Scene-Aware Urban Design: A Human–AI Recommendation Framework Using Co-Occurrence Embeddings and Vision-Language Models

Rodrigo A. Gallardo
Department of Architecture
Massachusetts Institute of Technology
Cambridge, MA 02139
ragallar@mit.edu

Oz Fishman
Department of Architecture
Massachusetts Institute of Technology
Cambridge, MA 02139
oof0928@mit.edu

Alexander Htet Kyaw
Department of Architecture
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
alexkyaw@mit.edu

Abstract

This paper introduces a human-in-the-loop computer vision framework that uses generative AI to propose micro-scale design interventions in public space and support more continuous, local participation. Using Grounding DINO and a curated subset of the ADE20K dataset as a proxy for the urban built environment, the system detects urban objects and builds co-occurrence embeddings that reveal common spatial configurations. From this analysis, the user receives five statistically likely complements to a chosen anchor object. A vision language model then reasons over the scene image and the selected pair to suggest a third object that completes a more complex urban tactic. The workflow keeps people in control of selection and refinement and aims to move beyond top-down master planning by grounding choices in everyday patterns and lived experience.

1 Introduction

The quality of public urban space is shaped not only by large-scale master planning but also by small, everyday interventions: a bench beneath a tree, a bike rack near a storefront, or a shade structure in a plaza. The results of these micro-scale decisions are core to the lived experience of the city Koolhaas [1995]. For residents, the ability to make such changes to their surroundings is constrained by regulation, resource access, or a lack of design knowledge Velasco and Floreán [2018]. In their absence, municipalities' design of these spaces is often generic or devoid of the spatial knowledge held by residents who use them daily. Our work would help residents make situated recommendations for the improvement of their surroundings, anchored on specific objects, and in a decision environment that is constructed from AI-generated suggestions of which additional objects would most likely create a more successful urban environment Kindberg et al. [2000], Agre [1994].

Rather than automating design, this hybrid model uses machine learning to surface possibilities to allow humans to provide interpretation, judgment, and intent. This collaboration raises timely questions: How can non-human agents help us to create decision environments that enable more active public participation? What kinds of local knowledge or spatial rituals can be preserved as

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

machine-generated suggestions enter civic processes? And how might systems like this redistribute design agency, enabling new forms of engagement with the built environment? Busquets [2007]

2 State of the Art

Tangible user interfaces that integrate "physical" manipulation with real-time simulation allow residents and planners alike to test zoning and land-use scenarios through accessible, hands-on interfaces that link directly to computational simulations Noyman [2022]. This foregrounding of spatial cause-and-effect democratizes access to the many decisions regarding (and the implications of) involved in macro-level interventions (such as neighborhood density, mobility networks, or district-wide zoning), but offers limited usability for the everyday, object-scale decisions that define how individuals inhabit public space, especially at scale. Emerging developments in extended reality and computer vision have enabled a new mode of interaction with the built environment, treating physical objects as responsive, queryable agents Kindberg et al. [2000]. Through real-time semantic segmentation and large language models, such systems allow users to direct prompts at material elements within their surroundings, effectively enabling interactions through augmented interfaces Liu et al. [2024], Dogan et al. [2024], Radford et al. [2021]. Recent research has also demonstrated the use of vision-language models to integrate real-world data and sustainability metrics into design pipelines. Gupta and Kyaw [2025] These interaction paradigms reframe urban elements as digital endpoints for dialog and design.

3 Methods

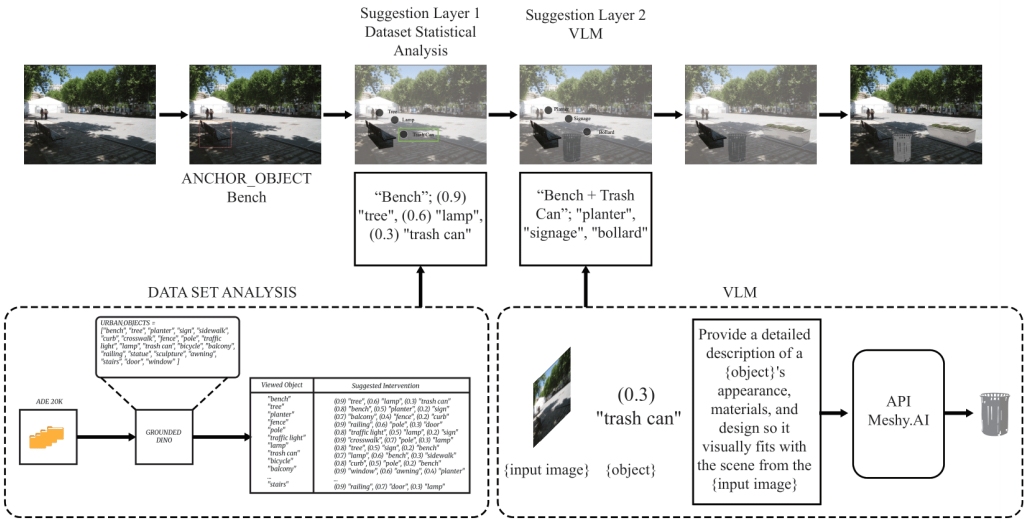

Figure 1: Overview Diagram

The proposed pipeline runs lightweight background object detection as scenes are processed. When a scene meets the criteria for a potential intervention, an anchor object is specified for that scene by the user. Once the anchor is set, the system initiates a two-branch decision process.

In the statistical branch, the system retrieves a short list of urban objects that most frequently co-occur with the anchor in socially active contexts, based on prior co-occurrence analysis using the ADE20K dataset. The user selects one of these statistically suggested co-occurring objects.

In the semantic branch, the selected anchor and co-occurring object pair is passed to a vision-language model, which proposes five additional objects that could be added to the scene using contextual and spatial cues. (see Appendix for the full system prompt).

This staged selection process, beginning with an anchor object chosen by the user and followed by a co-occurring object suggested by the system, ensures that the resulting set of objects is grounded in

real-world spatial relationships. The objective is to enable the creation of urban tactics composed of multiple interacting objects.

Once the five additional recommendations are generated, 3D models are produced for each object to give the user a visual preview. To ensure these models are visually integrated with the existing scene, we employ a vision language model (VLM). Given an image of the scene, the model is prompted to "Provide a detailed description of an object's appearance, materials, and design that could visually fit with the scene from the input image." The output description is then used as input to a text-to-3D generation model, producing scene-consistent 3D assets that the user can select and situate in the urban space.

In its current stage, user participation occurs at two levels: (1) the user selects an anchor object and chooses among system-suggested co-occurring elements, and (2) the user may iteratively accept, reject, or re-prompt VLM-generated recommendations before committing them to the scene. This iterative loop preserves user intent, ensuring that each regenerated scene remains grounded in human selection and interpretation rather than autonomous replacement.

3.1 Dataset Analysis

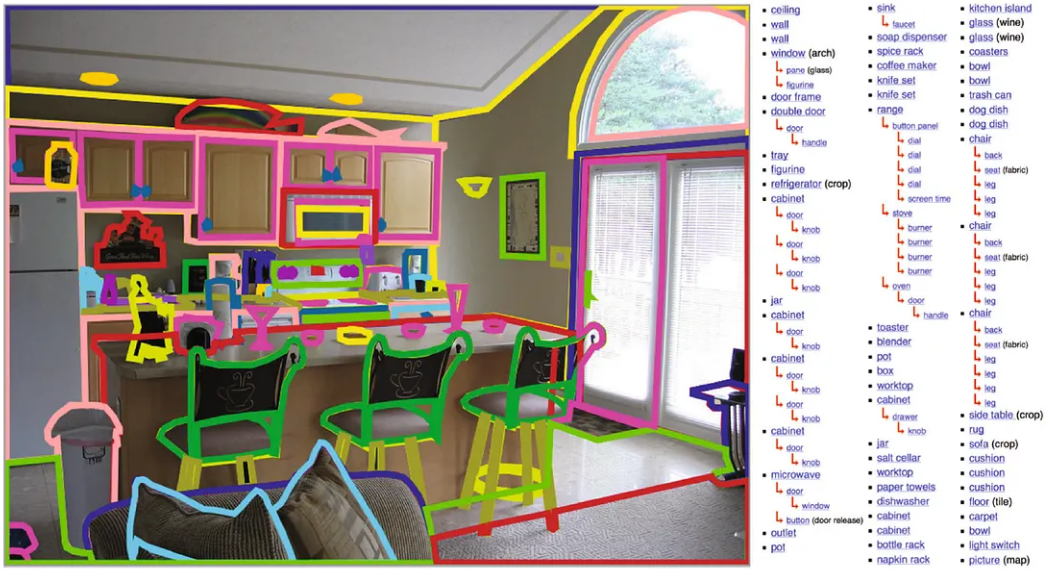

Figure 2: ADE20K image with listed objects and their parts

The framework's object analysis relies on running the Grounding DINO model ADE20K dataset. ADE20K offers over 20,000 labeled images spanning indoor and outdoor environments, with dense semantic segmentation and scene categorization Zhou et al. [2017]. Its breadth makes it one of the few large-scale datasets with labeled public space scenes that include furniture, vegetation, signage, and infrastructure.

Grounding DINO was selected for its open-vocabulary, zero-shot detection capabilities Liu et al. [2024]. However, the model's reliance on text prompts introduces potential inconsistency in detection quality, especially in visually cluttered or low-resolution images. Preliminary testing also revealed false positives in highly textured scenes (e.g., planters confused with trash cans), highlighting the importance of visual post-processing and vocabulary refinement.

Specifically, scenes are only included if Grounding DINO detects five or more people with confidence above a specified threshold. This heuristic approximates scenes with higher levels of public activity, based on the assumption that activated space is correlated with human presence. While this filter does not guarantee spatial proximity or engagement, it serves as a practical proxy for identifying images where the co, occurrence of urban objects may relate to actual use and behavior.

This two-stage filtering process, first by scene category, then by population threshold, produces a set of approximately 900 images, each annotated with object detection results. Object detection is performed using Grounding DINO model. This approach allows for the detection of urban elements beyond closed, class labels, which is essential for analyzing micro-scale features often absent from conventional object detection datasets.

3.2 Co-Occurrence Aggregation and Embedding

Figure 3: 900 Image Co-Occurrence Matrix

Once urban objects are detected in each filtered scene, the system constructs a pairwise co-occurrence matrix to capture which elements frequently appear together. For every qualifying image, all unordered pairs of detected urban objects are recorded in a symmetric matrix indexed by object class. If a scene includes a bench, a tree, and a trash can, for example, the matrix entries for (bench, tree), (bench, trash can), and (tree, trash can) are incremented. This aggregation process is repeated across the dataset to produce a global co-occurrence matrix spanning all scenes of interest.

$$P(o_j \mid o_i) = \frac{\text{count}(o_i \wedge o_j)}{\text{count}(o_i)}$$

The raw counts in the matrix are then normalized to create co-occurrence probability vectors for each object. Each row of the matrix is divided by its sum, yielding a vector that reflects the likelihood of other objects appearing in a scene given the presence of a particular object. These vectors function as contextual embeddings, not learned via backpropagation, but constructed empirically from real, world spatial data.

$$o = \big[P(o_1 \mid o),\, P(o_2 \mid o),\, \ldots,\, P(o_n \mid o)\big]$$

3.3 Vision Language Model

Vision-language models can read style, materials, and spatial cues directly from images. After the user picks an item from the co-occurrence list, the model receives three inputs: the full scene image, tight crops and normalized boxes for the anchor and the selected co-occurring object, and a short scene summary with scene type, a five-color palette, dominant materials, and a coarse near–mid–far depth sketch.

The model returns five candidates in a compact schema: object type, key materials and finish, approximate size, color hints, simple placement guidance relative to the anchor or pair, and a one-sentence justification. This format is short and easy to parse, while still carrying the information needed for the next stage. To keep suggestions practical, we filter out items that the current scene cannot support. For example, a crosswalk is often associated with a bench, but it is removed when the image does not show a street edge or an intersection. Similar feasibility checks look at ground type, clearance, and obvious access constraints before ranking the final five.

3.4 Mesh Generation

Off-the-shelf text-to-3D models can produce meshes that match a target style when given a clear, structured prompt. In our workflow, the VLM output is converted into a concise brief and sent to the Meshy API noa [2025a] for a first-pass mesh. The system is model agnostic, so other generators can be swapped in and benchmarked for speed, geometry fidelity, UV coverage, and texture quality.

We normalize scale to real-world units, set the pivot at ground level, and generate a low-poly LOD for AR. If a mesh fails, we regenerate with a tightened prompt.

4 Results

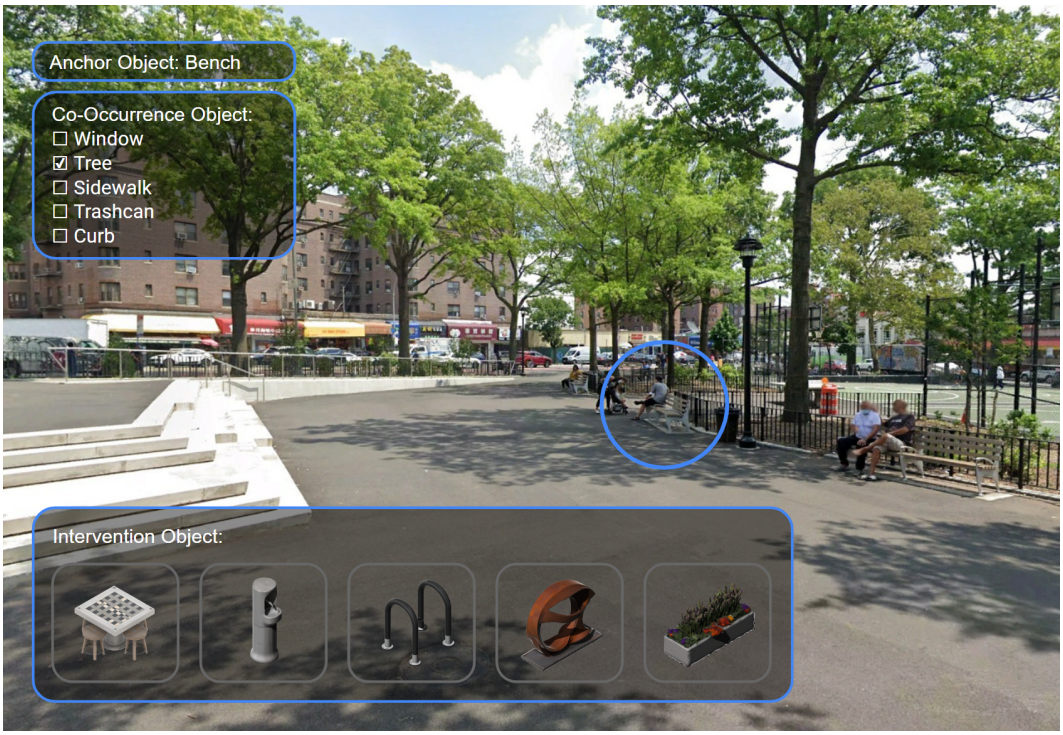

Figure 4: Pilot Interface in Urban Scene

A pilot version of the interface was created to demonstrate the two steps involved in completing an urban tactic, using a bench in the image as the anchor object. A similar interface could be implemented in an AR environment with a headset, giving the user the ability to place and adjust models in 3D space.

Across varied urban scenes, the pipeline produced placeable suggestions. In Layer 1, the co-occurrence embedding provided in Appendix A.1 is used to provide object pairs that can complements to the anchor object, such as benches with trees, trash cans, signs, and lamp posts. We also tested the VLM with various prominent anchor objects and various co-occurrence pairings with VLM generated recommendations, as detailed in Appendix A.3.

Examples from four selected scenes with different prominent anchor objects are shown in Appendix A.2, where the VLM produced contextually specific suggestions such as bus stops, kiosks, and wayfinding signage. These are the kind of items that a predefined co-occurrence list alone would be not be able to recommend. The ability to use a VLM in this step is critical because it can combine visual context, spatial cues, and object semantics to generate recommendations that go beyond statistical pairings, adapting to the specific scene and producing more functionally relevant, locally appropriate, and visually coherent suggestions.

Failure modes included weak detections in cluttered scenes, occasional near-duplicate or policy-sensitive suggestions in Layer 2, and certain object generation errors from the text-to-3D stage. In particular, Meshy AI sometimes failed to capture all prompt details. For example, chess table patterns were rendered inaccurately, and a drinking fountain described with a lower basin did not reproduce that feature correctly (Figure 5). These limitations highlight the need for a tighter vocabulary, region-aware guidance, stronger negative examples, and improvements in the downstream 3D generation stage.

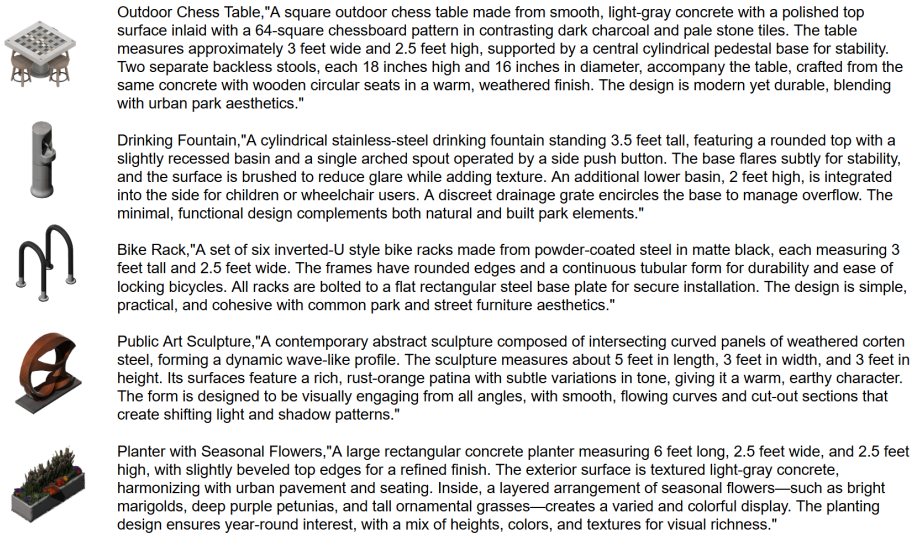

Figure 5: Mesh generated using VLM generted recommendation and discription

5 Limitations and Future Work

The current system estimates object co-occurrence based on pixel distance in 2D images, which does not capture true 3D spatial relationships and may reduce precision when proposing site-specific interventions. In addition, the dataset is biased toward certain geographic and cultural contexts, limiting generalization. Future directions include integrating 3D scene understanding from RGB-D or multi-view video, adapting the interface for continuous use on AR headsets, expanding datasets for greater diversity, and incorporating post-deployment feedback to refine suggestions. There is also potential to link the system with civic reporting platforms noa [2025b], allowing proposed interventions to be shared with local authorities and, over time, letting real-world outcomes inform future recommendations.

Vision–language models (VLMs) face several limitations when applied to recommendation tasks in urban design contexts. They may struggle to account for nuanced, location-specific social, cultural, or legal constraints, leading to suggestions that, while visually plausible, may be socially inappropriate or non-compliant with local regulations. VLMs also often lack access to critical non-visual factors such as spatial measurements, traffic patterns, and accessibility requirements, which are essential for determining the practical suitability of an object in a given location. Furthermore, biases in training data, particularly if it is dominated by imagery from specific regions such as North America or Europe can result in recommendations that overlook local architectural styles, materials, and street usage patterns in other regions. Finally, VLMs cannot directly assess installation feasibility, meaning they are unable to consider site-specific constraints such as terrain, underground utilities, or budget.

Future iterations will ideally include a participatory evaluation, where users can directly edit proposed scenes. This will allow measurement of how the AI's recommendations influence, align with, or diverge from user intent, helping to refine regeneration fidelity and interpretability.

6 Conclusion

This framework is both a technical system and a design argument. As a system, it operationalizes co-occurrence embeddings and semantic reasoning within a human-in-the-loop AR interface, enabling residents to propose micro-scale interventions grounded in statistical spatial patterns and their own situated knowledge. As an argument, it contends that the city's smallest parts (benches, trees, lights) are as consequential to urban life as its master plans, yet remain largely inaccessible to participatory design at scale. By placing computational intelligence in the immediacy of the site, the framework transforms the urban environment itself into a living model, where design emerges from dialogue between people, objects, and algorithms. In doing so, it reframes AI as a collaborator in everyday spatial authorship, embedding lived experience directly into civic processes and advancing a model of local spatial intervention.

References

API Platform - Meshy, August 2025a. URL https://www.meshy.ai/en/api.

Snap Send Solve, 2025b. URL https://www.snapsendsolve.com/how-it-works.

Philip E. Agre. Surveillance and capture: Two models of privacy. *The Information Society*, 10(2):101–127, April 1994. ISSN 0197-2243. doi: 10.1080/01972243.1994.9960162. URL https://doi.org/10.1080/01972243.1994.9960162. Publisher: Routledge _eprint: https://doi.org/10.1080/01972243.1994.9960162.

Joan Busquets. Defining the Urbanistic Project:. 2007.

Mustafa Doga Dogan, Eric J. Gonzalez, Karan Ahuja, Ruofei Du, Andrea Colaço, Johnny Lee, Mar Gonzalez-Franco, and David Kim. Augmented Object Intelligence: Making the Analog World Interactable with XR-Objects. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, October 2024. doi: 10.1145/3654777.3676379. URL http://arxiv.org/abs/2404.13274. arXiv:2404.13274 [cs].

Richa Gupta and Alexander Htet Kyaw. Insights Informed Generative AI for Design: Incorporating Real-world Data for Text-to-Image Output, June 2025. URL http://arxiv.org/abs/2506.15008. arXiv:2506.15008 [cs].

T. Kindberg, J. Barton, J. Morgan, G. Becker, D. Caswell, P. Debaty, G. Gopal, M. Frid, V. Krishnan, H. Morris, J. Schettino, B. Serra, and M. Spasojevic. People, places, things: Web presence for the real world. In *Proceedings Third IEEE Workshop on Mobile Computing Systems and Applications*, pages 19–28, Los Alamitos, CA, USA, 2000. IEEE Comput. Soc. ISBN 978-0-7695-0816-0. doi: 10.1109/MCSA.2000.895378. URL http://ieeexplore.ieee.org/document/895378/.

Rem Koolhaas. Whatever Happened to Urbanism? 1995.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, July 2024. URL http://arxiv.org/abs/2303.05499. arXiv:2303.05499 [cs].

Ariel Noyman. An Urban modeling and Simulation Platform. 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Carlos Zedillo Velasco and Diego Pérez Floreán. Housing and SDG's In Mexico. 2018.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.544. URL http://ieeexplore.ieee.org/document/8100027/.

A Technical Appendix

A.1 Co-Occurence Matrices

Figure 6: Co Ocurrence Matrices per sub folder

The system was tested on a curated subset of the ADE20K dataset, filtered to include approximately 300 urban scenes with high pedestrian density. For each scene, urban object detections were processed using Grounding DINO, and co, occurrence statistics were aggregated into a global matrix across all scenes of interest. The resulting matrix revealed clear and intuitive spatial associations among frequently detected objects.

For example, the object class “bench” co, occurred most often with “tree,” “trash can,” “sign,” and “lamp post.” Similarly, “planter” frequently appeared with “fence,” “balcony,” and “sidewalk.” These pairings align with common public space configurations and validate the hypothesis that co, occurrence vectors can reflect real, world design patterns.

A.2 Embeddings

Table 1: Top 5 Co-Occurrence Embeddings

Anchor Object	Embedding 1	Embedding 2	Embedding 3	Embedding 4	Embedding 5
bench	window	tree	sign	traffic light	crosswalk
tree	traffic light	window	sidewalk	door	planter
planter	tree	sidewalk	window	balcony	traffic light
sign	traffic light	window	crosswalk	tree	sidewalk
sidewalk	window	traffic light	tree	planter	sign
curb	window	sign	traffic light	sidewalk	crosswalk
crosswalk	traffic light	window	sign	tree	sidewalk
fence	window	sidewalk	tree	planter	traffic light
pole	window	traffic light	tree	sign	crosswalk
traffic light	sign	window	tree	crosswalk	sidewalk
lamp	window	door	tree	stairs	sidewalk
trash can	window	tree	traffic light	sign	door
bicycle	window	traffic light	pole	fence	sidewalk
balcony	planter	tree	sidewalk	fence	door
railing	window	pole	bicycle	fence	sidewalk
stairs	traffic light	window	sidewalk	tree	door
door	tree	window	traffic light	sidewalk	crosswalk
window	traffic light	sidewalk	pole	fence	bicycle

A.3 System Prompt

User Prompt You are given an input image of an urban scene along with an anchor object and a co-occurrence object. Based on the scene in the image, propose five options for a third object that would plausibly fit into the scene in a socially active public space.

For each object, provide:

Object – The object's name.

Description – A detailed paragraph that can be directly used for text-to-3D generation. This description must specify the object's appearance, materials, approximate scale, color palette that matches the scene, style cues, likely placement relative to the anchor and co-occurrence objects, and functional details such as geometry, key components, and reasonable dimensions. Ground every detail in the visual context of the provided image so that the 3D model will integrate naturally into the scene. However, do not specify where the object is supposed to be placed in the scene. It should be an independent object without a background. Avoid brand names, unsafe elements, or features that block primary circulation.

Only output the result as a CSV file with exactly two columns: Object and Description. Do not include any other commentary or formatting.

Inputs:

Scene image: filename

Anchor object: object1

Co-occurrence object: object2

A.4 Additional Examples and Scene Outputs from VLM-Generated Recommendations and Descriptions for 3D Generative AI.

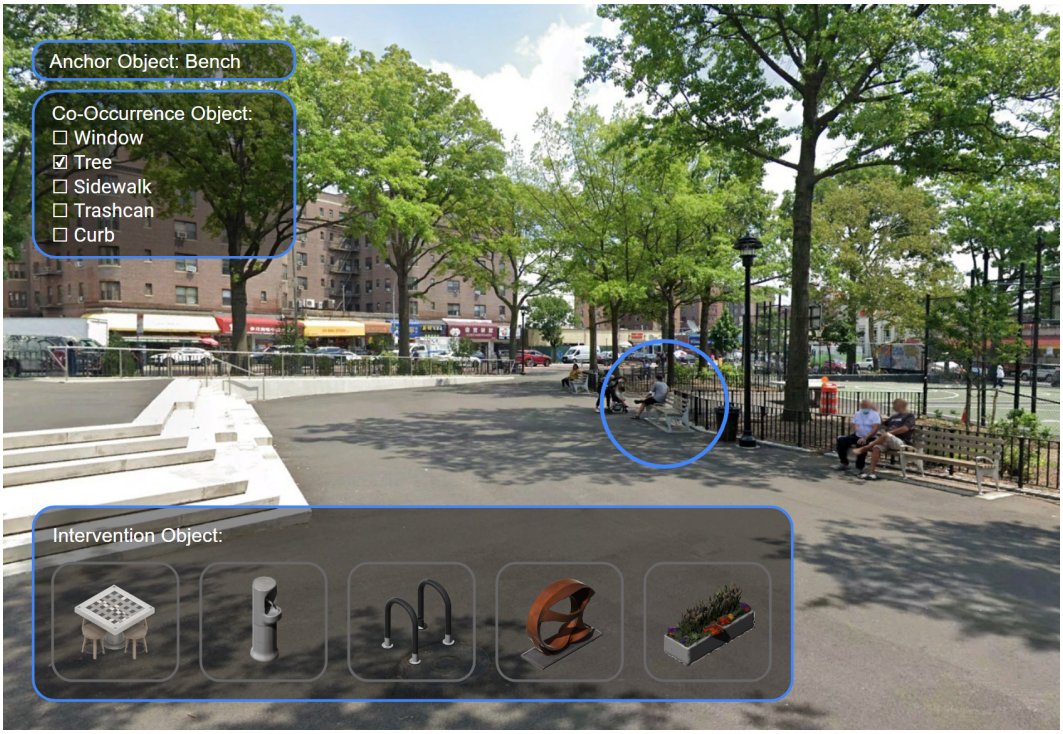

Figure 7: Pilot Interface in Urban Scene 2

Outdoor Chess Table,"A square outdoor chess table made from smooth, light-gray concrete with a polished top surface inlaid with a 64-square chessboard pattern in contrasting dark charcoal and pale stone tiles. The table measures approximately 3 feet wide and 2.5 feet high, supported by a central cylindrical pedestal base for stability. Two separate backless stools, each 18 inches high and 16 inches in diameter, accompany the table, crafted from the same concrete with wooden circular seats in a warm, weathered finish. The design is modern yet durable, blending with urban park aesthetics."

Drinking Fountain,"A cylindrical stainless-steel drinking fountain standing 3.5 feet tall, featuring a rounded top with a slightly recessed basin and a single arched spout operated by a side push button. The base flares subtly for stability, and the surface is brushed to reduce glare while adding texture. An additional lower basin, 2 feet high, is integrated into the side for children or wheelchair users. A discreet drainage grate encircles the base to manage overflow. The minimal, functional design complements both natural and built park elements."

Bike Rack,"A set of six inverted-U style bike racks made from powder-coated steel in matte black, each measuring 3 feet tall and 2.5 feet wide. The frames have rounded edges and a continuous tubular form for durability and ease of locking bicycles. All racks are bolted to a flat rectangular steel base plate for secure installation. The design is simple, practical, and cohesive with common park and street furniture aesthetics."

Public Art Sculpture,"A contemporary abstract sculpture composed of intersecting curved panels of weathered corten steel, forming a dynamic wave-like profile. The sculpture measures about 5 feet in length, 3 feet in width, and 3 feet in height. Its surfaces feature a rich, rust-orange patina with subtle variations in tone, giving it a warm, earthy character. The form is designed to be visually engaging from all angles, with smooth, flowing curves and cut-out sections that create shifting light and shadow patterns."

Planter with Seasonal Flowers,"A large rectangular concrete planter measuring 6 feet long, 2.5 feet wide, and 2.5 feet high, with slightly beveled top edges for a refined finish. The exterior surface

Figure 8: Pilot Interface in Urban Scene 3

is textured light-gray concrete, harmonizing with urban pavement and seating. Inside, a layered arrangement of seasonal flowers—such as bright marigolds, deep purple petunias, and tall ornamental grasses—creates a varied and colorful display. The planting design ensures year-round interest, with a mix of heights, colors, and textures for visual richness."

Park Bench,"A classic slatted park bench made from light beige painted wood with a black powder-coated steel frame. The bench measures approximately 6 feet in length and 3 feet in height, with a gently curved backrest and armrests at both ends for comfort. The wooden slats are evenly spaced and smooth-finished, offering durability and weather resistance. The steel frame has rounded legs with flat feet for stable ground contact. The color palette of muted beige and black harmonizes with both natural greenery and urban surroundings."

Drinking Fountain,"A modern drinking fountain crafted from brushed stainless steel, standing 3.5 feet tall with a cylindrical base and a slightly flared bottom for stability. The top has a shallow, concave basin with a chrome water spout activated by a push button. An additional side-mounted spout at 2 feet high allows accessibility for children or wheelchair users. The brushed finish reduces glare and resists fingerprints, while a circular drainage grate at the base prevents water pooling."

Bike Rack,"A modular bike rack system made from matte black powder-coated steel, featuring five inverted-U hoops aligned in a row. Each hoop stands 3 feet tall and 2.5 feet wide, made from a continuous tubular frame with smooth bends. The hoops are mounted on a rectangular steel base plate with concealed anchor bolts. The design is minimal and functional, complementing typical park fixtures and providing secure locking points for bicycles."

Information Kiosk,"A freestanding information kiosk with a rectangular frame made from dark bronze-finished aluminum, standing 6.5 feet tall and 3.5 feet wide. The top section features a clear, weather-resistant acrylic display case for maps, schedules, or community notices, with an internal LED strip for illumination. The lower section has a lockable storage compartment for maintenance tools or brochures. The overall design is sleek, durable, and styled to integrate with urban park elements."

Planter with Seasonal Flowers,"A large square planter measuring 4 feet on each side and 2.5 feet high, made from textured light-gray concrete with chamfered edges. Inside, a mix of seasonal plants

Figure 9: Pilot Interface in Urban Scene 4

such as vibrant yellow marigolds, deep purple petunias, and tall ornamental grasses create a layered, colorful display. The concrete's neutral tone complements surrounding elements, while the plant arrangement provides year-round visual interest with varied heights and textures."

Public Seating Bench,"A contemporary public bench with a frame constructed from galvanized steel and seating surfaces made of sustainably sourced hardwood slats. The bench measures 6 feet in length, 2 feet in depth, and 3 feet in height, featuring a gently angled backrest for ergonomic comfort. The steel frame has a matte silver finish to complement surrounding urban fixtures, while the warm-toned wood slats contrast subtly with the cool pavement. The bench design incorporates slightly curved armrests and flat steel legs with base plates for stability and durability in high-traffic public spaces."

Bicycle Rack,"A modular bicycle rack made of black powder-coated tubular steel, designed in a wave pattern for both aesthetic appeal and multiple locking points. Each wave loop stands 3 feet tall and extends 2 feet in width, with smooth, continuous curves to prevent damage to bicycles. The rack is mounted on a heavy-duty steel base rail with concealed anchor bolts, ensuring secure installation. Its clean, minimalist style and durable materials match the modern urban streetscape and resist weathering from sun and rain."

Drinking Fountain,"A sleek stainless-steel drinking fountain standing 3.5 feet high, with a cylindrical pedestal base that tapers slightly toward the top. The fountain head has a push-button spout at a comfortable angle for adults, with a secondary, lower-mounted spout 2 feet high for children and wheelchair users. The brushed steel finish reduces glare and hides fingerprints, while a small grated drain surrounds the base to prevent puddling. The streamlined design fits seamlessly with contemporary city infrastructure."

Public Art Sculpture,"A small-scale abstract public art piece made from powder-coated aluminum panels in interlocking geometric shapes. The sculpture measures approximately 4 feet in height and 3 feet in width, with bold color accents in muted red, golden yellow, and steel gray to complement the tones of surrounding buildings and fixtures. The surfaces are smooth with a matte finish, and the structure is anchored to a flat steel base plate. Its compact size and modern aesthetic make it both a focal point and an unobtrusive addition to a busy streetscape."

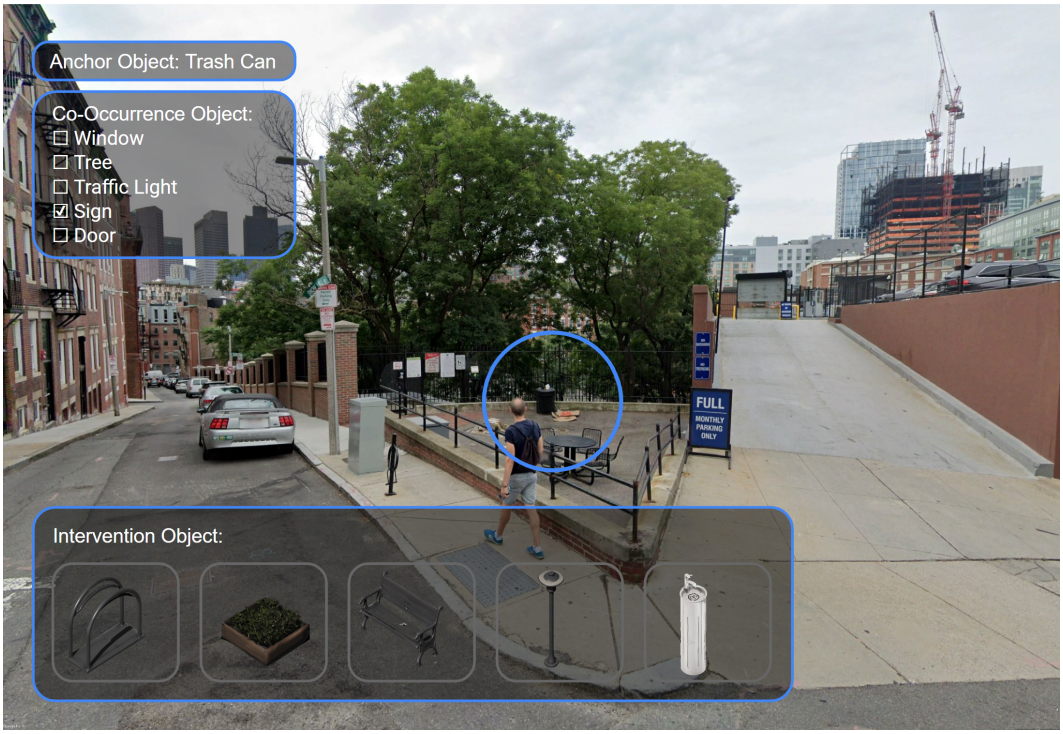

Figure 10: Pilot Interface in Urban Scene 5

Wayfinding Signpost,"A freestanding wayfinding signpost made from dark bronze anodized aluminum with a rectangular column design. Standing 8 feet tall and 1.5 feet wide, it features high-contrast printed panels displaying directional arrows, maps, and neighborhood points of interest. The top section includes an LED-lit panel for nighttime visibility, while the base is reinforced with a stainless-steel kick plate for durability. The clean lines and muted color palette ensure the signpost integrates harmoniously into an urban pedestrian environment."

Bicycle Rack,"A black powder-coated steel bicycle rack with a simple inverted U-shape design, suitable for securing up to two bikes. The rack is made from round steel tubing approximately 5 cm in diameter, bent into a smooth arch with two vertical legs fixed to a rectangular steel base plate. The finish is matte to reduce glare, and the height is about 80 cm with a width of 90 cm, proportionate to pedestrian-scale street furniture. The style matches the utilitarian aesthetic of the surrounding urban environment, with clean lines and durability for outdoor use."

Planter Box,"A rectangular planter box constructed from weathered dark-brown cedar wood planks reinforced with black steel corner brackets. Measuring approximately 120 cm in length, 45 cm in width, and 50 cm in height, it contains low-maintenance greenery such as ornamental grasses and small flowering plants in muted greens and pale yellows that blend with the scene's natural tones. The base has concealed drainage slots, and the top edges are slightly chamfered for a finished look, integrating natural texture into the urban setting."

Public Bench,"A three-seat metal bench with a powder-coated black steel frame and horizontal slats forming the seat and backrest. The bench is 180 cm long, 60 cm deep, and 80 cm high, with armrests at each end for comfort and accessibility. The slats are slightly curved for ergonomic support, and the style mirrors the robust, low-maintenance design typical of city street furniture. The matte finish complements the dark metal tones present in the surrounding railings and trash receptacle."

Street Lamp,"A tall, slender street lamp post approximately 3.5 meters in height, made from galvanized steel painted in a dark matte gray finish to match nearby metalwork. The lamp head features a slightly flared conical shade with a frosted glass diffuser, providing warm white LED lighting. The post tapers gently toward the top, with a simple base plate and bolted flanges for stability. Its

Figure 11: Pilot Interface in Urban Scene 6

minimalistic, functional design blends with the mix of historic brick facades and modern urban infrastructure."

Drinking Fountain,"A pedestal-style drinking fountain made from cast stainless steel with a brushed finish to resist fingerprints and weathering. The column stands about 95 cm tall with a round basin at the top and a polished chrome push-button spout. A small drainage grate is integrated into the basin to prevent standing water. The cylindrical body is sleek with subtle vertical grooves for visual interest, harmonizing with the modern utilitarian character of nearby street furniture while offering a practical amenity."

Bus Stop Shelter,"A rectangular bus stop shelter with a powder-coated steel frame in dark gray, featuring three vertical side panels made from tempered glass with a slight greenish tint. The roof is flat and slightly overhangs, constructed from opaque polycarbonate sheets to provide shade while allowing diffused light. The shelter measures approximately 3 meters in length, 1 meter in depth, and 2.4 meters in height. Inside, a single wooden bench with metal legs is mounted along the back panel. The design is minimalistic yet functional, blending with the urban streetscape mixed materials of concrete, asphalt, and painted metal."

Street Vendor Cart,"A mobile street vendor cart constructed from stainless steel with a polished surface, equipped with a canopy in red-and-white striped fabric for shade. The cart measures about 1.8 meters in length, 0.8 meters in width, and 1.5 meters in height to the canopy top. It includes side shelves for displaying goods, a small storage compartment beneath the main counter, and rubberized wheels for mobility. The style is vibrant and practical, complementing the colorful facades and signage in the surrounding area."

Planter Barrier,"A rectangular concrete planter serving as a traffic barrier, finished in a light gray tone with a slightly rough surface texture. The planter is approximately 1.2 meters long, 0.5 meters wide, and 0.6 meters tall. It contains hardy green shrubs and small flowering plants in yellow and white, adding natural contrast to the asphalt and concrete. Its utilitarian form integrates greenery while doubling as a subtle protective element for pedestrians."

Public Bench,"A metal-framed public bench with horizontal wooden slats stained in a medium brown tone, designed to seat up to three people. The bench is about 1.8 meters long, 0.6 meters deep, and

0.8 meters high, with armrests at each end for comfort. The frame is powder-coated in dark gray, echoing the muted tones of street infrastructure. The slats are slightly curved for ergonomic support, providing both aesthetic appeal and practical durability in an outdoor urban environment."

Information Kiosk,"A freestanding rectangular kiosk made of powder-coated aluminum in dark green, with a lockable glass front panel displaying local maps, transit schedules, and event posters. The kiosk measures approximately 1.5 meters in height, 0.8 meters in width, and 0.3 meters in depth. The base has a reinforced steel plate for stability, and the top features a small overhanging canopy to protect the display from rain and direct sunlight. Its color and proportions are consistent with other street fixtures, making it blend naturally into the urban scene."

A.5 VLM Suggestions

Table 2: VLM Top 5 (Stacked) for All Anchors

Anchor Object	Embedding Object	VLM Third Objects
balcony	door	crosswalk hydrant bollard fence awning
balcony	fence	traffic light kiosk bicycle rack info board curb
balcony	planter	bicycle rack curb bollard bench sidewalk
balcony	sidewalk	curb lamp trash can crosswalk hydrant
balcony	tree	window sidewalk door crosswalk curb
bench	crosswalk	info board lamp bollard awning street art
bench	sign	info board sidewalk trash can street art planter
bench	traffic light	tree lamp door planter bike lane marking
bench	tree	bike lane marking bicycle rack seating wall stairs awning
bench	window	trash can bicycle rack curb bike lane marking street art
bicycle	fence	door bollard

Anchor Object	Embedding Object	VLM Third Objects
		awning bike lane marking curb
bicycle	pole	awning kiosk bollard crosswalk bicycle rack
bicycle	sidewalk	hydrant fence crosswalk tree seating wall
bicycle	traffic light	crosswalk sidewalk awning lamp stairs
bicycle	window	bicycle rack hydrant bollard planter fence
crosswalk	sidewalk	stairs traffic light bicycle rack door trash can
crosswalk	sign	planter lamp curb stairs fence
crosswalk	traffic light	trash can window sidewalk bicycle rack fence
crosswalk	tree	fence seating wall bench traffic light awning
crosswalk	window	tree bench sign trash can curb
curb	crosswalk	planter awning door bollard fence
curb	sidewalk	traffic light kiosk lamp

Anchor Object	Embedding Object	VLM Third Objects
		hydrant street art
curb	sign	door seating wall crosswalk window lamp
curb	traffic light	seating wall sidewalk fence awning crosswalk
curb	window	planter bench bollard bicycle rack street art
door	crosswalk	kiosk awning planter bollard sign
door	sidewalk	bench tree crosswalk curb bike lane marking
door	traffic light	planter crosswalk sign trash can bollard
door	tree	window awning planter bench street art
door	window	awning sign curb bollard sidewalk
fence	planter	curb street art bike lane marking lamp info board
fence	sidewalk	hydrant sign trash can street art door
fence	traffic light	window street art awning crosswalk

Anchor Object	Embedding Object	VLM Third Objects
		info board
fence	tree	curb bike lane marking traffic light trash can door
fence	window	bench bicycle rack door lamp info board
lamp	door	traffic light kiosk tree window info board
lamp	sidewalk	curb awning hydrant stairs bollard
lamp	stairs	bollard seating wall kiosk street art fence
lamp	tree	sign stairs trash can bollard kiosk
lamp	window	bike lane marking info board traffic light planter tree
planter	balcony	street art trash can tree fence lamp
planter	sidewalk	hydrant crosswalk street art kiosk traffic light
planter	traffic light	door awning fence trash can stairs
planter	tree	fence trash can info board sign window

Anchor Object	Embedding Object	VLM Third Objects
planter	window	fence bench bike lane marking traffic light lamp
pole	crosswalk	sidewalk bench curb bike lane marking kiosk
pole	sign	stairs curb info board hydrant seating wall
pole	traffic light	planter bike lane marking bicycle rack hydrant tree
pole	tree	trash can traffic light window bollard fence
pole	window	awning bicycle rack stairs bollard sidewalk
railing	bicycle	bollard traffic light hydrant fence stairs
railing	fence	window kiosk info board tree stairs
railing	pole	fence bollard bicycle rack trash can crosswalk
railing	sidewalk	traffic light kiosk trash can sign seating wall
railing	window	hydrant street art bicycle rack traffic light bench
sidewalk	planter	crosswalk

Anchor Object	Embedding Object	VLM Third Objects
		tree trash can bike lane marking street art
sidewalk	sign	seating wall planter traffic light awning street art
sidewalk	traffic light	awning window planter trash can kiosk
sidewalk	tree	kiosk crosswalk window traffic light door
sidewalk	window	curb trash can info board street art planter
sign	crosswalk	trash can stairs bench traffic light window
sign	sidewalk	curb trash can hydrant planter info board
sign	traffic light	hydrant window bike lane marking sidewalk curb
sign	tree	fence bicycle rack sidewalk planter bench
sign	window	awning bench crosswalk bike lane marking bicycle rack
stairs	door	fence curb bollard bike lane marking sign
stairs	sidewalk	bollard kiosk

Anchor Object	Embedding Object	VLM Third Objects
		window bike lane marking lamp
stairs	traffic light	window lamp hydrant tree info board
stairs	tree	sign bollard traffic light window fence
stairs	window	bicycle rack bench bike lane marking planter kiosk
traffic light	crosswalk	bike lane marking awning door trash can curb
traffic light	sidewalk	planter awning info board bike lane marking tree
traffic light	sign	sidewalk info board door lamp kiosk
traffic light	tree	hydrant sidewalk bollard lamp bike lane marking
traffic light	window	door bollard stairs planter fence
trash can	door	tree seating wall window bike lane marking fence
trash can	sign	crosswalk tree stairs bike lane marking kiosk
trash can	traffic light	window tree stairs

Anchor Object	Embedding Object	VLM Third Objects
		crosswalk street art
trash can	tree	bench curb info board kiosk bike lane marking
trash can	window	info board tree bicycle rack street art bollard
tree	door	lamp crosswalk street art kiosk planter
tree	planter	awning kiosk fence bollard sidewalk
tree	sidewalk	awning bike lane marking kiosk bench bicycle rack
tree	traffic light	awning bicycle rack stairs kiosk trash can
tree	window	kiosk stairs crosswalk bicycle rack sign
window	bicycle	sign hydrant crosswalk awning planter
window	fence	bollard stairs street art bike lane marking info board
window	pole	awning hydrant trash can curb traffic light
window	sidewalk	lamp stairs traffic light hydrant

Anchor Object	Embedding Object	VLM Third Objects
		awning
window	traffic light	fence trash can bollard door bike lane marking

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Sections 1 and Abstract) clearly state the contribution, and the claims match the methods, experiments, and scope of the work.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses dataset bias, 2D-based co-occurrence limitations, and hardware accessibility constraints, along with possible future improvements.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not contain formal theorems or proofs, focusing instead on methodological and experimental contributions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 and Appendix X describe dataset filtering, object detection settings, co-occurrence computation, and AR integration in sufficient detail for reproduction.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Anonymized code and data preprocessing scripts are included in the supplemental material with instructions for reproducing the main results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Dataset splits, model settings, confidence thresholds, and prompt details are described in Section 3 and Appendix X.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results include mean $\pm$ standard deviation over multiple runs, capturing variability from different scene samples and random seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix X specifies GPU type, memory, and average runtime for object detection, embedding computation, and AR preview rendering.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work follows the NeurIPS Code of Ethics, with transparency in dataset use, acknowledgment of bias, and discussion of societal impacts.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section X discusses benefits for community-driven design and risks such as reinforcing biased associations or privacy concerns from video capture.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The system does not release high-risk generative models or datasets and uses only publicly available datasets with controlled prompts.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models, including ADE20K and Grounding DINO, are cited with their licenses and source URLs.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release a new dataset or model.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects research requiring IRB approval.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM use in the semantic branch for contextual object recommendations is described in Section 3.1 as part of the core methodology.