APPENDIX

A. Background: Large Multimodal Models (LMMs)

Large Language Models (LLMs) are highly effective at generating natural language, while Large Multimodal Models (LMMs) enhance this ability by integrating visual understanding. These models combine a pre-trained LLM with a visual encoder (e.g., CLIP, SigLIP) to extract visual features f and use an adapter W to translate these features into the language space. Following the training paradigm of a general LMM, this relationship can be expressed as:

$$C = \{x_1, x_2, \ldots, x_l\} \tag{1}$$

$$x_t = LMM(f_T(x_{t-1}) + W(f_I(I))) \tag{2}$$

Where C represents the tokens, l represents the number of tokens, f_T is the feature representation from the text modality, f_I is the feature representation from the image modality.

B. Perceptual scores

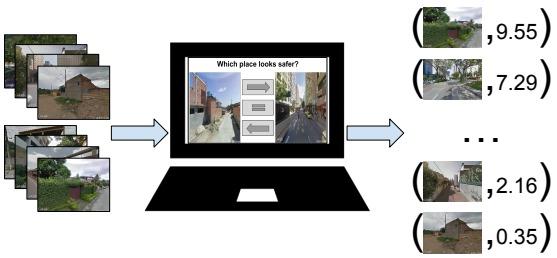


Fig. 2. The data pipeline, starting with street view images evaluated by humans through an online survey, then calculated the perceptual score for each image.

Figure 2 illustrates the processing pipeline to obtain image comparisons. For each comparison between images i and j in the category k (e.g., safe), we define *intensity of perception* of the image i as the percentage of times that the image was selected and is affected by the intensity of the compared images j.

$$W_i = \frac{w_i}{w_i + d_i + l_i} \tag{3}$$

$$L_i = \frac{l_i}{w_i + d_i + l_i} \tag{4}$$

$$Q_{i,k} = \frac{10}{3}(W_{i,k} + \frac{1}{n_i}(\sum_{j}^{n_i} W_{j,k}) - \frac{1}{m_i}(\sum_{j}^{m_i} L_{j,k}) + 1) \tag{5}$$

The Equation 5 represents the perceptual score of image i, referred to as the *Q-score*, and denoted $Q_{i,k}$, within category k. Here, $W_{i,k}$ (Equation 3) and $L_{i,k}$ (Equation 4) represent the win and loss rates of image i in category k. In addition, n_i is the number of images j that image i has won against, and m_i is the number of images j that image i has lost to. Finally, following previous studies on visual assessment [30], [36], the perceptual score Q 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. This scaling is achieved by adding a constant value of 1 and multiplying by $\frac{10}{3}$.

C. Percpetual score distribution

We calculate the safety perception scores twice using the algorithm described in Appendix B. Figure 3 (a) shows the distribution when using all unique IDs stored in the dataset. Here, we observe that most sample images have a score of 3.33, which may be due to the number of comparisons between images and the resulting wins and losses. Figure 3 (b) presents the distribution of the perceptual scores after identifying and aggregating images by their ID. Specifically, we identify different IDs that correspond to the same image and location and then group them. After this adjustment, we observe that the distribution is smoother and appears more balanced.

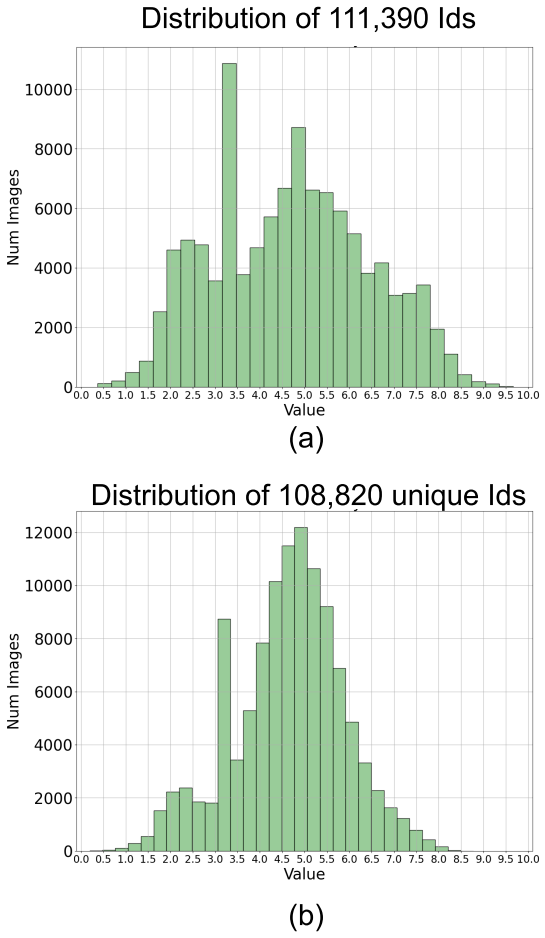


Fig. 3. Safety score distribution in both scenarios: (a) using all 111,390 image IDs; (b) mapping all repetitions to unique 108,820 IDs.

D. 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 [23], [56], 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:
```

E. 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."
```

F. Model comparisons in classification and regression tasks

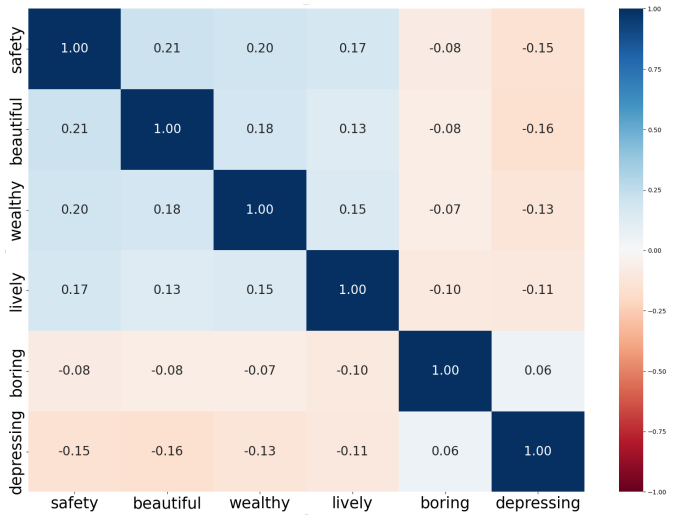
TABLE II
ACCURACY REPORT USING BINARY CLASSIFICATION IN SAFE CATEGORY

Model	Acc
DSAPN+ResNet [52]	64.87
MTDRALN-LC [24]	65.07
MTDRALN-TC [24]	65.82
VGG+ImageNet [27]	65.72
VGG-GAP+ImageNet [27]	66.09
VGG+Places365 [27]	66.46
VGG-GAP+Places365 [27]	66.96
VGG19+ImageNet [3]	67.01
PSPNet+SVR [53]	70.63
DeiT+ResNet50 [38]	71.16
ViT-nn [26]	71.29
ViT-nn+OneFormer [26]	75.68
UrbanVLM (LlaVA+CLIP)	82.45

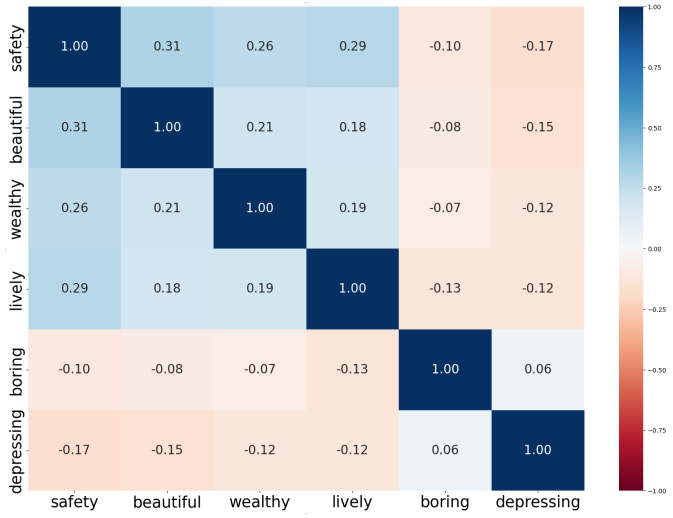
G. Correlation matrix of perceptual scores

TABLE III
REGRESSION RESULTS IN SAFE CATEGORY

Model	R^2	RMSE
PSPNet-Regressor [53]	0.25	–
Fine-Tuned BERT [21]	0.42	–
FPN-based regressor [19]	0.52	–
DeepLabV3+ regressor [19]	–	2.16
DeepLabV3+ regressor [50]	–	2.91
SFB5+ConvNeXt-B+RF [57]	0.67	1.29
VIT+SegFormer+RF [10]	0.76	1.75
UrbanVLM (LlaVA+CLIP)	0.88	1.04



(a)



(b)

Fig. 4. (a) Correlation matrix of the six perceptual scores computed using the ground truth annotations and the "strength of schedule" algorithm, and (b) correlation matrix of the perceptual scores predicted by our UrbanVLM.