UrbanAlign: Post-hoc Semantic Calibration for VLM-Human Preference Alignment

Yecheng Zhang[1,⋆,†], Rong Zhao[2,⋆], Zhizhou Sha[1], Yong Li[3], Lei Wang[4,†], Ce Hou[3], Wen Ji[5], Hao Huang[1], Yunshan Wan[6], Jian Yu[7], Junhao Xia[1], Yuru Zhang[8], Chunlei Shi[9,†]

[1]Tsinghua University, China [2]University College London, UK [3]HKUST, China [4]Peking University, China [5]Southwest Jiaotong University, China [6]Zhejiang University, China [7]University of Texas at Austin, USA [8]Renmin University of China, China [9]Southeast University, China

⋆Equal contribution. †Corresponding authors. wanglei2025@pku.edu.cn, zhangyec23@mails.tsinghua.edu.cn, 230238514@seu.edu.cn

Abstract. Vision-language models (VLMs) can describe urban scenes in rich detail, yet consistently fail to produce reliable human preference labels in domain-specific tasks such as safety assessment and aesthetic evaluation. The standard fix, fine-tuning or RLHF, requires large-scale annotations and model retraining. We ask a different question: *can a frozen VLM be aligned with human preferences without modifying any weights?* Our key insight is that VLMs are strong concept extractors but poor decision calibrators. We propose a three-stage post-hoc pipeline that exploits this asymmetry: (i) interpretable evaluation dimensions are automatically mined from consensus exemplars; (ii) an Observer–Debater–Judge chain extracts robust concept scores from the frozen VLM; and (iii) locally-weighted ridge regression on a hybrid manifold calibrates these scores to human ratings. Applied as **UrbanAlign** on Place Pulse 2.0, the framework reaches 70.8% accuracy (κ=0.41) across six perception categories, outperforming all baselines by +9.6 pp and zero-shot VLM by +14.1 pp, with full interpretability and zero weight modification.

Keywords: Human Preference Alignment · Post-hoc Concept Bottleneck · Vision-Language Model · Urban Perception

1 Introduction

Large vision-language models (VLMs) excel at describing visual scenes but consistently fail to produce reliable preference labels for domain-specific tasks. This *VLM–human preference alignment gap* appears across domains, from urban perception [25, 48] to aesthetic quality [24, 41], and the standard remedy is model adaptation: fine-tuning, LoRA, or RLHF [29], all of which modify model weights

and require domain-specific training data and non-trivial compute. We ask a different question: *can a frozen VLM be aligned with human preferences in a new domain, without modifying any model weights?*

The key observation is that VLMs are already strong *concept extractors* but poor *decision calibrators* [25, 48]: they can identify rich visual elements in a scene (e.g., building facades, vegetation, and street infrastructure), yet their mapping from these features to discrete comparison labels is poorly aligned with human judgement boundaries (Fig. 1). This mirrors the finding of Concept Bottleneck Models (CBMs) [17] that routing predictions through interpretable concept scores and then a calibrated linear layer dramatically outperforms end-to-end classification. Such a bottleneck can be retrofitted *post-hoc* onto any frozen backbone [45], and recent work further shows that concepts can be discovered directly from VLMs without manual labels [27, 43].

A naïve alternative is to prompt the VLM as a single-shot scorer, but this suffers from well-documented biases: position effects, anchoring, and inconsistent self-evaluation [22,41]. When asked to compare two scenes directly, a single VLM call conflates perception and judgement: the model must simultaneously identify relevant visual cues, weigh their relative importance, and produce a discrete label, all within a single forward pass. This end-to-end compression discards the structured reasoning that human raters implicitly perform when evaluating complex, multi-dimensional scenes. Multi-agent debate protocols decompose complex judgements into complementary reasoning steps and have been shown to mitigate such biases through structured disagreement [7, 39], yet these techniques have primarily targeted factual reasoning and text generation; their potential for domain-specific visual preference alignment remains unexplored.

This principle is operationalised as a pipeline that keeps the VLM entirely frozen, with three tightly coupled stages (Fig. 2): **(i)** *concept mining*, where interpretable evaluation dimensions are discovered by the VLM from consensus exemplars; **(ii)** *structured scoring*, where an Observer–Debater–Judge multi-agent chain extracts robust continuous concept scores from the frozen VLM; **(iii)** *geometric calibration*, where locally-weighted ridge regression [6] on a hybrid visual–semantic manifold aligns concept scores to human ratings. An end-to-end optimization loop unifies these stages, automatically selecting the dimension set that maximises calibrated accuracy via temperature-scheduled trials. We instantiate this framework as **UrbanAlign** on Place Pulse 2.0 [33], a large-scale pairwise comparison dataset covering six categories of urban perception, where subjective judgements [12, 40] defy the feature-to-label mappings of existing approaches [8, 26]. UrbanAlign substantially outperforms both supervised baselines and zero-shot VLM scoring across all six categories, with full dimension-level interpretability and zero model-weight modification, demonstrating that a frozen VLM already contains sufficient perceptual knowledge for human-aligned urban assessment.

Contributions.

1. **End-to-end concept mining.** We design a loop that automatically discovers interpretable evaluation dimensions from consensus exemplars and refines

them via a temperature-scheduled search, forming a concept bottleneck for perception prediction.
2. **Multi-agent structured scoring.** We introduce an Observer–Debater–Judge deliberation chain that extracts robust continuous concept scores from a frozen VLM, reducing single-agent bias through structured disagreement.
3. **Local geometric calibration.** We propose locally-weighted ridge regression on a hybrid visual–semantic manifold that calibrates concept scores against human ratings with per-sample interpretability.

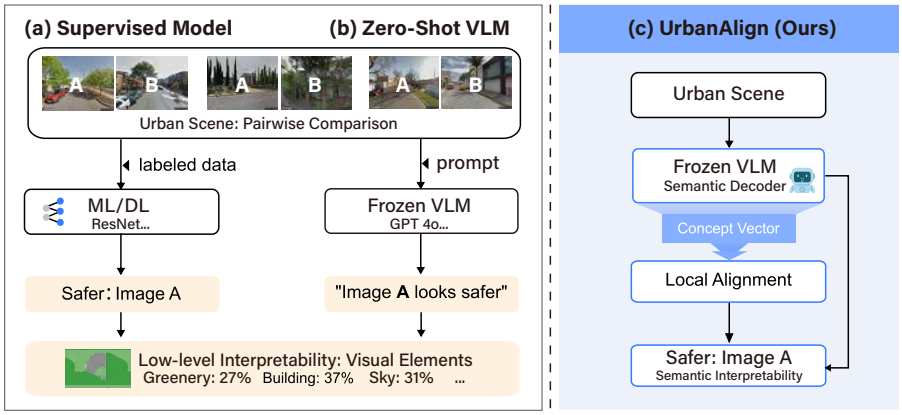

Fig. 1: Three approaches to urban perception: **(a)** supervised baselines (e.g., Siamese networks, segmentation regression) require labelled data and lack interpretable intermediates; **(b)** zero-shot VLM prompting is training-free but produces opaque, uncalibrated judgements; **(c)** UrbanAlign routes predictions through VLM-discovered semantic dimensions for interpretable, locally calibrated comparison.

2 Related Work

Urban perception and preference data. Pairwise comparison has a long history in psychometrics [3, 37]. Place Pulse collects over one million pairwise comparisons of street-view images across six perceptual dimensions [33]. Siamese networks [8], multi-task learning [47], and continuous score regression [26] have been trained on these labels, but map directly from pixels to abstract percepts without interpretable mid-level representations. Cross-cultural evaluations further show that such models transfer poorly to non-Western cities [14], motivating cost-effective local calibration rather than large-scale re-annotation. Pairwise preference datasets now extend to 10-dimension urban perception [30] and photographic aesthetics [24], paralleling the learning-to-rank paradigm [15], yet the resulting models remain black-box rankers. Foundation models have been used

to generate synthetic training corpora [23, 44], but these operate on unimodal text with discrete labels and do not address the multi-dimensional, subjective nature of visual preference. UrbanAlign instead synthesises *visual* pairwise data via structured dimension scoring and post-hoc geometric calibration, requiring no weight modification.

VLM-based scoring and multi-agent reasoning. Systematic evaluation reveals that VLMs capture broad perceptual trends but oversimplify multi-dimensional human judgements [48]; benchmarking on urban perception confirms stronger alignment on objective properties (*e.g.*, building density) than subjective ones (*e.g.*, safety, boringness) [25]. Contrastive pretraining on satellite imagery [42] and zero-shot urban function inference [11] further demonstrate both the power and limits of VLMs for urban tasks; these studies position VLMs as end-to-end classifiers, whereas we reposition them as *structured semantic decoders* whose outputs require post-hoc calibration. Pairwise evaluation aligns better with human opinions than pointwise scoring [22, 41], yet single-shot scoring still suffers from position effects and anchoring biases. Self-consistency sampling [39] and multi-agent debate [7, 19, 21] mitigate these issues for factual QA and mathematical reasoning, but applying structured deliberation to continuous visual scoring with domain-specific semantics remains unexplored. UrbanAlign bridges this gap with an Observer–Debater–Judge chain that separates observation, argumentation, and judgement into distinct reasoning stages for pairwise dimension scoring.

VLM preference alignment. Recent VLM alignment methods can be grouped by whether they modify model weights. RLHF-based approaches train reward models and fine-tune the VLM via policy optimisation [29]; extensions augment this pipeline with factual grounding [35] and vision-guided reinforcement that eliminates human annotation entirely [46]. Robust visual reward models leverage auxiliary textual preference data to improve reward quality under scarce visual annotations [38], while implicit preference mining from user-generated content broadens the training signal beyond curated labels [36]. All these methods require modifying VLM weights. A parallel line calibrates VLM confidence post-hoc via semantic perturbation at the token level [49], but targets per-query confidence rather than domain-specific preference alignment. UrbanAlign occupies a distinct position: it requires *zero weight modification*, instead routing predictions through an external concept bottleneck with local geometric calibration—combining the interpretability of CBMs with the locality of retrieval-augmented prediction.

Concept bottleneck models and post-hoc calibration. Modern neural networks are known to be poorly calibrated [9]; temperature scaling and Platt scaling partially address this, but domain-specific preference alignment requires richer structural remedies. Concept Bottleneck Models (CBMs) route predictions through interpretable concept scores, enabling human-auditable reasoning and concept-level intervention [17]. Post-hoc retrofitting onto frozen backbones [45] and automated concept discovery from VLMs [27, 43] eliminate manual concept engineering, making CBMs applicable without task-specific annotation of concept labels. Our semantic dimensions serve as an analogous interpretable

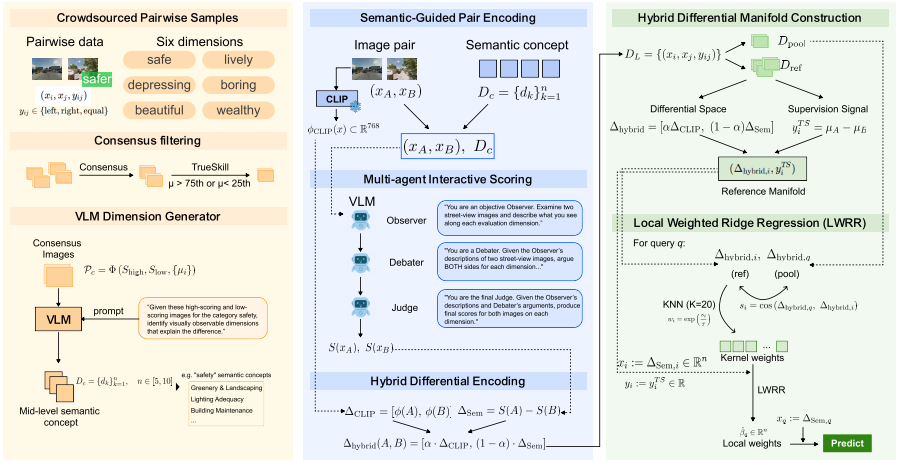

Fig. 2: UrbanAlign framework. Stage 1 mines semantic dimensions from consensus exemplars; Stage 2 distils dimension scores via Observer–Debater–Judge; Stage 3 calibrates via LWRR on the hybrid differential manifold. $\mathcal{D}_{\text{ref}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing$.

bottleneck, but operate *externally* on VLM-synthesised data rather than as an internal network layer [20, 34]: the VLM generates continuous concept scores, and a separate calibration step maps them to human-aligned predictions.

The calibration mechanism draws on locally-weighted regression [1,6,32] and can be viewed as a visual analogue of retrieval-augmented generation (RAG) [18]: for each query pair, the system retrieves the K most similar reference pairs from the hybrid manifold and locally fits a calibration function, adapting dimension weights to the manifold neighbourhood. Unlike global linear probes or fixed concept weights, this local fitting allows the relative importance of dimensions to vary smoothly across the perceptual manifold, capturing the context-dependent nature of human preference.

3 Method

3.1 Problem Formulation

Given an urban image set $\mathcal{X}=\{x_1, \ldots, x_N\}$ and a perception category d (*e.g.*, wealthy), we aim to predict human pairwise preferences:

$$y_{ij} = \begin{cases} \text{left} & \text{if } x_i \succ x_j \text{ on dimension } d, \\ \text{right} & \text{if } x_i \prec x_j, \\ \text{equal} & \text{if } x_i \approx x_j. \end{cases} \tag{1}$$

The labelled pool $\mathcal{D}_L=\{(x_i, x_j, y_{ij})\}$ contains crowdsourced pairwise comparisons (from Place Pulse 2.0). We split $\mathcal{D}_L$ into a *reference set* $\mathcal{D}_{\text{ref}}$ (for LWRR cal-

ibration), a *validation set* $\mathcal{D}_{\text{val}}$ (for dimension and hyperparameter search), and a *test set* $\mathcal{D}_{\text{test}}$ (for final evaluation), with strict isolation: $\mathcal{D}_{\text{ref}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing$.
Notation. $\phi: \mathcal{X} \to \mathbb{R}^{768}$ denotes a pre-trained CLIP encoder (ViT-L/14) [31]. TrueSkill [10] converts pairwise comparisons to continuous ratings μ_i per image.

3.2 Framework Overview

The pipeline consists of three stages unified by an end-to-end optimization loop (Fig. 2). *Concept mining* (Sec. 3.3) discovers interpretable evaluation dimensions from a handful of consensus exemplars and optimises them via a temperature-scheduled search. *Multi-agent structured scoring* (Sec. 3.4) extracts continuous dimension scores from a frozen VLM through Observer–Debater–Judge deliberation, producing hybrid visual–semantic feature vectors. *Local manifold calibration* (Sec. 3.5) aligns these scores to human ratings via locally-weighted ridge regression on the hybrid differential manifold. Data isolation ($\mathcal{D}_{\text{ref}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing$) ensures that calibration, validation, and evaluation data never overlap.

3.3 Concept Mining and Dimension Optimization (Fig. 2a)

Motivation. Instead of asking a VLM "Which image looks more wealthy?" (a poorly calibrated end-to-end query), we first decompose the abstract percept into interpretable, continuously scorable sub-dimensions. This is analogous to defining the concept set in a CBM [17]: the dimensions form an interpretable bottleneck through which all subsequent predictions must pass.

TrueSkill consensus sampling. We convert pairwise comparisons to continuous scores via TrueSkill [10]. We sample *high-consensus* ($\mu > \tau_h,\ \sigma < \sigma_{\text{med}}$) and *low-consensus* ($\mu < \tau_l,\ \sigma < \sigma_{\text{med}}$) exemplars, where τ_h and τ_l are the 75th and 25th percentiles. A small set per group forms the consensus set $\mathcal{S}_{\text{consensus}}$.

Dimension extraction. A structured prompt presents the high/low exemplars with their TrueSkill scores and PCA-compressed CLIP embeddings. The VLM outputs $n \in [5, 10]$ dimensions in JSON, each containing a name, description, and high/low indicators (exact schema and prompts in supplementary Section 4). For the *wealthy* category, the extracted dimensions are: *Façade Quality*, *Vegetation Maintenance*, *Pavement Integrity*, *Vehicle Quality*, *Building Modernity*, *Infrastructure Condition*, *Street Cleanliness*, and *Lighting Quality*.

These dimensions satisfy three requirements: *visual observability* (scorable from street-view images), *measurability* (1–10 continuous scale), and *universality* (applicable across cities).

End-to-end dimension optimization. The dimensions discovered above depend on the VLM's sampling temperature and the particular consensus exemplars presented. We close the loop with an automated search over dimension sets, optimising each perception category independently. Let $\mathcal{D}_c^{(t)}$ denote the dimension set generated at trial t for category c. Each trial runs the full pipeline

(dimension extraction → multi-agent scoring → LWRR calibration) on the held-out validation split $\mathcal{D}_{\text{val}}$. The per-category optimal set is:

$$\mathcal{D}_c^* = \arg \max_{t \in \{1,\ldots,T\}} \text{Acc}_c\big(\text{LWRR}\big(\text{Score}(\mathcal{D}_{\text{val}}^c,\, \mathcal{D}_c^{(t)}),\, \mathcal{D}_{\text{ref}}^c\big)\big). \tag{2}$$

The search proceeds in two phases: *Explore* (high τ_{gen}): diverse, independently generated dimension sets; each category accumulates its best. *Converge* (low τ_{gen}): per-category best dimensions are preserved except for 1–2 targeted replacements (*mutation mode*), enabling local refinement. Since categories may peak at different trials, the final output *assembles* per-category optima: $\{\mathcal{D}_c^*\}_{c=1}^{C}$.

3.4 Multi-Agent Structured Scoring (Fig. 2b)

From classification to feature extraction. Rather than querying a VLM for a one-shot discrete label, we extract a *hybrid feature vector*:

$$h(x) = [\phi_{\text{CLIP}}(x),\ S(x)], \tag{3}$$

where $S(x) \in [1, 10]^n$ is the vector of semantic dimension scores produced by the VLM, fusing low-level visual patterns with high-level domain semantics analogously to a post-hoc CBM [45].

Observer–Debater–Judge deliberation. Inspired by multi-agent debate [7, 19] and self-consistency [39], we decompose reasoning into three sequentially chained agents:

- **Observer**: describes visually observable details per dimension, without judgement, thereby suppressing confirmation bias.
- **Debater**: argues both for HIGH and LOW scores on each dimension, exploring opposing perspectives.
- **Judge**: synthesises descriptions and arguments to produce final scores $S(x) \in [1, 10]^n$.

Variance reduction. The three-agent chain reduces dimension score variance from σ_S^2 to at most $\sigma_S^2/(1+2(1-\rho))$, where $\rho \in [0, 1]$ is inter-agent correlation; when agents reason independently ($\rho \to 0$), variance drops by up to $3\times$ (proof in supplementary Section 5).

Intensity-based winner determination. For each pair (x_A, x_B), we also record the *intensity*: $I = |S(x_A) - S(x_B)|$ summed across dimensions. Pairs with intensity below a threshold σ_I are labelled "equal"; otherwise the VLM's predicted winner is retained.

3.5 Local Manifold Calibration (Fig. 2c)

This stage is the core algorithmic contribution: a post-hoc calibration layer analogous to the linear predictor in a CBM [17], but *locally adaptive* to account for heterogeneous perception patterns across the visual-semantic manifold.

Hybrid Differential Space. CLIP, pre-trained on web-scale image-text pairs, provides strong general features but exhibits distribution shift on street-view scenes. We augment it with semantic dimension scores to form the hybrid differential encoding:

$$\Delta_{\text{hybrid}}(A,B) = \big[\alpha \cdot \overline{\Delta}_{\text{CLIP}}(A,B),\ (1-\alpha) \cdot \overline{\Delta}_{\text{Sem}}(A,B)\big], \tag{4}$$

where $\overline{\Delta}_{\text{CLIP}}$ uses L2-normalised CLIP embeddings, $\Delta_{\text{Sem}}{=}S(A){-}S(B)$ is normalised to $[0,1]$ by dividing by 10, and $\alpha \in [0,1]$ balances visual and semantic components.

LWRR Algorithm. Our calibration extends locally-weighted regression [1, 6, 32] to the hybrid differential space defined above.
Step 1: Build reference manifold. From $\mathcal{D}_{\text{ref}}$, compute hybrid differential vectors and TrueSkill score differences $y_i^{\text{TS}}{=}\mu_A{-}\mu_B$. Mirror augmentation doubles the reference set: for each (A,B,y), add $(B,A,-y)$.
Step 2: Neighbourhood search. For query $\Delta_q^{\text{hybrid}} \in \mathcal{D}_{\text{test}}$, compute cosine similarities to all reference points and select the K nearest neighbours $\mathcal{N}_K$.
Step 3: Kernel weighting. $w_i = \exp(s_i/\tau)$ for $i \in \mathcal{N}_K$, where s_i is cosine similarity and τ is kernel bandwidth.
Step 4: Local ridge regression. Solve for local dimension weights:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i \in \mathcal{N}_K} w_i \big(y_i^{\text{TS}} - \mathbf{w}^\top \Delta_{\text{Sem},i}\big)^2 + \lambda \|\mathbf{w}\|^2, \tag{5}$$

with closed-form solution $\hat{\mathbf{w}} = (X^\top W X{+}\lambda I)^{-1} X^\top W y$, where $W{=}\text{diag}(w_1,\ldots,w_K)$ and X stacks the semantic differentials.
Step 5: Prediction. The calibrated score difference $\hat{\delta}_q = \hat{\mathbf{w}}^\top \Delta_{\text{Sem}}^q$; a local goodness-of-fit statistic provides per-prediction interpretability by measuring how much TrueSkill variance is explainable by the mid-level dimensions in each neighbourhood.
Step 6: Statistical re-inference.

$$\hat{y} = \begin{cases} \text{equal} & \text{if } |\hat{\delta}| < \varepsilon \text{ or } c_{\text{eq}} > \theta, \\ \text{left} & \text{if } \hat{\delta} > 0, \\ \text{right} & \text{otherwise,} \end{cases} \tag{6}$$

where ε is a score-difference threshold and $c_{\text{eq}}{=}|\{i \in \mathcal{N}_K : l_i{=}\text{equal}\}|/K$ measures local equal-consensus.

Theorem 1 (Local Calibration Bound). *(Adapted from classical locally-weighted regression [4, 6] to the hybrid VLM-concept manifold.) Under the local model $y_i^{\text{TS}} = w^*(q)^\top \Delta_{\text{Sem},i} + \epsilon_i$ within $\mathcal{N}_K(q)$, with zero-mean noise of variance σ^2, the LWRR estimator satisfies:*

$$\mathbb{E}\big[\|\hat{w} - w^*(q)\|^2\big] \ \le\ \frac{\sigma^2\, \text{tr}(H_\lambda^{-1} X^\top W^2 X\, H_\lambda^{-1})}{K} \ +\ \lambda^2 \|w^*\|^2, \tag{7}$$

where $H_\lambda = X^\top W X + \lambda I$*. At optimal* $\lambda^* = \sigma\sqrt{n/K}/\|w^*\|$*, this simplifies to* $O(\sigma\|w^*\|\sqrt{n/K})$*. The key implication for our setting is that with* n=5–8 *concept dimensions and* K=20 *neighbours, the system is well-overdetermined* ($K \gg n$)*, ensuring small estimation variance; in contrast, a global model incurs irreducible bias from manifold heterogeneity (supplementary Theorem 2).* Proof in supplementary Section 5.

The following decomposition provides intuitive guidance for why each pipeline stage is necessary, though it is approximate rather than a formal bound:

$$\text{Acc} \;\approx\; \Phi\left(\frac{\|w^*\|\,\bar{\delta}}{\sqrt{\sigma^2_{\text{LWRR}} + \sigma^2_S/\eta_M}}\right), \tag{8}$$

where $\bar{\delta}$ is the mean signal strength (maximised by concept mining), σ^2_S is dimension-score variance (reduced by multi-agent scoring with η_M-fold gain), and σ^2_{LWRR} is calibration error (minimised by LWRR). Each pipeline stage tightens a specific term: Stage 1 maximises $\bar{\delta}$, Stage 2 minimises σ^2_S/η_M, and Stage 3 minimises σ^2_{LWRR}; a formal end-to-end bound is given in supplementary Corollary 2. Full pseudocode in supplementary Section 8; additional theory (local vs. global calibration, E2E convergence) in Section 5.

4 Experiments and Results

4.1 Setup

Dataset. Place Pulse 2.0 [33]: 110,688 street-view images, 1.17M pairwise comparisons, six categories. We use all pairs with $\geq$2 independent votes, sampling 449/category (capped by *boring*); $N \geq 3$ results (72.2%) in Supp. J.

Data protocol. Each category is partitioned into $\mathcal{D}_{\text{ref}}$ (60%), $\mathcal{D}_{\text{val}}$ (20%), and $\mathcal{D}_{\text{test}}$ (20%); $\mathcal{D}_{\text{ref}} \cap \mathcal{D}_{\text{val}} \cap \mathcal{D}_{\text{test}} = \varnothing$ (stratified split, seed = 42). $\mathcal{D}_{\text{ref}}$ provides TrueSkill ratings for LWRR calibration; $\mathcal{D}_{\text{val}}$ serves dimension search (Eq. (2)) and hyperparameter selection; $\mathcal{D}_{\text{test}}$ is held out for final evaluation. After excluding human "equal" labels, 75–78 test pairs per category remain.

Implementation. CLIP ViT-L/14 (768-d) for visual features; PCA to 8D for consensus prompts. GPT-4o [28] as the default VLM; VLM-backbone-agnostic design is verified with Qwen2.5-VL-72B [2] (Sec. 4.5). TrueSkill priors: μ=25, σ=8.33, draw probability 0.10; 5 images per consensus group (10 total). Stage 2: Observer τ=0.3, Debater τ=0.5, Judge τ=0.1; single-shot modes τ=0.0. Stage 3 defaults: K=20, τ_{kernel}=1.0, λ=1.0, ε=0.8, θ=0.6, α=0.3. E2E dimension search: T=15 trials with patience-based early stopping after 5 non-improving trials; explore phase τ_{gen}=0.85→1.0, converge phase τ_{gen}=0.7→0.5. Total VLM cost: ~$68 for the experimental dataset (4,819 pairs, Mode 4); wall-clock time ~2–4 h with concurrent API access; CLIP feature extraction requires a one-time GPU pass (~10 min), after which all pipeline stages run on CPU only (supplementary Section 6). For zero API cost, we deployed Qwen2.5-VL-72B (Apache-2.0) locally on 4×RTX 4090 GPUs, achieving equivalent accuracy (Sec. 4.5). The main

results (Tab. 1) use *per-category* optimised hyperparameters from a combined random search (N=1,000 configurations sampled from a 50,000-element grid on $\mathcal{D}_{\text{val}}$; supplementary Section 3), as the optimal K (10–50) and α (0.2–0.7) vary substantially across categories.

Baselines. *ResNet-50 Siamese*: twin-tower on frozen ResNet-50 features [16]; *CLIP Siamese*: same architecture on CLIP ViT-L/14 features; *Seg. + CLIP Regression*: SegFormer segmentation proportions + PCA-reduced CLIP, classified by Gradient Boosting [5]; *Zero-shot VLM*: GPT-4o with forced binary prompt (no dimensions or multi-agent reasoning). All supervised baselines use the reference split with mirror augmentation.

4.2 Main Results

Table 1 and Fig. 1 present the main comparison; Fig. 3 provides per-pair qualitative evidence. Key findings:

(1) UrbanAlign achieves 70.8% average accuracy (κ=0.41), outperforming the strongest same-feature control (Global Ridge on Δ_{sem}: 61.2%) by **+9.6 pp** and zero-shot VLM (56.7%) by +14.1 pp, with 81.6% (κ=0.63) on *safety*; the gap over same-feature controls confirms that local calibration—not feature quality alone—is the key differentiator. (2) LWRR calibration provides +9.9 pp average boost (60.9%→70.8%), with the largest effect on *safety* (+18.1 pp) and *depressing* (+12.4 pp). (3) The gap between *safety* (81.6%) and *boring* (65.3%) reflects perceptual grounding: safety correlates with concrete cues [13, 40], while boringness depends on subtler absence-of-stimulation signals. (4) Human split-half agreement on Place Pulse ranges from 78–87% across categories (supplementary Section 13); on *safety*, UrbanAlign matches this ceiling (81.8% vs. 81.6%), while on other categories a 6–16 pp gap remains. Test-set bootstrap 95% CIs (1,000 resamples with fixed optimised parameters) average [61.0%, 80.0%], confirming stability despite moderate test-set sizes (75–78 excl-equal pairs per category); all lower bounds exceed the 56.7% zero-shot baseline.

4.3 Dimension-Level Discriminability

Table 2 shows discriminability varies substantially: *wealthy* (avg. 68.0%) and *safety* (64.7%) are highest, while *depressing* (42.1%) is weakest. For *wealthy*, top dimensions (*Façade Quality*, *Street Cleanliness*, *Pavement Integrity*, all 71.2%) align with broken-windows theory [40]. LWRR's aggregate accuracy (74.0%) exceeds any single dimension, confirming the multi-dimensional bottleneck captures complementary information.

Why does the pipeline outperform end-to-end alternatives? Each stage tightens a specific term in the accuracy decomposition (Eq. (8)): concept mining maximises $\bar{\delta}$; multi-agent scoring reduces σ_S^2; LWRR minimises calibration error locally. The hybrid space (α=0.3) is critical: CLIP provides neighbourhood structure while semantic scores supply domain signals.

Table 1: Comparison on Place Pulse 2.0 across all six perception categories. All values reported as excl-equal (incl-equal). UrbanAlign outperforms all baselines including same-feature controls (70.8% vs. 61.2% best baseline).

	Safety		Beautiful		Lively	
Method	Acc.(%)	κ	Acc.(%)	κ	Acc.(%)	κ
ResNet50 Siamese [16]	54.8 (54.0)	0.16 (0.18)	58.6 (56.0)	0.18 (0.19)	45.7 (42.5)	−0.07 (−0.06)
CLIP Siamese [31]	66.7 (64.4)	0.34 (0.31)	61.4 (57.3)	0.23 (0.20)	48.2 (44.8)	−0.01 (−0.01)
Seg. Regression [5]	61.9 (59.8)	0.25 (0.23)	57.1 (53.3)	0.15 (0.13)	51.9 (48.3)	0.04 (0.04)
GPT-4o zero-shot [28]	61.9 (59.8)	0.23 (0.21)	62.9 (58.7)	0.25 (0.21)	49.4 (46.0)	0.01 (0.01)
BT Logistic [3]†	65.5 (63.2)	0.31 (0.29)	58.6 (54.7)	0.17 (0.15)	56.8 (52.9)	0.17 (0.15)
Global Ridge†	66.7 (64.4)	0.33 (0.31)	61.4 (57.3)	0.22 (0.19)	58.0 (54.0)	0.19 (0.17)
UrbanAlign (Ours)	**81.6** (76.5)	**0.63** (0.56)	**72.4** (67.9)	**0.45** (0.40)	**64.5** (60.5)	**0.28** (0.25)
	Wealthy		Boring		Depressing	
Method	Acc.(%)	κ	Acc.(%)	κ	Acc.(%)	κ
ResNet50 Siamese [16]	47.0 (44.8)	0.01 (0.00)	48.7 (45.0)	−0.00 (−0.02)	53.3 (50.0)	0.06 (0.05)
CLIP Siamese [31]	57.8 (55.2)	0.19 (0.17)	50.0 (47.5)	0.06 (0.07)	58.7 (55.0)	0.19 (0.16)
Seg. Regression [5]	61.5 (58.6)	0.26 (0.23)	62.2 (57.5)	0.22 (0.19)	54.7 (52.5)	0.09 (0.11)
GPT-4o zero-shot [28]	60.2 (57.5)	0.21 (0.19)	48.7 (45.0)	0.06 (0.05)	57.3 (53.8)	0.14 (0.12)
BT Logistic [3]†	68.7 (65.5)	0.38 (0.35)	50.0 (46.3)	0.01 (0.01)	58.7 (55.0)	0.17 (0.15)
Global Ridge†	69.9 (66.7)	0.40 (0.37)	54.1 (50.0)	0.10 (0.08)	57.3 (53.8)	0.15 (0.13)
UrbanAlign (Ours)	**74.4** (71.6)	**0.47** (0.43)	**65.3** (60.5)	**0.30** (0.26)	**66.7** (64.2)	**0.33** (0.30)

*Excl-equal excludes "equal" labels; model "equal" predictions count as errors. $\mathcal{D}_{\text{test}}$: 75–78 pairs/cat. Same GPT-4o backbone for UrbanAlign and zero-shot. †Trained on UrbanAlign's Δ_{sem} with $\mathcal{D}_{\text{ref}}$ labels.

Interpretability. LWRR weights reveal per-pair top contributors (e.g. Building Modernity w=2.30, Vegetation Maintenance w=1.72), directly actionable for urban planners. Figure 3 visualises per-pair reasoning; "Corrected" cases show LWRR overturning incorrect raw predictions. A survey of 126 urban perception papers confirms 80% of our auto-mined dimensions appear in existing literature; 20 domain experts selected top-5 dimensions per category from a mixed pool, with our dimensions achieving 85.9% selection rate.

4.4 Ablation Study

We ablate four scoring modes crossing two factors: input context (single-image vs. pairwise) and reasoning protocol (single-shot vs. multi-agent), yielding Modes 1–4 (Tab. 4).

Component analysis. Structured dimension scoring (Mode 2 raw, 61.7%) already exceeds zero-shot (56.7%) by +5.0 pp while creating an interpretable concept bottleneck. Multi-agent deliberation (Mode 4 raw, 60.9%) produces similar accuracy but lower dimension-score variance. LWRR leverages this for +9.9 pp (70.8%). Table 1 uses original Stage-1 dimensions with per-category optimised LWRR hyperparameters (N=1,000 random search on $\mathcal{D}_{\text{val}}$; Supp. C); E2E dimension search achieves 69.3% (Supp. F).

Fig. 3: Qualitative examples across six perception categories. **A** denotes the human-selected winner (green border). *Centre chart*: per-dimension concept scores for image A (green) and image B (blue) on the left half, and LWRR contribution towards A on the right half; the y-axis lists dimension names and bar length represents each dimension's contribution to $\hat{\delta}_q$ (Eq. (5)); positive LWRR bars support A, negative bars support B. The summary line reports the raw score comparison and calibrated LWRR delta; "Corrected" marks cases where calibration overturns an incorrect raw prediction. *Right chart*: baseline confidence bars (orange = conf. A, purple = conf. B, grey = equal); UrbanAlign is shown in bold.

Table 2: Per-dimension discriminative power (Mode 4). "Power" = fraction (%) of test pairs where the score difference on that single dimension correctly predicts the human winner (50% = chance). n = test pairs; "avg" = mean power across all dimensions.

Safety (n=283, avg 64.7%)		**Beautiful** (n=125, avg 56.7%)		**Lively** (n=259, avg 53.0%)	
Dimension	Power	Dimension	Power	Dimension	Power
Greenery & Landscaping	71.2	Greenery & Nat. Elem.	64.4	Human Presence	59.6
Lighting Adequacy	71.2	Color Coordination	62.2	Building Diversity	57.7
Building Maintenance	69.2	Pedestrian Amenities	55.6	Street Furniture	57.7
Pedestrian Infrastructure	65.4	Street Cleanliness	55.6	Color & Lighting	53.8
Street Activity	57.7	Architectural Coherence	55.6	Commercial Activity	53.8
Visibility Clarity	53.8	Visual Complexity	46.7	Public Art & Signage	46.2
				Vegetation & Greenery	42.3
Wealthy (n=117, avg 68.0%)		**Boring** (n=91, avg 43.8%)		**Depressing** (n=95, avg 42.1%)	
Dimension	Power	Dimension	Power	Dimension	Power
Façade Quality	71.2	Activity Presence	52.1	Street Cleanliness	47.9
Street Cleanliness	71.2	Color Vibrancy	50.0	Visual Clutter	45.8
Pavement Integrity	71.2	Street Furn. Presence	47.9	Façade Condition	43.8
Vegetation Maintenance	69.2	Architectural Diversity	43.8	Lighting Quality	39.6
Lighting Quality	67.3	Visual Complexity	39.6	Greenery Presence	33.3
Infrastructure Condition	67.3	Spatial Enclosure	37.5		
Vehicle Quality	65.4	Vegetation Variety	35.4		
Building Modernity	61.5				

5–10 dims/category; E2E loop explores alternatives over 15 trials.

Table 3: Multi-agent scoring at equal API budget (6-cat avg, excl. "equal"). All rows use identical LWRR hyperparameters; only the scoring mechanism differs. †5-cat (*depressing* third run incomplete).

Scoring method	Calls/pair	Raw	+LWRR
1× Mode 2 (pairwise single-shot)	1	61.7	63.6
3× Mode 2 (majority vote)†	3	62.3	64.6
Mode 4 ODJ (multi-agent)	**3**	60.9	**70.8**

Is multi-agent scoring necessary at equal budget? (Tab. 3) Mode 4 uses three VLM calls per pair (Observer, Debater, Judge). To isolate the value of *structured* deliberation from the value of simply sampling more, we compare against a same-budget baseline: running the single-shot pairwise prompt (Mode 2) three times and aggregating by majority vote. All three variants share identical LWRR hyperparameters. After calibration, Mode 4 leads the single-shot Mode 2 by **+7.2 pp** (70.8% vs. 63.6%), whereas naive 3× repetition adds only +1.0 pp (63.6%→64.6%). The Observer–Debater–Judge chain produces lower-variance dimension scores, which is what lets LWRR extract a much larger calibration gain—confirming that the structure of the deliberation, not the call count, drives the improvement.

2×2 factorial (Tab. 4). Multi-agent reasoning improves accuracy by +5.0 pp (single-image) and +18.9 pp (pairwise). Pairwise context alone adds only +1.0 pp

but *amplifies* multi-agent reasoning (+16.3 pp, Mode 3→4). The two factors interact synergistically: Mode 4 exceeds Mode 1 by +21.3 pp, far exceeding the sum of individual effects (per-category breakdown in Supp. Table 13).

LWRR calibration gain (Tab. 5). +9.9 pp average (60.9%→70.8%), largest on *safety* (+18.1 pp) and *depressing* (+12.4 pp). Full per-mode breakdown in supplementary Section 7.

Table 4: 2×2 factorial averaged across all six categories (accuracy% / κ, post-LWRR). Per-category results in supplementary Table 13.

	Single-Shot	Multi-Agent	Multi-Agent Gain
Single-Image	50.9 / 0.08	55.9 / 0.18	+5.0
Pairwise	51.9 / 0.10	**70.8 / 0.41**	+18.9
Pairwise Gain	+1.0	+16.3	—

Table 5: LWRR calibration gain (Mode 4, excl. "equal", values as acc.% / κ). All categories improve; avg. +9.9 pp.

	Safety	Beaut.	Lively	Wealthy	Boring	Depress.	Avg
Raw	63.5 / .27	66.1 / .33	53.5 / .07	63.2 / .26	65.0 / .30	54.3 / .09	60.9 / .22
Aligned	81.6 / .63	72.4 / .45	64.5 / .28	74.4 / .47	65.3 / .30	66.7 / .33	70.8 / .41
Δ (pp)	**+18.1**	**+6.3**	**+10.9**	**+11.2**	**+0.3**	**+12.4**	**+9.9**

4.5 Sensitivity Analysis

Full grids in supplementary Section 3. $K_{\max}$ is most sensitive (+4.3 pp; K=20 best on average); α=0.3 (70% semantic) is optimal but per-category optima differ; τ_{kernel} and λ_{ridge} have limited effect (<2.5 pp); scoring thresholds ε and θ are robust. The E2E dimension loop (Eq. (2), 12/15 trials) yields 69.3% via per-category assembly; the main pipeline (70.8%) uses optimised calibration hyperparameters with original dimensions (supplementary Sections 3, 6).

Backbone generalization. Replacing GPT-4o with Qwen2.5-VL-72B [2] (deployed locally on 4×RTX 4090) and re-optimising Stage 3 hyperparameters yields 72.5% (κ=0.45) vs GPT-4o 72.2% (Δ=0.3 pp; same $N\geq 3$ data, per-backbone optimised LWRR; supplementary Section 10), confirming VLM-backbone-agnostic design.

Bootstrap stability. To quantify evaluation uncertainty, we perform 1,000 test-set bootstrap resamples with fixed optimised parameters. The 95% CIs are: safety [72.4, 89.5]%, beautiful [63.1, 81.6]%, lively [53.9, 73.7]%, wealthy [65.4, 83.4]%, boring [54.7, 74.7]%, depressing [56.4, 76.9]%. Every CI contains the reported point estimate and every lower bound exceeds the 50% majority baseline, confirming that results are stable despite moderate test-set sizes (75–78 excl-equal pairs per category).

Raw vs. calibrated trade-off. Mode 4 raw accuracy (60.9%) is marginally below Mode 2 (61.7%) by design: the ODJ chain trades raw winner prediction

for lower-variance dimension scores, which in turn yields a far larger LWRR calibration gain (+9.9 pp vs. +1.9 pp; Tab. 3).

Cross-dataset and cross-domain generalization. We run the full pipeline end-to-end on two datasets it was never designed for. On **SPECS** [30] (400 images, 10 categories, four absent from Place Pulse; 60 reference pairs/cat), UrbanAlign exceeds zero-shot on **8/10** categories (avg. 67.5%) and beats all supervised baselines. On **AVA** [24] (255K aesthetically-rated photographs, a different domain), it beats zero-shot at every difficulty tier (+4.1 pp overall, 76.7% vs. 72.6%) using five auto-mined aesthetic dimensions. All dimensions are mined automatically; full results in supplementary Section 11.

5 Conclusion

We presented **UrbanAlign**, a post-hoc calibration framework that transforms frozen VLMs from unreliable annotators (56.7%) into structured perception decoders (70.8%, κ=0.41) via interpretable semantic dimensions and locally-adaptive geometric calibration, without modifying any model weights. Its three stages—concept mining [17], multi-agent scoring [7], and locally-weighted ridge regression on a hybrid manifold—are mutually reinforcing: on Place Pulse 2.0 the combination surpasses all baselines, with the largest LWRR gains on *safety* (+18.1 pp) and *depressing* (+12.4 pp), while the per-pair LWRR weights remain directly interpretable.

Cross-dataset generalization. The pipeline generalises beyond Place Pulse to SPECS and AVA (Sec. 4.5; supplementary Section 11): UrbanAlign exceeds zero-shot on 8/10 SPECS categories and at every AVA difficulty level (+4.1 pp), with all dimensions mined automatically.

Limitations and future work. Urban perception labels can encode societal biases; UrbanAlign's interpretable dimension weights expose but do not eliminate such biases, and deployment for planning decisions should be paired with equity-aware review. Natural next steps include extending to cross-cultural settings [14, 30] where demographic factors modulate perception, and generalising to other pairwise-preference domains such as aesthetic quality. Code and data splits are publicly available. Extended discussion is in supplementary Section 12.

Acknowledgements. We thank the anonymous reviewers for their constructive feedback.

References

1. Atkeson, C.G., Moore, A.W., Schaal, S.: Locally weighted learning. Artificial Intelligence Review **11**, 11–73 (1997)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)

3. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika **39**(3–4), 324–345 (1952)
4. Caponnetto, A., De Vito, E.: Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics **7**(3), 331–368 (2007)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
6. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association **74**(368), 829–836 (1979)
7. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: ICML (2024)
8. Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C.A.: Deep learning the city: Quantifying urban perception at a global scale. In: ECCV. pp. 196–212 (2016)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. vol. 70, pp. 1321–1330 (2017)
10. Herbrich, R., Minka, T., Graepel, T.: TrueSkill: A Bayesian skill rating system. In: NeurIPS. pp. 569–576 (2006)
11. 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**(7) (2024)
12. Jacobs, J.: The Death and Life of Great American Cities. Random House (1961)
13. Jeffery, C.R.: Crime prevention through environmental design. American Behavioral Scientist **14**, 598–610 (1971)
14. Jin, R., Cai, C.: Plausible or misleading? evaluating the adaption of the Place Pulse 2.0 dataset for predicting subjective perception in Chinese urban landscapes. Habitat International **157** (2025)
15. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. pp. 133–142 (2002)
16. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop (2015)
17. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML. pp. 5338–5348 (2020)
18. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NeurIPS. pp. 9459–9474 (2020)
19. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. In: EMNLP (2024)
20. Liu, Y., Zhang, T., Gu, S.: Hybrid concept bottleneck models. In: CVPR (2025)
21. Liu, Y., Cao, J., Li, Z., He, R., Tan, T.: Breaking mental set to improve reasoning through diverse multi-agent debate. In: ICLR (2025)
22. Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulić, I., Korhonen, A., Collier, N.: Aligning with human judgement: The role of pairwise preference in large language model evaluators. In: COLM (2024)
23. Meng, Y., Michalski, M., Huang, J., Zhang, Y., Abdelzaher, T., Han, J.: Tuning language models as training data generators for augmentation-enhanced few-shot learning. In: ICML (2023)
24. Murray, N., Marchesotti, L., Perronnin, F.: AVA: A large-scale database for aesthetic visual analysis. In: CVPR. pp. 2408–2415 (2012)
25. Mushkani, R.: Do vision-language models see urban scenes as people do? an urban perception benchmark. arXiv preprint arXiv:2509.14574 (2025)

26. Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.A.: Streetscore—predicting the perceived safety of one million streetscapes. In: CVPR Workshop on Web-scale Vision and Social Media (2014)
27. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: ICLR (2023)
28. OpenAI: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
29. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS. pp. 27730–27744 (2022)
30. Quintana, M., Gu, Y., Liang, X., Hou, Y., Ito, K., Zhu, Y., Abdelrahman, M., Biljecki, F.: Global urban visual perception varies across demographics and personalities. Nature Cities **2**(11), 1092–1106 (2025)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
32. Ruppert, D., Wand, M.P.: Multivariate locally weighted least squares regression. The Annals of Statistics **22**(3), 1346–1370 (1994)
33. Salesses, P., Schechtner, K., Hidalgo, C.A.: The collaborative image of the city: Mapping the inequality of urban perception. PLoS ONE **8**(7), e68400 (2013)
34. Srivastava, D., Yan, G., Weng, T.W.: VLG-CBM: Training concept bottleneck models with vision-language guidance. In: NeurIPS (2024)
35. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented RLHF. In: Findings of ACL (2024)
36. Tan, Z., Li, Z., Liu, T., Wang, H., Yun, H., Zeng, M., Chen, P., Zhang, Z., Gao, Y., Wang, R., Nigam, P., Yin, B., Jiang, M.: Aligning large language models with implicit preferences from user-generated content. In: ACL (2025)
37. Thurstone, L.L.: A law of comparative judgment. Psychological Review **34**(4), 273–286 (1927)
38. Wang, C., Gan, Y., Huo, Y., Mu, Y., Yang, M., He, Q., Xiao, T., Zhang, C., Liu, T., Du, Q., Yang, D., Zhu, J.: RoVRM: A robust visual reward model optimized via auxiliary textual preference data. In: AAAI (2025)
39. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: ICLR (2023)
40. Wilson, J.Q., Kelling, G.L.: Broken windows: The police and neighborhood safety. Atlantic Monthly **249**(3), 29–38 (1982)
41. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In: ICML (2024)
42. Yan, Y., Wen, H., Zhong, S., Chen, W., Chen, H., Wen, Q., Zimmermann, R., Liang, Y.: UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In: WWW (2024)
43. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: CVPR (2023)
44. Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., Kong, L.: ZeroGen: Efficient zero-shot learning via dataset generation. In: EMNLP (2022)
45. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: ICLR (2023)

46. Zhan, Y., Zhu, Y., Zheng, S., Zhao, H., Yang, F., Tang, M., Wang, J.: Vision-R1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. arXiv preprint arXiv:2503.18013 (2025)
47. 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)
48. Zhang, Y., Zhao, R., Huang, Z., Wang, X., Ma, Y., Long, Y.: GenAI models capture urban science but oversimplify complexity. arXiv preprint arXiv:2505.13803 (2025)
49. Zhao, Y., Zhang, R., Xiao, J., Hou, R., Guo, J., Zhang, Z., Hao, Y., Chen, Y.: Object-level verbalized confidence calibration in vision-language models via semantic perturbation. arXiv preprint arXiv:2504.14848 (2025)