, Human Evaluation, Urban Planning.**

I. Introduction

Urban studies increasingly emphasize public perception, where subjective assessments of factors like safety, liveliness, and beauty influence urban planning and design [46], [56], [60]. Among these, safety perception has long been a central concern [31], [33], [35]. Early urban sociology, particularly the Broken Windows Theory [50], posits that visible signs of disorder, such as broken windows, contribute to a perceived lack of safety, leading to increased crime and environmental degradation [13], [39]. This theory has inspired numerous studies that expand on these ideas, linking physical environments to perceptions of crime and safety, and emphasizing how subjective safety experiences influence behavior and quality of life [21], [32], [45].

In recent years, the use of street view imagery (SVI) services has enabled researchers to collect data on human evaluations of urban perception through online platforms such as StreetScore [31], Wmodi [1], UrbanGems [36], Scenic-or-not [44], SenseCityVity [41], Places for play [14], and City-SAFE [5], among others. However, traditional methods of collecting these perceptions—primarily through human surveys or volunteer-based ratings—can be time-consuming, limited in scope, and difficult to scale [6], [56]. One prominent framework for capturing urban perception is the Place Pulse 2.0 dataset [9], [31], [38], which categorizes urban spaces across six key perception categories: safety, liveliness, wealth, beauty, and their opposites (e.g., not depressing, not boring).

Additionally, the rise of machine learning methods has spurred numerous studies, including studies using SVI with crime records [10], [47], graffiti presence [15], [48], demographic factors [4], [19], emotional perceptions [26], [29], landscape scenicness and beauty [36], [44], and deep learning-based approaches to estimate the winner between image comparisons [9], [25]. Recent advancements in vision-language models (VLMs) like LlaVA [18], BLIP-2 [17], CLIP [37], and SigLIP [53] have shown great potential in bridging visual content and natural language understanding. By leveraging large-scale datasets of images paired with textual descriptions, these models enable human-like comprehension of visual data, opening new possibilities for applications in urban perception.

Contributions This research explores the use of vision-language models for image caption generation, aiming to categorize and rate street view images. Our contributions are as follows: (i) We propose a methodology to leverage the benefits of vision-language models for describing, rating, and classifying street view image perception. (ii) We evaluate the impact of incorporating image descriptions and contrastive learning into urban perception tasks by using vision-language models across six urban perception categories. (iii) We develop an Urban Vision-Language Model (UrbanVLM), which unifies street view images, contextual information, and descriptions to achieve accurate predictions of human perception ratings. For supplemental material and source code [1].

[1]http://www.visualdslab.com/papers/UrbanVLM

II. Related works

A. Urban Perception and Computer Vision

Urban perception plays a key role in urbanism and planning, focusing not only on creating accurate prediction models [28], [41] but also on understanding the urban environment and its effects on residents [7], [49]. The primary goal is to develop models that capture a city's visual identity and define its uniqueness. Questions like "*What makes Paris look like Paris?*" [8], "*What makes an outdoor space beautiful?*" [44], "*What makes London appear beautiful, quiet, and happy?*" [36], and "*What makes a place feel safe?*" [27] are central to this research. Additionally, some studies incorporate supplementary data, such as crime and robbery statistics [2], [43].

The MIT Media Lab introduced a significant dataset in urban perception, the MIT Place Pulse dataset [31], [38]. This dataset inspired researchers to map urban perception scores, and feature extraction techniques such as GIST, DeCAF, and ImageNet were used to train image representations along with their respective perceptual scores [31], [33]. Other studies have sought to extract more detailed information about the visual appearance of images using complex methods, such as convolutional neural networks (CNNs) [9], [35]. Additionally, segmentation techniques have been employed to analyze the presence of visual elements and their correlation with safety perception [51], [57] or apply explanation methods to understand the relationship between model predictions and human perception [25], [30].

B. Multimodal Models in Urban Perception

Although previous studies have applied computer vision techniques to urban analysis, including the addition of information such as text captioning using LSTM [23] and BERT [22], or combining street view images with comments obtained from social networks (e.g., Twitter) [42]. Recent research has begun leveraging multimodal models (see Appendix A for definition). For instance, some studies employ models like Siamese networks and GPT-4V to compare SVI and rank them [59]. [24] compares the visual appeal and functionality of streets using the GPT-4 model alongside human evaluations. Similarly, [12] utilizes the CLIP model to infer urban street functionality through zero-shot learning. Furthermore, [16] builds on the Scenic-or-Not dataset [44], [58], extending it with manually added image annotations to infer scenicness using the CLIP model.

However, our work stands out by leveraging an ensemble of vision-language models to integrate image-text generation with urban perception evaluations. Using designed prompts, we generate descriptions that capture visual appearance descriptions, then fine-tune them alongside images with multimodal models. This approach enables robust classification of urban perception in street view imagery (SVI), aligning model predictions with human evaluations to provide deeper insights into urban spaces.

III. Methodology

Our methodology is composed of the following steps: (i) human evaluations quantification; we begin with an exploratory data analysis, quantifying urban perception scores derived from the human evaluations; (ii) image descriptions generation, by using vision-language models we design prompts to obtain insights from the visual appearance of the image evaluated; (iii) UrbanVLM, this model integrates an image-text generation model and a contrastive language-image model to classify the perceived safety of street images.

A. Human evaluations quantification

We study the MIT Place Pulse 2.0 dataset, which contains approximately 1.22 million pairwise comparisons across 111,390 images from 56 cities, including image IDs, coordinates, and the respective winners. Following previous works, we focus our evaluation on the safety category. To preprocess comparisons and assign perceptual scores to each image, we apply the "strength of schedule" algorithm [34], which estimates the Q-score using the win rate and loss rate for each image with pairwise comparisons from the following equation:

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

$$Penalty_i^k = \frac{1}{l_i^k} \sum_{j=1}^{n_2} \frac{l_i^k}{w_i^k + d_i^k + l_i^k} \tag{2}$$

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

In the Equations 1 to 3, w_i^k, d_i^k, and l_i^k represent the number of times image i has been selected as the winner, equal, or loser respectively; n_1 and n_2 represent the number of times image i wins and loses a comparison; $Award$ is the average win rate where image i won the comparison; $Penalty$ is the average loss rate where image i lost the comparison. The final Q-score is scaled to fit a range from 0 to 10, where an image with a score close to zero is perceived as very unsafe, and a score close to 10 is perceived as very safe [31], [32], [38].

Additionally, unlike previous works, we conduct a data exploration analysis and find that about 2,471 coordinates have multiple image IDs, indicating repeated comparisons at the same duplicated location and leading to sample imbalances across cities (see Appendix B).

B. Street view imagery captioning

Since Place Pulse 2.0 lacks volunteer information (e.g., gender, age, location, nationality) and images have no descriptions, we extended it by generating captions for the 111,390 images using vision-language models, referred to as image descriptions. Our goal is to assess how these descriptions enhance AI-driven urban perception. We use and compare BLIP-2 and LLaVA to generate two types of descriptions: (i) visual appearance and (ii) evoked feelings (e.g., safety, boredom, depression). BLIP was also evaluated but produced lower-quality descriptions. See Appendix C for prompt details.

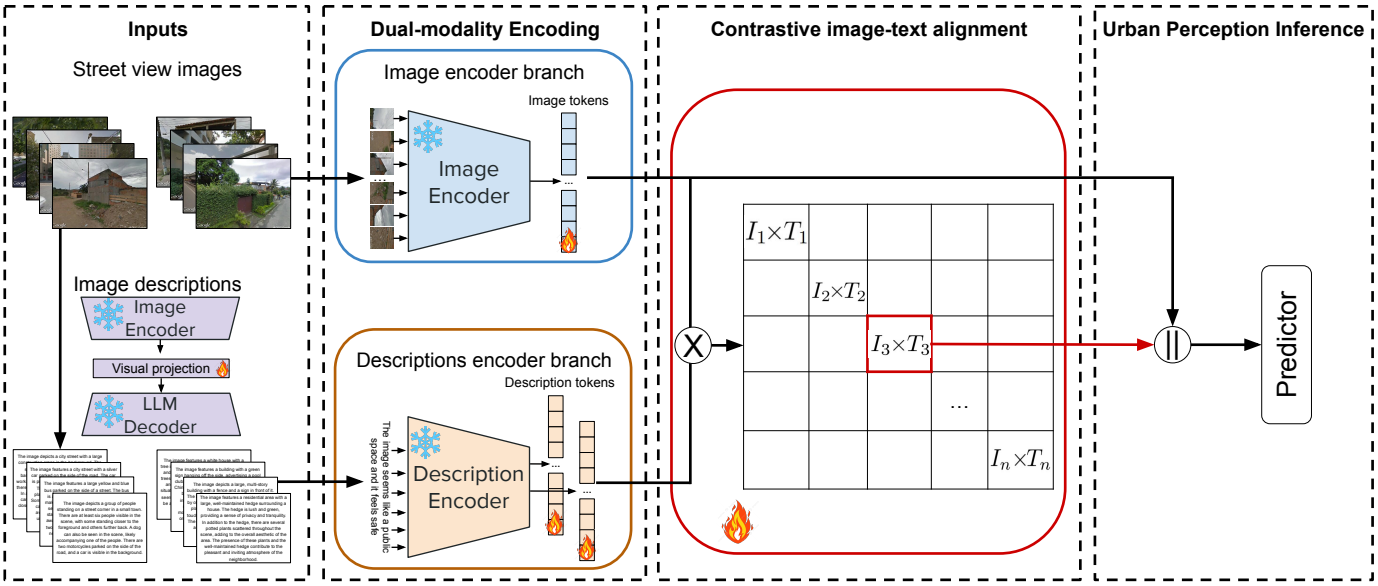


Fig. 1. Our proposed UrbanVLM multimodal model comprises four main components: (i) image description generation using a vision-language model with freeze layers except for the visual projections such as LlaVA and BLIP2, (ii) dual-modality encoding using vision-text encoding models such as CLIP or SigLIP, (iii) contrastive learning to align image-text descriptions, and (iv) classification and regression heads to infer the urban perception labels and scores.

C. UrbanVLM

We present UrbanVLM, a multimodal framework for urban perception, using the Place Pulse 2.0 dataset to classify and regress the perception of street view images. Our approach leverages large pre-trained models for image description generation (LLaVA and BLIP-2) and contrastive learning (CLIP and SigLIP).

As illustrated in Figure 1, UrbanVLM consists of four stages: (i) **Image description**: images are processed through a text generation model; (ii) **Dual-modality encoding**: images and descriptions are embedded into a shared latent space and either combined or used for contrastive training; (iii) **Contrastive image-text alignment**: the contrastive model identifies which description better matches the image; and (iv) **Classification and regression**: the model is fine-tuned for binary classification (e.g., safe vs. unsafe) and perception score regression (0–10 scale).

Mathematical Formulation

Let $\mathcal{D} = \{(I_i, y_i)\}_{i=1}^{N}$ denote the dataset, where: $I_i \in \mathbb{R}^{H\times W\times 3}$ is the i-th image and the label $y_i \in [0, 10]$ (for regression) or $y_i \in \{0, 1\}$(for classification).

We use a vision-language model function f_{vlm} to generate an intermediate visual encoding $V_i \in \mathbb{R}^V$ and the corresponding textual description $T_i \in \mathbb{R}^L$ (L is the token length), where T_i^+ means a positive description and T_i^- means a negative description.

Then, for each tuple of (I_i, T_i^+, T_i^-), we generate encodings using a contrastive model. The encodings are f_i and f_t. Where $f_i(I_i) \in \mathbb{R}^m$ is the image encoding and $f_t(T_i^+)$ and $f_t(T_i^-) \in \mathbb{R}^n$ are positive and negative encodings. The objective is to use contrastive learning to learn a shared representation between the image and the two descriptions (positive and negative), and then use classification learning to predict the label (”safe” or ”not safe”) based on the image and descriptions, following a similar approach for regression.

Contrastive learning: To align the image and text encodings, we apply contrastive loss (InfoNCE loss), which ensures that related image-text pairs are close together in the shared latent space, while unrelated pairs are far apart. The contrastive loss is given by:

$$\mathcal{L}_{\text{constrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_{\text{i}}(I_i), f_{\text{t}}^{+}(T_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_{\text{i}}(I_i), f_{\text{t}}(T_j))/\tau)}$$

Where $\text{sim}(f_i, f_t) = \frac{f_i^\top f_t}{\|f_i\|\|f_t\|}$ is the cosine similarity and τ is a temperature parameter that controls the smoothness of the softmax function with a value of 0.07.

Classification task: We use the shared representations from image and text encodings to predict the label. The image and text encodings are concatenated and passed through a classification head, which outputs a probability distribution over the two classes. The classification loss is computed using the cross-entropy loss function:

$$\hat{y}_i = W_{\text{i}} \cdot [f_{\text{image}}(I_i) \| f_{\text{text}}(T_i)] + b_{\text{i}}$$

$$\mathcal{L}_{\text{class}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(Tanh(\hat{y}_i))$$

Regression task: We predict the continuous perception scores using the concatenated image-text encodings and the original continuous values $\hat{y}_i$. The regression loss is computed using the mean squared error (MSE):

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

TABLE I
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

Total Loss Function: The total loss function is a weighted sum of the contrastive loss, classification loss, and regression loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \lambda_{\text{class}}\mathcal{L}_{\text{class}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}$$

where λ_{class} and λ_{reg} are hyperparameters controlling the contribution of each loss.

IV. EXPERIMENTS & DISCUSSIONS

In this section, we conduct extensive experiments to investigate the following Research Questions (RQ):

- **RQ1**: Can UrbanVLM outperform previous baseline methods (classification and regression)?
- **RQ2**: How does each component (e.g., image-text generation and contrastive learning) contribute to UrbanVLM?
- **RQ3**: How does automated text generation impact the results?

A. Experimental settings

We incorporate two widely used open-source multimodal models, LlaVA (llava-1.5-7b) and BLIP-2 (blip2-opt), available from Huggingface, to generate diverse and descriptive textual representations of street view images, focusing on features relevant to urban perception.

For contrastive learning, we use CLIP with a ViT-B/32 backbone and SigLIP with the SoViT-400m architecture. For all experiments, the dataset is split into 75% for training and validation and 25% for testing. Experiments are conducted on an NVIDIA RTX 3090 GPU with limited VRAM. To optimize training, we use float16 precision, freeze lower layers, and fine-tune only the higher layers, including the classification and regression heads.

B. RQ1: Performance evaluation

To validate the effectiveness of the proposed UrbanVLM, we compare its performance with that of previous models. Table I reports the binary classification accuracy for the safe category. Notably, most prior work reports only accuracy, disregarding other performance metrics. As shown in the table, UrbanVLM significantly surpasses baseline methods employing single-granularity models, achieving substantial improvements. It outperforms the best baseline, [27], by 6.87%.

TABLE II
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

Table II reports the R^2 and RMSE metrics, which are the most commonly used regression metrics in previous studies. We observe that UrbanVLM outperforms the best prior model [11] by 0.08 in R^2. These results demonstrate that multi-granularity approaches generally outperform single-granularity models. This improvement can be attributed to the incorporation of fine-grained information derived from street view image descriptions, which enriches the learning process.

C. RQ2: Ablation studies

We conduct ablation studies to investigate the effectiveness of different components in UrbanVLM on the Place Pulse dataset safety category, including text generation, contrastive methods, and their absence. Specifically, we evaluate the performance of training the visual projections in the text generation model, the contrastive method, and both combined. In Table III, the term *zero-shot* indicates that we first generate descriptions for images and test the model without any additional training; here, CLIP and SigLIP are used for inference. For this purpose, we use LLaVA and BLIP-2 to generate image descriptions and define zero-shot prompts to infer the perceptual score and the corresponding category (e.g., safety, boring) of the street view images. See Appendix D for more details about these prompts.

The term **w/o description & contrastive** refers to *Dual-modality encoding*, which indicates that we use the generated descriptions and concatenate the image-text encodings to train the classification and regression heads, as well as the image and description token projections, without fine-tuning the text generation or contrastive learning components. The term **w/o contrastive & dual-modality** refers to the *Visual projections* of the text generation model, which means that we do not perform any contrastive learning on the encodings, nor do we train the projections (i.e., image and description tokens).

The term **w/o description & dual-modality** refers to the absence of *description generation* and *dual-modality encoding*. In this setup, we focus solely on training the image-text alignment for contrastive learning to determine the best match, and only the classification and regression heads are trained.

The term **only heads** means that we only train the classification and regression heads. Finally, the term **UrbanVLM** refers

TABLE III
ABLATION STUDY ON URBANVLM PLACE PULSE DATASET

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

to the training of all components: visual projection, image and description tokens, contrastive learning, and heads.

Table III provides a detailed analysis of the impact of various techniques on performance across both classification and regression tasks. The results indicate that integrating contrastive learning into the heads leads to an average improvement of 3% in classification accuracy. This enhancement is accompanied by significant improvements in regression metrics, highlighting the versatility of this approach. Furthermore, incorporating visual projections from text generation yields a substantial 6% average improvement in classification accuracy, suggesting that these projections help capturie richer features from the multimodal data. The most pronounced improvement, however, comes from the dual-modality approach, which leads to a 9% average increase in accuracy across both tasks, underscoring the value of combining visual and textual information for improved model performance.

In terms of task-specific performance, contrastive learning is found to particularly benefit classification tasks when used in conjunction with SigLIP, thereby enhancing the model's ability to distinguish between classes. On the other hand, CLIP excels in regression tasks, as evidenced by its superior performance in these metrics. Interestingly, the relationship Root Mean Square Error (RMSE) $\approx$ Mean Absolute Error (MAE) suggests that the model's errors are likely distributed more uniformly, rather than being influenced by large, skewed outliers.

When examining the optimal model combinations for specific tasks, we observe that LlaVA+SigLIP performs best for classification tasks, suggesting that the fusion of these two components provides the most effective model for distinguishing between different classes. For regression tasks, however, LlaVA+CLIP outperforms all other configurations, highlighting the particular strength of this combination in predicting continuous values.

D. RQ3: Qualitative analysis

To evaluate the effectiveness of the captions generated by UrbanVLM, we performed a qualitative analysis by comparing the descriptions produced by BLIP, BLIP-2, and LlaVA. Example captions from each model are shown in Table IV. We assessed 50 randomly selected image-caption pairs to gauge the models' performance. Our analysis revealed that BLIP's captions were not accurate enough to be considered reliable—for instance, it described an image from the Philippines, even though the Philippines was not included in the dataset—leading us to exclude it from further evaluation.

In contrast, LlaVA consistently generated more detailed and contextually appropriate captions, capturing subtle aspects of the street view images. This higher quality of captioning likely contributed to better alignment with textual features, which in turn enhanced the model's performance in urban perception tasks. As a result, LlaVA outperformed the other models in all experiments.

Figure 2 presents the correlation matrices for six different perceptual scores, which serve as a measure of how various attributes are related to each other. Panel (a) displays the ground-truth correlations derived from human evaluations, providing a benchmark for assessing the model's performance. In contrast, panel (b) shows the correlations predicted by the UrbanVLM model.

The results reveal that the predicted correlations closely align with the human-annotated ones, indicating that UrbanVLM successfully captures the complex relationships between the perceptual attributes. This strong agreement suggests that the model can discern patterns in urban perception that are similar to those recognized by humans. Furthermore, the

TABLE IV
IMAGE DESCRIPTION GENERATION SAMPLE

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.

alignment of these correlation matrices highlights the model's capability to understand and replicate human-like judgments about urban environments.

V. LIMITATIONS

A. Data related

One limitation of this research lies in the inherent subjectivity of the Place Pulse 2.0 dataset, which is based on human perceptions of urban environments (e.g., safety, beauty, and liveliness). These perceptions can vary significantly across individuals due to cultural, social, and demographic differences. Consequently, the dataset's labels may not generalize well to diverse populations or regions outside the dataset's original scope.

This limitation is further compounded when using multimodal models, such as BLIP-2 and LlaVA, for image description generation, as these models may inherit or amplify biases from the data used during pretraining or fine-tuning.

Similarly, vision-language models such as CLIP and SigLIP, which are fine-tuned on these subjective labels, potentially reinforce and propagate existing biases present in the dataset. As a result, the model's outputs may reflect skewed or culturally specific perceptions of urban environments, which calls for cautious interpretation and the need for more diverse, representative training data in future research.

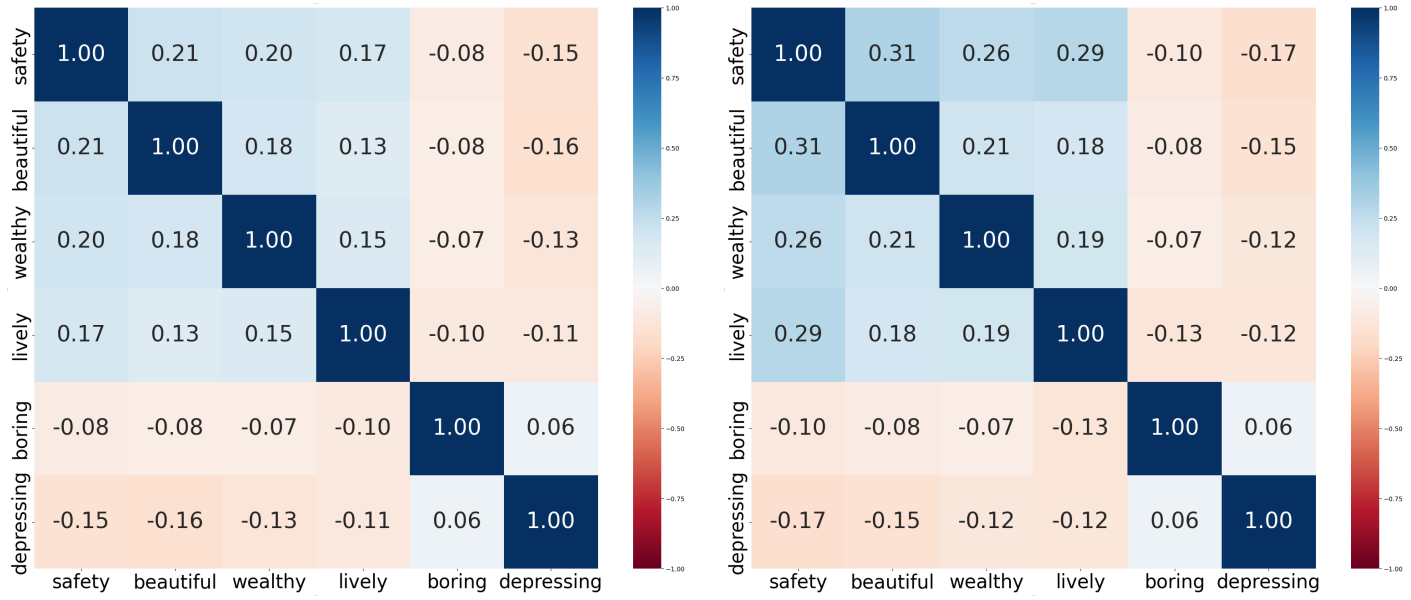
B. Model-related

Additionally, while leveraging high-performance models such as LlaVA, BLIP-2, and CLIP or SigLIP provides state-of-the-art capabilities for multimodal understanding and urban perception analysis, the reliance on a single Nvidia RTX 3090 GPU introduces significant limitations. This hardware constraint restricts the ability to experiment with larger or more complex model architectures, as well as to perform extensive hyperparameter tuning, ablation studies, or training over larger datasets.

As a result, the capacity to explore the full potential of these models is limited, and the outcomes may not reflect the best possible performance for classification and regression tasks. Moreover, the computational burden of processing a large dataset is substantial; generating image captions for more than 100,000 street view images takes approximately two weeks of continuous processing, making it challenging to iterate quickly or test multiple variants of the pipeline.

Furthermore, generating meaningful descriptions and identifying visual-semantic features that contribute to perceptual attributes (e.g., boredom, safety, or liveliness) is inherently influenced by the pretraining data and objectives of these foundation models. Since these models are trained on broad internet-scale datasets, they may capture and prioritize stereotypical or contextually shallow features that do not fully align with the specific urban settings or cultural nuances present in the dataset.

This can result in oversimplified or misleading representations of urban scenes, where subtle indicators of safety or beauty are either overlooked or misinterpreted. Such limitations underscore the need for increased computational resources to support more extensive fine-tuning, larger model experimentation, and faster processing pipelines—factors that are essential for enhancing the fidelity and relevance of the generated descriptions in the urban perception domain.



(a) "Strength of schedule" algorithm

(b) UrbanVLM predictions

Fig. 2. Correlation matrix of the six perceptual scores computed using human-based evaluations and UrbanVLM scores predictions.

VI. Conclusions

This paper presents UrbanVLM, a novel multimodal pipeline for urban perception that combines image-text generation, contrastive learning, dual-modality encoding, and task optimization (classification and regression) to predict urban perception labels and scores. By using contrastive loss to align image and text encodings, and jointly training image and description tokens, as well as visual projections, the model benefits from rich multimodal representations.

Integrating these models enables the automation of analyzing large volumes of Street View imagery with remarkable precision, yielding results closely aligned with human evaluations. This capability empowers researchers and policymakers to assess urban environments at an unprecedented scale, significantly reducing the time and effort required by traditional survey-based methods while preserving the depth and nuance of human judgment.

Acknowledgement

This work was supported by the National Council for Scientific and Technological Development (CNPq, grant #311144/2022-5), Carlos Chagas Filho Foundation for Research Support of Rio de Janeiro State (FAPERJ, grant #E-26/204.593/2024), São Paulo Research Foundation (FAPESP, grant #2021/07012-0) and the School of Applied Mathematics at Fundação Getulio Vargas.

References

[1] Acosta, S.F., Camargo, J.E.: City safety perception model based on visual content of street images. 2018 IEEE International Smart Cities Conference (ISC2) pp. 1–8 (2018)

[2] Arietta, S.M., Efros, A.A., Ramamoorthi, R., Agrawala, M.: City forensics: Using visual elements to predict non-visual city attributes. IEEE transactions on visualization and computer graphics (2014)

[3] Baltruaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 423–443 (2017)

[4] Beaucamp, B., Leduc, T., Tourre, V., Servieres, M.C.J.: The whole is other than the sum of its parts: Sensibility analysis of 360° urban image splitting. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2022)

[5] Costa, G.: City-safe: Estimating urban safety perception (2019)

[6] Dai, S., Li, Y., Stein, A., Yang, S., Jia, P.: Street view imagery-based built environment auditing tools: a systematic review. International Journal of Geographical Information Science pp. 1–22 (2024)

[7] De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., Lepri, B.: The death and life of great italian cities: a mobile phone data perspective. In: Proceedings of the 25th international conference on world wide web. pp. 413–423 (2016)

[8] Doersch, C., Singh, S., Gupta, A.K., Sivic, J., Efros, A.A.: What makes paris look like paris? ACM Transactions on Graphics (TOG) **31**, 1 – 9 (2012)

[9] Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city: Quantifying urban perception at a global scale. In: ECCV (2016)

[10] Glaeser, E.L., Kominers, S.D., Luca, M., Naik, N.: Big data and big cities: The promises and limitations of improved measures of urban life. Economic Inquiry **56**(1), 114–137 (2018)

[11] Huang, J., Qing, L., Han, L., Liao, J., Guo, L., Peng, Y.: A collaborative perception method of human-urban environment based on machine learning and its application to the case area. Eng. Appl. Artif. Intell. **119**, 105746 (2023)

[12] Huang, W., Wang, J., Cong, G.: Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. International Journal of Geographical Information Science **38**, 1414 – 1442 (2024)

[13] Keizer, K., Lindenberg, S., Steg, L.: The spreading of disorder. Science (New York, N.Y.) **322**, 1681–5 (12 2008)

[14] Kruse, J., Kang, Y., Liu, Y.N., Zhang, F., Gao, S.: Places for play: Understanding human perception of playability in cities using street view images and deep learning. Comput. Environ. Urban Syst. **90**, 101693 (2021)

[15] Lavi., B., Tokuda., E., Moreno-Vera., F., Nonato., L., Silva., C., Poco., J.: 17k-graffiti: Spatial and crime data assessments in são paulo city. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 968–975. SciTePress (2022)

[16] Levering, A., Marcos, D., Jacobs, N., Tuia, D.: Prompt-guided and multimodal landscape scenicness assessments with vision-language models. PLOS ONE **19** (2024)

[17] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

[18] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)

[19] Liu, X., Chen, Q., Zhu, L., Xu, Y., Lin, L.: Place-centric visual urban perception with deep multi-instance regression. Proceedings of the 25th ACM international conference on Multimedia (2017)

[20] Liu, Y., Chen, M., Wang, M., Huang, J., Thomas, F., Rahimi, K., Mamouei, M.: An interpretable machine learning framework for measuring urban perceptions from panoramic street view images. iScience **26** (2023)

[21] Lynch, K.: Reconsidering the image of the city pp. 151–161 (1984)

[22] Ma, H., Wu, D.: A natural language processing-based approach: mapping human perception by understanding deep semantic features in street view images (2023)

[23] Ma, Z.: Deep exploration of street view features for identifying urban vitality: A case study of qingdao city. Int. J. Appl. Earth Obs. Geoinformation **123**, 103476 (2023)

[24] Malekzadeh, M.S., Willberg, E.S., Torkko, J., Toivonen, T.: Urban visual appeal according to chatgpt: Contrasting ai and human insights (2024)

[25] Min, W., Mei, S., Liu, L., Wang, Y., Jiang, S.: Multi-task deep relative attribute learning for visual urban perception. IEEE Transactions on Image Processing **29**, 657–669 (2020)

[26] Moreno-Vera, F.: Understanding safety based on urban perception. In: International Conference on Intelligent Computing. pp. 54–64. Springer (2021)

[27] Moreno-Vera, F., Brandoli, B., Poco, J.: What makes a place feel safe? analyzing street view images to identify relevant visual elements. In: International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (2024)

[28] Moreno-Vera, F., Lavi, B., Poco, J.: Quantifying urban safety perception on street view images. In: International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (2021)

[29] Moreno-Vera, F., Lavi, B., Poco, J.: Urban perception: Can we understand why a street is safe? In: Mexican International Conference on Artificial Intelligence. pp. 277–288. Springer (2021)

[30] Moreno-Vera., F., Medina., E., Poco., J.: WSAM: Visual explanations from style augmentation as adversarial attacker and their influence in image classification. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 830–837. SciTePress (2023)

[31] Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: StreetScore: predicting the perceived safety of one million streetscapes. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops

[32] Nasar, J.L.: The evaluative image of the city (1998)

[33] Ordonez, V., Berg, T.L.: Learning high-level judgments of urban perception. European Conference on Computer Vision (ECCV) (2014)

[34] Park, J., Newman, M.: A network-based ranking system for us college football. Journal of Statistical Mechanics: Theory and Experiment **2005**, P10014 – P10014 (2005)

[35] Porzi, L., Rota Bulò, S., Lepri, B., Ricci, E.: Predicting and understanding urban perception with convolutional neural networks. In: ACM international conference on Multimedia (10 2015)

[36] Quercia, D., O'Hare, N., Cramer, H.: Aesthetic capital: what makes london look beautiful, quiet, and happy? Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (2014)

[37] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[38] Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: Mapping the inequality of urban perception. PLOS ONE (2013)

[39] Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T.: Assessing "neighborhood effects": Social processes and new directions in research. Annual review of sociology **28**(1), 443–478 (2002)

[40] Sangers, R., van Gemert, J.C., van Cranenburgh, S.: Explainability of deep learning models for urban space perception (2022)

[41] Santani, D., Ruiz-Correa, S., Gática-Pérez, D.: Looking south. ACM Transactions on Social Computing **1**, 1 – 23 (2018)

[42] Santos, F.A., Silva, T.H., Loureiro, A.A.F., Villas, L.A.: Uncovering the perception of urban outdoor areas expressed in social media. 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) pp. 120–127 (2018)

[43] Sengupta, N., Vaidya, A., Evans, J.: In her shoes: Gendered labelling in crowdsourced safety perceptions data from india. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (2023)

[44] Seresinhe, C.I., Preis, T., Moat, H.S.: Using deep learning to quantify the beauty of outdoor places. Royal Society Open Science **4** (2017)

[45] Skogan, W.G.: Disorder and decline: Crime and the spiral of decay in American neighborhoods. Univ of California Press (1992)

[46] Song, H.: Street view imagery: Ai-based analysis method and application. Applied and Computational Engineering (2024)

[47] Stalidis, P., Semertzidis, T., Daras., P.: Examining deep learning architectures for crime classification and prediction. Forecasting **3**(4), 741–762 (2021)

[48] Tokuda, E.K., Silva, C.T., Jr., R.M.C.: Quantifying the presence of graffiti in urban environments pp. 1–4 (2019)

[49] Wendt, M.: The importance of death and life of great american cities (1961) by jane jacobs to the profession of urban planning. New Visions for Public Affairs **1**, 1–24 (2009)

[50] Wilson, J.Q., Kelling, G.L.: Broken windows. Atlantic monthly **249**(3), 29–38 (1982)

[51] Xu, X., Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., Luo, D.: Associations between street-view perceptions and housing prices: Subjective vs. objective measures using computer vision and machine learning techniques. Remote. Sens. **14**, 891 (2022)

[52] Yao, Y., Liang, Z., Yuan, Z., Liu, P., Bie, Y., Zhang, J., Wang, R., Wang, J., Guan, Q.: A human-machine adversarial scoring framework for urban perception assessment using street-view images. International Journal of Geographical Information Science **33**, 2363 – 2384 (2019)

[53] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)

[54] Zhang, C., Wu, T., Zhang, Y., Zhao, B., Wang, T., Cui, C., Yin, Y.: Deep semantic-aware network for zero-shot visual urban perception. International Journal of Machine Learning and Cybernetics **13**, 1197 – 1211 (2021)

[55] Zhang, F., Hu, M., Che, W., Lin, H., Fang, C.: Framework for virtual cognitive experiment in virtual geographic environments. ISPRS Int. J. Geo Inf. **7**, 36 (2018)

[56] Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., Ratti, C.: Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. Annals of the American Association of Geographers pp. 1–22 (2024)

[57] Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C.: Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning **180**, 148–160 (2018)

[58] Zhang, Y., Xiong, X., Yang, S., Zhang, Q., Chi, M., Wen, X., Zhang, X., Wang, J.: Enhancing the visual environment of urban coastal roads through deep learning analysis of street-view images: A perspective of aesthetic and distinctiveness. PLOS ONE **20** (2025), https://api.semanticscholar.org/CorpusID:275541765

[59] Zhanga, J., Lia, Y., Fukudab, T., Wang, B.: Revolutionizing urban safety perception assessments: Integrating multimodal large language models with street view images (2024)

[60] Zhao, X., Lu, Y., Lin, G.: An integrated deep learning approach for assessing the visual qualities of built environments utilizing street view images. Engineering Applications of Artificial Intelligence **130**, 107805 (2024)

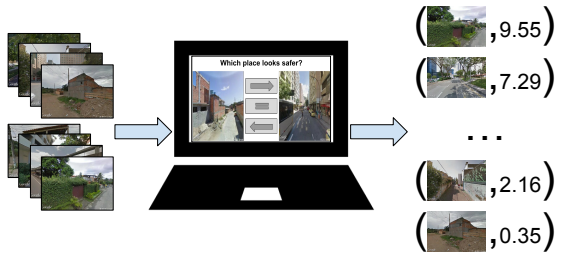
APPENDIX

A. Background: Large Multimodal Models (LMMs)

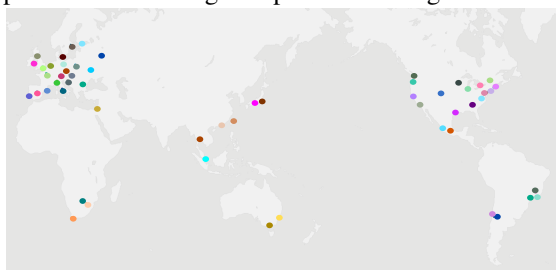
Multimodal models are machine learning models that combine information from two or more distinct types of data, referred to as modalities. Typical modalities include text, images, audio, and video. The objective is to learn representations that capture complementary signals from each modality, leading to improved performance in complex tasks compared to single-modality models [3]. Formally, given inputs $x_1, x_2, \ldots, x_n$ from different modalities, a multimodal model aims to learn a function $z = f(x_1, x_2, \ldots, x_n)$ that integrates these inputs into a unified representation.

Applications of multimodal models include tasks such as image captioning, visual question answering, and audio-visual speech recognition. The combination of modalities enables models to reason across different types of information, resulting in more robust and context-aware predictions. Recent advancements in vision-language models, such as BLIP-2 [17] and LLaVA [18], leverage large pre-trained vision encoders and language models to generate text conditioned on image features. These models align visual and textual encodings in a shared space, enabling tasks like image captioning, visual question answering, and perception-based description generation. Contrastive learning objectives, such as those used in CLIP [37], further enhance multimodal alignment by encouraging representations of matching image-text pairs to be close in the encoding space while pushing apart mismatched pairs.

B. Percpetual score distribution



(a) Pipeline for obtaining comparisons through online surveys



(b) Cities included in this dataset

Fig. 1. Place Pulse 2.0 dataset: Pipeline to obtain data and cities included in this evaluation.

Figure 1 illustrates the processing pipeline used to obtain image comparisons between images i and j in category k (e.g., safety). Additionally, we present the cities included in the survey, noting that the United States is represented by a greater number of cities compared to other countries. We calculate the safety perception scores twice. Figure 2 (a) shows the distribution using all unique IDs in the dataset, where we observe that most images have a score of 3.33. This likely results from the number of comparisons and their corresponding wins and losses. Figure 2 (b) presents the distribution of perceptual scores after aggregating images by their ID. Specifically, we group images corresponding to the same location and ID. After this adjustment, the distribution becomes smoother and more balanced.

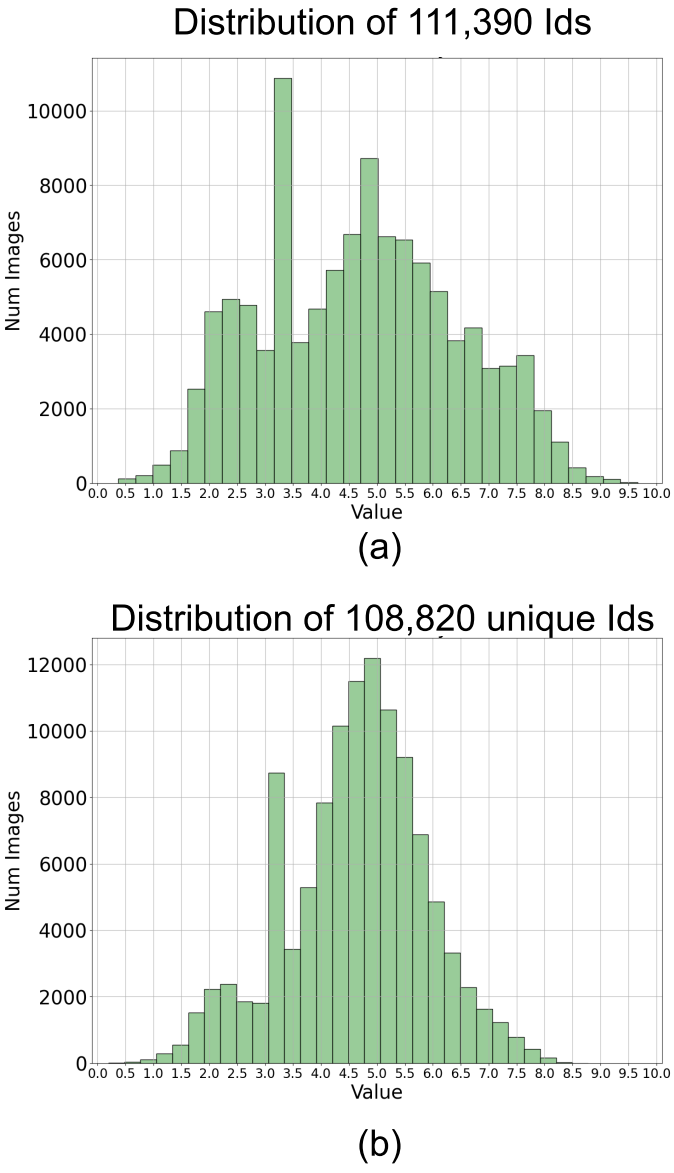


Fig. 2. Safety score distribution in both scenarios: (a) using all 111,390 image IDs; (b) mapping all repetitions to unique 108,820 IDs.

C. Prompts to generate image descriptions

We define two main prompts: (i) *"You are an ordinary observer analyzing a street view image. Please describe this*

image focusing in the visual appearance." for LlaVA model and *"Describe this image and its visual appearance."* for BLIP-2. This prompt is focused on providing a general description of the image. (ii) The second prompt focuses on describing the feelings or perceptions evoked by the image. Based on prior work [24], [59], we incorporate parameters such as city, country, and the category being assessed.

For the second type, we use two different prompt configurations depending on the model being used:

LlaVA

Prompts structure for LlaVA focused on the category:

```
Imagine you are an observer
analyzing a street view image.
But you know about some
demographic factors and crime rates
in the city {city}, {country}.

Based on the street view image provided,
please describe the factors that contribute
to making this street view image
feel {category}.

Consider elements such as the
visual appearance,
environment, colors, structures,
infrastructure, well-maintained level,
daylight, and any human or
social factors.
```

BLIP-2

Due to the token limitation in both models, we use a reduced prompt:

```
Question: What make this street view image
from {city}, {country}, feel {category}?
Consider aspects like the environment,
well-maintained, daylight, and architecture.
Answer:
```

D. Prompts for zero-Shot evaluations

When studying the ablation case without image-description generation, we provide definitions to help the models assign scores and determine the appropriate category (e.g., defining what constitutes a safe street).

```
Safety: "A well-lit, calm area with visible
security features like police or cameras,
and no signs of danger."

Not safety: "A poorly lit, isolated area with
signs of neglect or danger, like vandalism
or suspicious individuals."

Lively: "A vibrant, bustling area with
lots of activity, pedestrians, and vehicles
creating an energetic atmosphere."

Not lively: "A quiet, empty area with little
activity, feeling dull and uninviting."

Boring: "A dull, inactive area with no
significant activity, feeling monotonous
and quiet."

Not boring: "A fast-paced, vibrant area
with energy, movement, and entertainment."

Wealthy: "An affluent area with
luxury shops, well-maintained
infrastructure, and grand buildings."

Not wealthy: "A neglected, impoverished
area with rundown buildings,
poor infrastructure, and visible poverty."

Depressing: "A neglected area with
rundown buildings, broken windows,
and a gloomy, isolated feel."

Not depressing: "A well-maintained,
lively area with clean streets,
greenery, and good lighting."

Beautiful: "A visually pleasing area
with lush greenery,
attractive architecture,
and scenic elements."

Not beautiful: "An unattractive area
with faded buildings, litter,
and a sense of decay."
```