CoStyler: Indoor Furniture Styling by Counterfactual Reasoning in a Hypergraph Field

Anonymous authors
Paper under double-blind review

Abstract

Fixed-layout indoor furniture styling requires selecting assets that form a coherent room while preserving furniture categories, positions, orientations, and scales. Existing approaches typically retrieve each asset independently or rely on static local relations, making them prone to shape, material, and color conflicts after scene composition. We introduce CoStyler, a scene-level structured selection framework built on a dynamic hypergraph style field. It uses a frozen multimodal large language model to extract style priors, maintains candidate distributions for furniture slots, and captures higher-order dependencies through style-conditioned hyperedge activation and weighting. Counterfactual style preference learning evaluates candidates as local substitutions using Mahalanobis energies to measure contextual compatibility. Training alternates between optimizing the style field and candidate logits; inference freezes the model and iteratively updates only room-specific candidate logits to resolve cross-slot conflicts. Experiments on 3D-FRONT show that CoStyler outperforms object- and scene-level retrieval baselines in furniture retrieval and scene-level style coherence.

1 Introduction

An asset that matches a target style in isolation may become conspicuously incompatible when placed with other furniture in a three-dimensional room. This observation motivates fixed-layout indoor furniture styling: given a target-style description, a prescribed room layout, and a large asset library, the system must assign one asset to each furniture slot while preserving its category, position, orientation, and scale. Unlike general indoor scene synthesis, this task cannot avoid an incompatible combination by changing the spatial arrangement. Its fundamental challenge is therefore to bridge object-level semantic relevance and scene-level aesthetic coherence. Addressing this challenge is important for virtual interior design, 3D content creation, and immersive embodied environments.

Existing approaches mainly follow two paradigms. Object-level cross-modal retrieval methods use pretrained representations such as CLIP, ULIP, and OpenShape (Radford et al., 2021; Xue et al., 2023; Liu et al., 2023) to measure the similarity between a style description and an individual asset. Although these representations retrieve semantically relevant furniture, they cannot determine whether an asset remains appropriate after scene composition, where quality also depends on its color, material, form, and spatial relation to surrounding furniture. Relation-based methods instead model a scene with graphs or hypergraphs (Wald et al., 2020; Lv et al., 2024; Feng et al., 2019; Jiang et al., 2019). However, most rely on fixed pairwise edges or a static topology and therefore cannot express higher-order furniture dependencies that vary with the layout and requested style. Both paradigms also typically terminate after a single retrieval or ranking pass, without revising local choices according to the evolving scene context.

Our key insight is that an isolated furniture asset carries only partial style semantics, whereas room style emerges from the spatial, functional, and global relations among multiple assets. Fixed-layout furniture styling should therefore be formulated as scene-level structured selection rather than a collection of independent retrieval problems. Under this formulation, the preferred candidate for one slot depends on the current choices at other slots, while each local change alters the global style context.

Based on this insight, we propose CoStyler, a counterfactual reasoning framework built on a dynamic hypergraph style field. A frozen multimodal large language model extracts structured room- and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

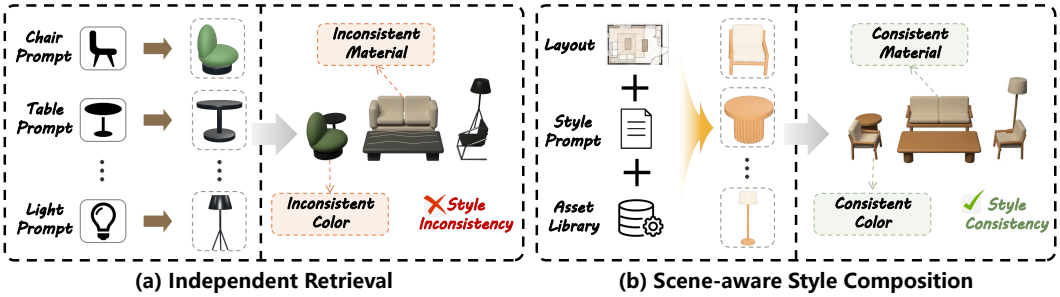


Figure 1: Comparison of two fixed-layout furniture-styling paradigms. (a) Independent object-level matching can combine individually relevant assets into a scene with shape, material, and color conflicts. (b) CoStyler jointly optimizes all slots to produce a coherent room-level assignment.

slot-level style priors from the open-ended request and fixed layout. Instead of committing each slot to a top-ranked asset, CoStyler maintains a distribution over its candidates throughout optimization, allowing scene context to continuously reshape local preferences. The dynamic hypergraph style field represents furniture slots as distributional nodes and conditionally activates and weights layout-induced hyperedges according to the target style. Local hyperedges capture spatially or functionally related furniture groups, while a global hyperedge summarizes the room, enabling higher-order style dependencies to emerge from local-to-global propagation.

To convert scene compatibility into candidate preferences, we introduce counterfactual style preference learning. Each candidate is treated as a local substitution in the current style field while the remaining slot distributions are held fixed. Candidate- and scene-level Mahalanobis energies then measure the contextual compatibility of the resulting configuration. Their learned dimension-wise tolerances distinguish restrictive style cues from weakly relevant variation, providing an adaptive measure of conflicts in color, material, and form. CoStyler alternates between two optimization stages during training: with candidate distributions fixed, preference ranking learns an energy field that assigns lower energy to coherent configurations; with the field fixed, the resulting counterfactual energies update the candidate logits. At inference, the multimodal large language model and learned style field remain frozen, and test-time optimization updates only the room-specific candidate logits. The furniture assignment can consequently improve with its evolving global context instead of terminating after one-shot retrieval.

Experiments on 3D-FRONT show that CoStyler achieves state-of-the-art furniture retrieval and scene-level style coherence. Ablation and qualitative analyses further support the effectiveness of dynamic higher-order reasoning, Mahalanobis energy, and iterative refinement.

Our contributions are as follows:

- We introduce a **dynamic hypergraph style field** that represents furniture slots as candidate distributions and uses style-conditioned hyperedge activation and propagation to capture local-to-global higher-order style dependencies.
- We propose **counterfactual style preference learning**, which evaluates each candidate as a local substitution in the current style field and uses candidate- and scene-level Mahalanobis energies to measure its contextual compatibility.
- We develop an **alternating optimization and test-time optimization strategy** that learns the style field and candidate logits in alternating stages, then freezes the field and iteratively refines room-specific candidate distributions at inference.

2 Related Work

3D Asset Retrieval. Vision–language pretraining provides shared semantic spaces for open-vocabulary 3D asset retrieval. CLIP establishes transferable image–text representations, while ULIP, ULIP-2, OpenShape, and Uni3D align language, images, and point clouds for scalable 3D representation learning (Radford et al., 2021; Xue et al., 2023; 2024; Liu et al., 2023; Zhou et al., 2024); Point-Bind and OpenScene extend multimodal alignment to richer 3D understanding (Guo et al.,

2023; Peng et al., 2023). These methods primarily score text–asset relevance and cannot determine whether locally plausible furniture remains coherent after composition in color, material, and form. MART (Yue et al., 2025) fine-tunes a multimodal retriever with interaction feedback and preference learning to account for the effectiveness of retrieved trajectories in downstream embodied tasks. Prior work learns style from furniture compatibility, real-world co-occurrence, or scene images (Liu et al., 2015; 2019; Weiss et al., 2020), and exploits local context or joint layout–object encodings for recommendation (Savva et al., 2017; Pan et al., 2025). For fixed-layout furniture selection, CoStyler uses frozen models to initialize retrieval priors, represents all slots as coupled candidate distributions, and jointly optimizes their style compatibility under a shared higher-order style field.

Indoor Scene Synthesis. Indoor scene synthesis has evolved from example-based arrangements and relation-graph modeling (Fisher et al., 2012; Wang et al., 2019) to autoregressive generation on large furnished-scene datasets (Fu et al., 2021; Paschalidou et al., 2021). Recent approaches use scene graphs, diffusion, and structured generation or editing to model layout, geometry, and appearance (Zhai et al., 2023; Tang et al., 2024; Ju et al., 2024; Zhao et al., 2024; Lin & MU, 2024; Zhai et al., 2024). Large language and vision–language models further support instruction following, commonsense constraints, and style-conditioned planning (Yang et al., 2024; Feng et al., 2023; Sun et al., 2025; Marshall et al., 2025; Pan & Liu, 2026; Berdoz et al., 2026), while related work studies multimodal generation, implicit representations, editable synthesis, and texturing (Yang et al., 2025; Liang et al., 2025; Zheng et al., 2025; Höllein et al., 2023; Huang et al., 2025). These methods generate or modify object categories, poses, geometry, or appearance. We address a complementary, more constrained problem: selecting real library assets that jointly realize a target style while categories, positions, orientations, and scales remain fixed.

Higher-Order Structured Inference. Scene graphs encode object relations for 3D scene understanding (Wald et al., 2020; Lv et al., 2024), but fixed pairwise edges cannot fully express the groupwise material, color, form, and spatial relations that define indoor style. Hypergraph neural networks model multi-node relations, and dynamic variants adapt higher-order connectivity to the input (Feng et al., 2019; Jiang et al., 2019). ED-HNN (Wang et al., 2023) models higher-order relations through equivariant hypergraph diffusion operators. Recent work further studies dynamic and multi-hop hypergraph reasoning (Zhou et al., 2023; Xie et al., 2025; Li et al., 2025); in other structured prediction domains, multi-scale graph inference and distributional candidate generation model cross-node dependencies and output uncertainty (Dang et al., 2021; 2022). In parallel, energy-based learning assigns low energy to compatible configurations (LeCun et al., 2006). Structured prediction energy networks and iterative energy minimization support global inference over coupled outputs (Belanger & McCallum, 2016; Du et al., 2022), while compositional energy models combine multiple constraints (Du et al., 2020). CoStyler applies these ideas to style-aware asset selection: the target style activates and weights layout-induced hyperedges, nodes represent candidate distributions rather than fixed features, and learned Mahalanobis energies evaluate counterfactual local substitutions.

3 METHOD

3.1 PROBLEM FORMULATION

Given a fixed layout, a target-style description, and an asset library $\mathcal{A}$, we select one asset for each of N furniture slots while preserving their categories, positions, orientations, and scales. Slot i has category c_i, geometric descriptor g_i, and a category-constrained candidate set $\mathcal{C}_i = \{x_i^k\}_{k=1}^K$. CoStyler maintains $p_i = \mathrm{softmax}(\alpha_i)$ with optimizable logits $\alpha_i \in \mathbb{R}^K$ and jointly optimizes all slot distributions to obtain a style-aligned, coherent assignment $\hat{\mathcal{X}} = \{x_i^{\hat{k}_i}\}_{i=1}^N$. Figure 2 presents the framework.

3.2 STYLE-PRIOR CANDIDATE INITIALIZATION

A frozen MLLM generates room-level style and slot-specific retrieval descriptions from the target description and fixed layout. Their embeddings z_r and z_i encode the shared palette, materials, and form language and their category-specific realization, respectively. The room description provides

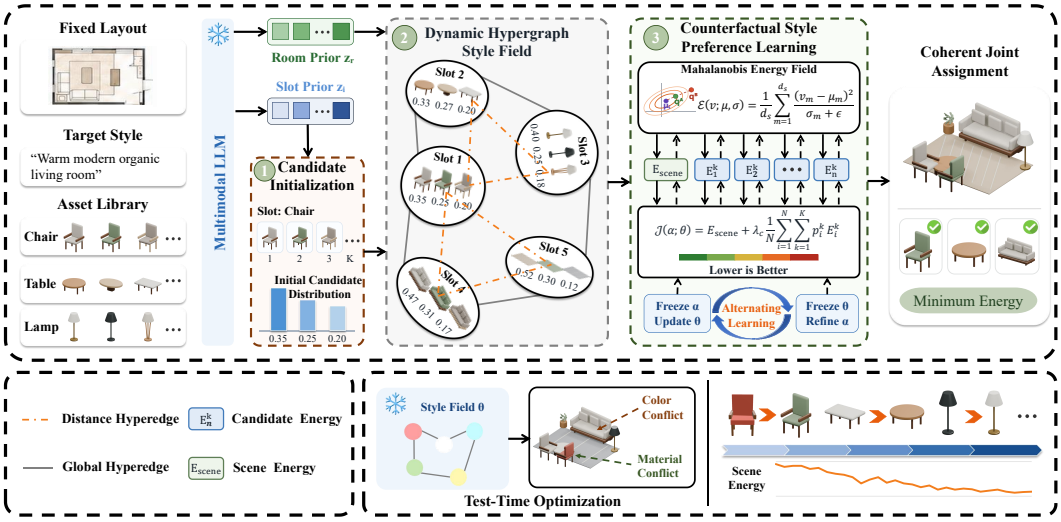


Figure 2: Overview of CoStyler. A frozen MLLM generates structured style priors and initializes candidate distributions. A dynamic hypergraph propagates higher-order context, while counterfactual substitutions are scored by Mahalanobis energy. Training alternates field and logit updates; TTO optimizes room-specific candidate logits.

a common style reference, while slot descriptions translate it into category-appropriate retrieval criteria, allowing different furniture categories to express the same target style in their own forms. A frozen multimodal encoder indexes multi-view asset renders and uses the slot description to retrieve the top-K category-compatible candidates and their similarities $\{s_i^k\}_{k=1}^K$ for slot i. We initialize candidate probabilities as $p_i^{k,0} = \exp(s_i^k/\tau)/\sum_{j=1}^K \exp(s_i^j/\tau)$ and the corresponding logits as $\alpha_i^{k,0} = \gamma \log(p_i^{k,0} + \epsilon)$; subsequent optimization uses $p_i = \mathrm{softmax}(\alpha_i)$. Here, τ is the retrieval temperature, γ controls initialization sharpness, and ϵ ensures numerical stability. For given retrieval similarities, a lower τ concentrates the initial probabilities $p_i^{k,0}$ on candidates with higher similarity, whereas a higher τ reduces probability differences between candidates, yielding a flatter distribution. Retrieval similarities provide an initial prior on candidate–description alignment but do not directly measure compatibility after scene composition.

3.3 Dynamic Hypergraph Style Field

Distributional slot nodes. Let $f_i^k \in \mathbb{R}^{d_f}$ denote the frozen visual–semantic feature of candidate x_i^k. Slot i has soft representation $\bar{f}_i = \sum_{k=1}^K p_i^k f_i^k$ and uncertainty $H(p_i) = -\sum_{k=1}^K p_i^k \log p_i^k$. A learnable encoder combines these with geometry and style priors to initialize $h_i^0 = \phi_n([\bar{f}_i, g_i, z_i, z_r, H(p_i)])$. The soft representation varies continuously with candidate probabilities, allowing the style field to account for asset choices before discretization. Entropy additionally describes distribution concentration, helping distinguish slots that favor a particular candidate from those with greater selection uncertainty.

Style-conditioned propagation. Following hypergraph message passing (Feng et al., 2019; Jiang et al., 2019), we construct $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from the fixed layout. Local hyperedges connect spatially or functionally related slots, and a global hyperedge covers the room. Node-to-edge aggregation and edge-to-node propagation exchange information among slots sharing a hyperedge to form group-level context. Local hyperedges provide context from related furniture groups, while the global hyperedge conveys room-wide information to each slot. Each edge has structural descriptor r_e; its activation and propagation strength depend on the target style and current candidate distributions. Let u_e^ℓ denote the edge state at layer ℓ and π_e^ℓ its activation probability. Given π_e^ℓ, we aggregate node messages as $m_e^\ell = \pi_e^\ell \sum_{i \in e} a_{i \to e}^\ell W_n h_i^\ell$ and update the edge state through a residual connection as $u_e^{\ell+1} = \mathrm{LN}(u_e^\ell + \phi_e(m_e^\ell))$. The hyperedge activation probability, propagation weight, and node

update are defined as

$$\begin{aligned} \pi_e^\ell &= \sigma\big(\phi_{\text{sel}}([\text{mean}_{i\in e}\, h_i^\ell, r_e, z_r])\big)\,, \\ \omega_e^\ell &= \pi_e^\ell \sigma\big(\phi_w([u_e^{\ell+1}, r_e, z_r])\big)\,, \\ h_i^{\ell+1} &= \text{LN}\left(h_i^\ell + \sum_{e\ni i} b_{e\to i}^\ell \omega_e^\ell W_e u_e^{\ell+1}\right). \end{aligned} \tag{1}$$

Here, ω_e^ℓ is the activation- and content-dependent edge weight. Attention $a_{i\to e}^\ell$ is normalized within each edge, and $b_{e\to i}^\ell$ over edges incident to node i. After L layers, h_i^L and u_g^L represent slot context and the global room state, allowing higher-order interactions to adapt to style requests and candidate distributions.

3.4 Counterfactual Style Preference Learning

For candidate x_i^k, we temporarily replace p_i with the one-hot vector $\mathbf{e}_k$, hold all other slot distributions fixed, and recompute the style field. Candidates are thus compared as local substitutions under the same remaining configuration.

We evaluate scene- and candidate-level compatibility. The room state is projected as $\tilde{u}_g = W_g u_g^L$, with a diagonal Gaussian prototype (μ_g, σ_g) predicted from z_r. The candidate representation is $q_i^k = \phi_c([f_i^k, h_i^L, z_r])$, with prototype (μ_i, σ_i) predicted from $[h_i^L, z_r]$. Both use the mean diagonal Mahalanobis energy

$$\mathcal{E}(v; \mu, \sigma) = \frac{1}{d_s} \sum_{m=1}^{d_s} \frac{(v_m - \mu_m)^2}{\sigma_m + \epsilon}. \tag{2}$$

Scene and candidate energies are $E_{\text{scene}} = \mathcal{E}(\tilde{u}_g; \mu_g, \sigma_g)$ and $E_i^k = \mathcal{E}(q_i^k; \mu_i, \sigma_i)$, respectively. Lower energy indicates greater compatibility; learned diagonal scales control tolerance to deviations along different style dimensions.

To avoid degeneracy from directly minimizing energy, we learn the field by ranking positive and negative scenes. Positives use ground-truth assets; negatives replace one or two slots with random candidates or, with probability ρ, embedding-similar hard negatives. The field objective is

$$\begin{aligned} \mathcal{L}_{\text{rank}} &= \max\big(0, \delta + E_{\text{scene}}^+ - E_{\text{scene}}^-\big)\,, \\ \mathcal{L}_{\text{sparse}} &= \frac{1}{|\mathcal{E}|} \sum_e \left(\pi_e^+ + \pi_e^-\right), \\ \mathcal{L}_\theta &= \mathcal{L}_{\text{rank}} + \lambda_s \mathcal{L}_{\text{sparse}}, \end{aligned} \tag{3}$$

where δ is the ranking margin and $\lambda_s \geq 0$ weights the sparsity penalty that discourages indiscriminate edge activation.

3.5 Training and Inference Strategy

Training. Each round alternates two stages: with training-room candidate distributions fixed, update the style-field parameters θ using Eq. equation 3 for T_θ steps; then freeze θ and perform T_α candidate-logit updates for each room by minimizing

$$\mathcal{J}(\boldsymbol{\alpha}; \theta) = E_{\text{scene}} + \frac{\lambda_c}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p_i^k E_i^k. \tag{4}$$

This objective combines scene energy with mean expected candidate energy, balanced by λ_c. Each candidate update recomputes the style field and energies from the current distributions, coupling local selection with global context. Algorithm 1 in Appendix D provides the complete training procedure.

Inference. Initialize test-room logits $\boldsymbol{\alpha}_*^0$ as in Section 3.2 and freeze the MLLM, multimodal encoder, and style-field parameters. Optimize only $\boldsymbol{\alpha}_*$ using Eq. equation 4 for T steps, recomputing the dynamic hypergraph and counterfactual energies after each update. The final assignment selects $\hat{k}_i = \arg\max_k p_i^{k,T}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset. We conduct experiments on 3D-FRONT (Fu et al., 2021), using 7,100 rooms for training and 800 rooms for testing. The test set covers bedrooms, living rooms, dining rooms, and libraries/studies. For each room, we preserve the original furniture categories, positions, orientations, and scales and vary only the selected assets, thereby constructing fixed-layout indoor styling instances. Target-style descriptions are derived from room images showing the original furniture. The description-generation prompt and subsequent style-prior extraction protocol are provided in Appendix B.

Evaluation metrics. We report slot-level exact-match accuracy against reference assets: *Init. R@1* for CoStyler's retrieval prior and *Final R@1* for each method's final predictions. The fixed top-15 candidate sets have 88.9% reference-asset recall, bounding Final R@1. GPT-4o rates rendered scenes against target-style descriptions on a five-point scale across four dimensions: *Aesthetic* (AES; visual quality, balance, and atmosphere), *Color & Material* (C&M; coordination of colors, textures, and materials), *Style Coherence* (SC; target-style alignment and inter-asset consistency), and *Realism & Geometry* (R&G; realism, scale compatibility, and geometric plausibility). Scores are averaged over five independent assessments per scene. GPT-4o serves solely as an evaluator, with no role in retrieval, training, or TTO; the full protocol is in Appendix C. Ten experts additionally rate 200 randomly sampled scenes on the same dimensions and scale.

Baselines. We compare CoStyler with representative 3D asset retrieval methods. ULIP (Xue et al., 2023), OpenShape (Liu et al., 2023), and Uni3D (Zhou et al., 2024) measure object-level text–asset relevance. MetaFind (Pan et al., 2025) additionally encodes existing objects and layout context, providing a stronger scene-aware baseline. Because MetaFind has not released its source code, we reproduce its architecture, training objective, and retrieval procedure following the paper. All methods use the same data split, asset library, category constraints, fixed layouts, target-style descriptions, and top-15 candidate sets.

Implementation. We instantiate the frozen MLLM with Qwen3-VL (Bai et al., 2025) to generate room- and slot-level style priors and use the frozen Qwen3-VL-Embedding model (Li et al., 2026) to retrieve fifteen category-compatible candidates per slot. Neither foundation model is updated during training or inference. Training is conducted on a single NVIDIA A800-SXM4-80GB GPU. The dynamic hypergraph and energy field are trained using the alternating procedure. During each alternating cycle, we perform one model-parameter update followed by five candidate-logit updates, using learning rates of 1.0×10^{-4} and 1.0×10^{-3}, respectively. At inference, the MLLM, multimodal encoder, and learned style field remain frozen; only room-specific candidate logits are updated for 200 TTO steps.

4.2 QUANTITATIVE RESULTS

Table 1 shows improvements in retrieval accuracy and scene quality from object-level retrieval to scene-aware retrieval and then to CoStyler's joint optimization. This progression suggests that fixed-layout furniture selection requires both individual alignment with the target style and compatibility among assets. MetaFind's gains over object-level methods demonstrate the value of scene context, while CoStyler's further improvements support the effectiveness of continually updating that context and jointly refining slot choices. Changing one slot can alter the suitability of assets at other slots, making it necessary to reassess local choices against the updated composition. Higher expert ratings across all four dimensions further indicate that these refinements improve overall visual quality. Table 2 examines this process across levels of scene complexity. Compared with bedrooms, which contain fewer slots, living rooms require coordinating more furniture relationships and therefore remain more challenging. Nevertheless, improvements over initial retrieval across all room types show consistent benefits from using scene context to refine local choices.

4.3 QUALITATIVE ANALYSIS

Figure 3 compares all methods under identical layouts and target styles. Object-level baselines often retrieve assets that match the prompt individually but become visually disruptive in the composed

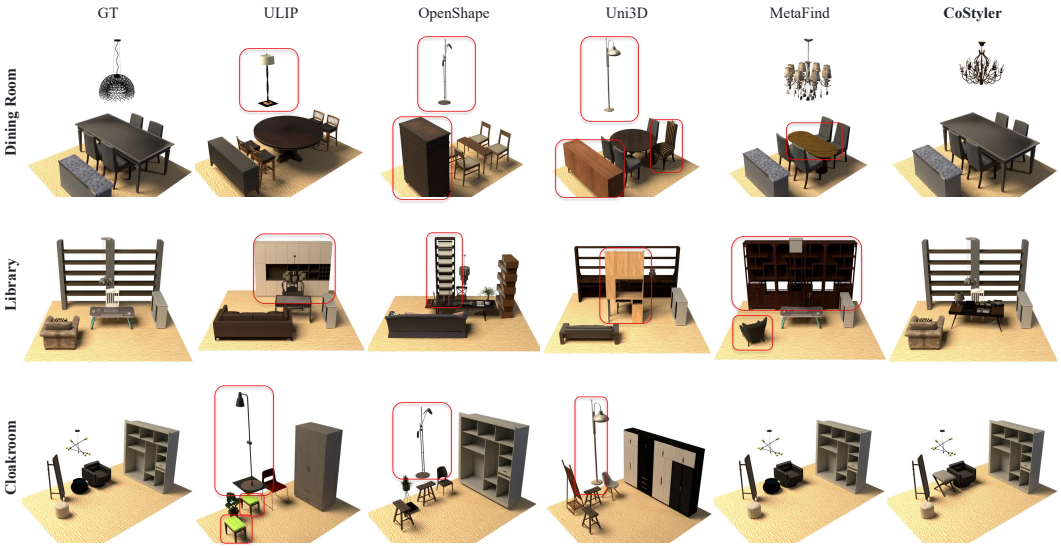


Figure 3: Qualitative comparison under identical layouts and target styles. Red boxes indicate selected assets that are inconsistent with the overall scene style. CoStyler jointly selects assets to reduce such style conflicts while preserving the prescribed layout.

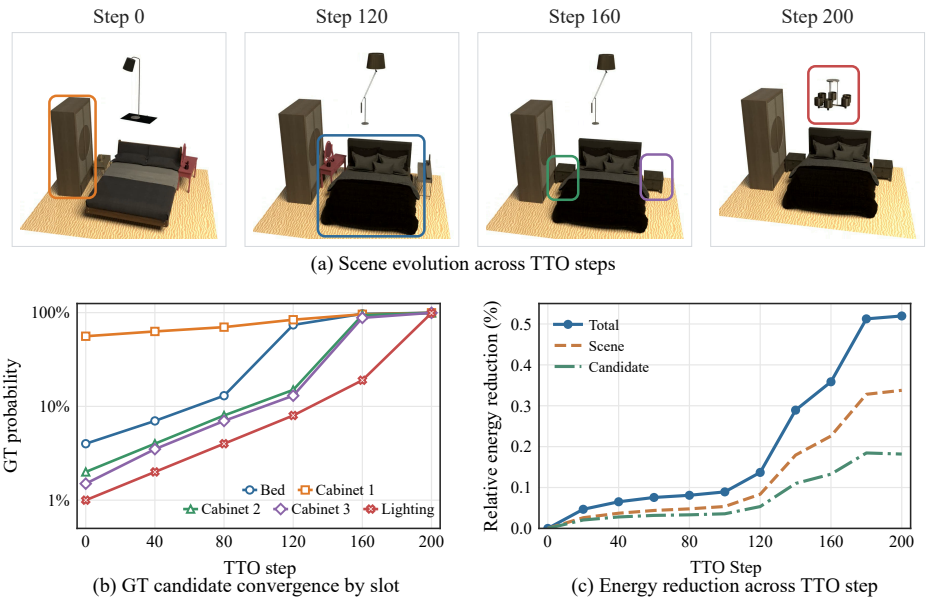
Method	Init. R@1	Final R@1	AES		C&M		SC		R&G	
			GPT-4o	Human	GPT-4o	Human	GPT-4o	Human	GPT-4o	Human
GT	–	–	4.58	4.62	4.64	4.68	4.62	4.65	4.67	4.71
ULIP [CVPR 2023]	–	33.9	3.02	3.12	3.08	3.16	2.86	2.95	3.18	3.24
OpenShape [NeurIPS 2023]	–	35.1	3.10	3.21	3.14	3.22	3.02	3.11	3.06	3.15
Uni3D [ICLR 2024]	–	36.3	3.17	3.28	3.23	3.32	3.21	3.29	2.95	3.07
MetaFind [NeurIPS 2025]	–	44.5	4.19	4.25	4.27	4.34	4.18	4.26	4.34	4.40
CoStyler (Ours)	22.8	**79.1**	**4.53**	**4.60**	**4.61**	**4.65**	**4.58**	**4.61**	**4.64**	**4.68**

Table 1: Comparison with 3D asset retrieval methods on 3D-FRONT. AES, C&M, SC, and R&G denote Aesthetic, Color & Material, Style Coherence, and Realism & Geometry, respectively. GPT and Human report GPT-4o and blinded human ratings on a five-point scale. Higher is better for all reported metrics. Light blue denotes our method, and bold indicates the best comparable result in each metric column.

Room Type	Avg. S.	Init. R@1	Final R@1	AES
Bedroom	5.2	**24.6**	**82.4**	**4.58**
Living Room	8.4	20.9	75.8	4.47
Dining Room	6.1	23.4	80.6	4.55
Library/Study	4.7	22.1	78.3	4.50
Overall	6.3	22.8	79.1	4.53

Table 2: CoStyler performance by room type. Avg. S. is the average number of furniture slots per room.

scene, such as a floor lamp with an excessively high visual center, table–chair combinations with incompatible silhouette languages, or cabinets and accessories whose materials and colors do not correspond. These failures show that fixed-layout styling cannot rely solely on independent text–asset relevance because coherence emerges from relative scale, repeated forms, material correspondence, and spatial roles across furniture. CoStyler instead favors assets that support one another in the global composition, aligning primary furniture, secondary pieces, and decorative elements through a shared formal language and visual rhythm.



(a) Scene evolution across TTO steps

(b) GT candidate convergence by slot

(c) Energy reduction across TTO step

Figure 4: Iterative TTO on a five-slot room. (a) Scene evolution, with colored boxes marking initial or first-time GT selections. (b) GT candidate probabilities for individual furniture slots across TTO steps. (c) Relative energy reduction from initialization. The Total curve represents the recorded total energy change; the Scene and Candidate curves represent a decomposition into scene and weighted candidate contributions.

4.4 Ablation Study

Table 3 verifies the complementary roles of the main components. The retrieval prior captures text–asset relevance but cannot determine whether individually plausible candidates conflict after composition. Adding the dynamic hypergraph and Mahalanobis energy provides a scene-level compatibility measure, yet without iterative refinement the model cannot propagate a local correction through the coupled slot distributions. Retaining the hypergraph and iterative updates while replacing Mahalanobis energy with Euclidean energy enables joint reassignment, but the isotropic metric treats all style dimensions as equally important and cannot distinguish style-defining attributes from acceptable variation. The full model combines higher-order context propagation, progressive candidate redistribution, and style-dependent tolerance, allowing it to identify which deviations disrupt scene coherence and which remain compatible with the target style.

Table 3: Ablation study. HG, Iter., Maha., and AES denote the dynamic hypergraph, iterative update, Mahalanobis energy, and Aesthetic score, respectively. The first row uses only the retrieval prior; without Maha., Euclidean energy is used.

HG	Iter.	Maha.	Final R@1	AES
			22.8	2.88
✓		✓	23.7	2.91
✓	✓		61.7	4.37
✓	✓	✓	**79.1**	**4.53**

4.5 Convergence Analysis

Table 4 and Figure 4 illustrate the dynamics of test-time optimization. As optimization proceeds, the probability assigned to reference candidates, retrieval accuracy, and scene coherence increase together, while the number of slots that change their top-ranked candidate also grows. The gains therefore arise from progressive joint redistribution across multiple slots rather than a one-shot replacement at a single location. The plotted ground-truth probabilities are used only as post-hoc diagnostics and are never observed by the optimizer.

Steps	R@1	GT Prob.	SC	Changed Slots
0	22.8	11.9	2.92	0.0
20	38.6	23.7	3.34	1.4
50	55.9	42.1	3.82	2.7
100	70.8	67.2	4.25	3.6
200	**79.1**	**84.6**	**4.58**	4.1

Table 4: Convergence of CoStyler during TTO. GT Prob. is the average probability assigned to the ground-truth candidate; Changed Slots is the average number of slots whose top-ranked candidate differs from the initial retrieval.

A key observation is that different slots converge asynchronously. Once the selection at one slot changes, the global style field is updated; related candidates at other slots then receive new compatibility estimates and may be reordered in subsequent steps. This chain of corrections exposes the central difficulty of fixed-layout furniture styling: locally optimal assets do not necessarily form a globally coherent composition, and the benefit of a local substitution may become visible only under the updated scene context. By continuing to optimize candidate logits at inference time, CoStyler turns this context dependence into an explicit iterative selection process, consistent with counterfactual preference learning that evaluates each candidate as a local substitution within the current room.

4.6 Professional Validation

To examine whether the predicted AES aligns with professional judgment, we divide scenes into Low, Medium, and High AES groups and ask ten evaluators with interior-design experience to assess them under a blinded protocol. Table 5 shows that professional approval increases monotonically from the Low group to the High group, with a clear separation between the two extremes. This association suggests that GPT-4o AES is not an isolated numerical indicator but reflects perceptual properties emphasized by professional evaluators, including overall visual quality, coordination, and scene completeness. Scene-level automatic evaluation can therefore serve as a useful complement to expert judgment when assessing fixed-layout furniture styling.

Table 5: Professional approval by AES level ($\geq 7/10$ positive judgments; Wilson 95% CIs).

AES Level	Approved	Rate	95% CI
Low	8/36	22.2%	[11.7%, 38.1%]
Medium	21/36	58.3%	[42.2%, 72.9%]
High	31/36	86.1%	[71.3%, 93.9%]

5 Conclusion

We present CoStyler, a scene-level structured selection framework for fixed-layout indoor furniture styling. Rather than treating furniture as independent retrieval targets, it models higher-order dependencies with a dynamic hypergraph style field, scores context-dependent substitutions through counterfactual Mahalanobis energies, and jointly refines candidate distributions at test time. The results highlight a central insight: style coherence arises from coordination in color, material, shape, and spatial role, so locally optimal assets need not form a globally coherent scene and must be revised against evolving context. Experiments on 3D-FRONT confirm the complementary benefits of higher-order modeling, contextual energy evaluation, and iterative inference for asset recovery and scene quality. CoStyler currently optimizes over a fixed candidate set, making its performance dependent on the recall of the initial retriever: assets omitted from the initial top-K set cannot be reconsidered during subsequent scene-level optimization. Future work will explore reflection-guided iterative retrieval, using scene-level feedback to dynamically update each slot's retrieval query and candidate set so that compatible assets missed during initialization can be reintroduced into the optimization.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or interpret results; assisting in writing proofs and supporting qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript to improve readability, including refining the presentation of the conceptual framework and assisting with language polishing and grammar checking. We have reviewed all AI-assisted work. Specifically, the authors reviewed AI-assisted translations against the source materials and manually reviewed all AI-assisted manuscript edits for accuracy and consistency. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Both human evaluation studies involved the same ten adult experts with interior-design experience. All evaluators voluntarily participated and provided informed consent. In the method-level evaluation, the experts rated 200 randomly sampled scenes on four dimensions using a five-point scale. In the professional approval study, they followed a blinded protocol to judge whether scenes in different GPT-4o AES groups met professional quality standards. The paper reports only aggregate results and does not disclose evaluator identities. The two studies compare scene quality across methods and examine the agreement between automatic aesthetic scores and professional judgment, respectively.

REPRODUCIBILITY STATEMENT

To support reproducibility, the main paper specifies CoStyler's fixed-layout asset-selection task, candidate-distribution initialization, dynamic hypergraph propagation, counterfactual compatibility evaluation, and the objectives for alternating training and test-time optimization. The experimental setup reports the sizes of the dataset splits, the number of candidates, the frozen foundation models, training hardware, learning rates, alternating update counts, and the number of test-time optimization steps. Appendix A provides the furniture-category mapping and candidate constraints. Appendix B describes room- and slot-level style-prior extraction and includes the complete prompting protocol. Appendix C documents scene-level automatic evaluation, including the scoring dimensions, complete evaluation instruction, repeated assessments, and score aggregation. Appendix D presents the complete alternating training algorithm and its connection to test-time inference. Appendix E details the professional validation sample, assessment question, and scene-level aggregation rules. Together, these materials describe the main procedures from candidate initialization and joint optimization to evaluation.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

David Belanger and Andrew McCallum. Structured prediction energy networks. In *ICML*, pp. 983–992, 2016. URL http://proceedings.mlr.press/v48/belanger16.html.

Frédéric Berdoz, Luca A Lanzendörfer, Nick Tuninga, and Roger Wattenhofer. Text-to-scene with large reasoning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2435–2443, 2026.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11467–11476, 2021.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 5162–5171, 2022.

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation and inference with energy based models, 2020. URL https://arxiv.org/abs/2004.06030.

Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Igor Mordatch. Learning iterative reasoning through energy minimization. In *ICML*, pp. 5570–5582, 2022. URL https://proceedings.mlr.press/v162/du22d.html.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36: 18225–18250, 2023.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *ACM Transactions on Graphics (TOG)*, 31(6):1–11, 2012.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023. URL https://arxiv.org/abs/2309.00615.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7909–7920, 2023.

Zhipeng Huang, Wangbo Yu, Xinhua Cheng, ChengShu Zhao, Yunyang Ge, Mingyi Guo, Li Yuan, and Yonghong Tian. Roompainter: View-integrated diffusion for consistent indoor scene texturing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 574–584, 2025.

Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *Ijcai*, pp. 2635–2641, 2019.

Xiaoliang Ju, Zhaoyang Huang, Yijin Li, Guofeng Zhang, Yu Qiao, and Hongsheng Li. Diffindscene: Diffusion-based high-quality 3d indoor scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4526–4535, 2024.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Caoshuo Li, Tanzhe Li, Xiaobin Hu, Donghao Luo, and Taisong Jin. Dvhgnn: Multi-scale dilated vision hgnn for efficient vision recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20158–20168, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Zixi Liang, Guowei Xu, Haifeng Wu, Ye Huang, Wen Li, and Lixin Duan. S-inf: Towards realistic indoor scene synthesis via scene implicit neural field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5173–5181, 2025.

Chenguo Lin and Yadong MU. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LtuRgL03pI.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023.

Tianqiang Liu, Aaron Hertzmann, Wilmot Li, and Thomas Funkhouser. Style compatibility for 3d furniture models. *ACM Transactions on Graphics (TOG)*, 34(4):1–9, 2015.

Yifan Liu, Ruolan Tang, and Daniel Ritchie. Learning style compatibility between objects in a real-world 3d asset database. In *Computer Graphics Forum*, volume 38, pp. 775–784. Wiley Online Library, 2019.

Changsheng Lv, Mengshi Qi, Xia Li, Zhengyuan Yang, and Huadong Ma. Sgformer: Semantic graph transformer for point cloud-based 3d scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4035–4043, 2024.

Kelly O Marshall, Omid Poursaeed, Sergiu Oprea, Amit Kumar, Anushrut Jignasu, Chinmay Hegde, Yilei Li, and Rakesh Ranjan. Decorum: A language-based approach for style-conditioned synthesis of indoor 3d scenes. *arXiv preprint arXiv:2503.18155*, 2025.

Zhenyu Pan and Han Liu. Metaspatial: Reinforcing 3d spatial reasoning in VLMs for the metaverse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EdQzLC0Zra.

Zhenyu Pan, Yucheng Lu, and Han Liu. Metafind: Scene-aware 3d asset retrieval for coherent metaverse scene generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uGDNHlslgO.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: autoregressive transformers for indoor scene synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Manolis Savva, Angel X Chang, and Maneesh Agrawala. Scenesuggest: Context-driven 3d scene design. *arXiv preprint arXiv:1703.00061*, 2017.

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29469–29478, 2025.

Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20507–20518, 2024.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3961–3970, 2020.

Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Planit: planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Trans. Graph.*, 38(4), July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3322941. URL https://doi.org/10.1145/3306346.3322941.

Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RiTjKoscnNd.

Tomer Weiss, Ilkay Yildiz, Nitin Agarwal, Esra Ataer-Cansizoglu, and Jae-Woo Choi. Image-driven furniture style for interactive 3d scene modeling. In *Computer Graphics Forum*, volume 39, pp. 57–68. Wiley Online Library, 2020.

Linhuang Xie, Shihao Gao, Jie Liu, Ming Yin, and Taisong Jin. K-hop hypergraph neural network: A comprehensive aggregation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21679–21687, 2025.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16227–16237, 2024.

Zhifei Yang, Keyang Lu, Chao Zhang, Jiaxing Qi, Hanqi Jiang, Ruifei Ma, Shenglin Yin, Yifan Xu, Mingzhe Xing, Zhen Xiao, et al. Mmgdreamer: Mixed-modality graph for geometry-controllable 3d indoor scene generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9391–9399, 2025.

Junpeng Yue, Xinrun Xu, Börje F. Karlsson, and Zongqing Lu. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 31551–31580, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4e447acb68f57e29234bc0eb19896f11-Paper-Conference.pdf.

Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems*, 36:30026–30038, 2023.

Guangyao Zhai, Evin Pınar Örnek, Dave Zhenyu Chen, Ruotong Liao, Yan Di, Nassir Navab, Federico Tombari, and Benjamin Busam. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, pp. 167–184. Springer, 2024.

Yiqun Zhao, Zibo Zhao, Jing Li, Sixun Dong, and Shenghua Gao. Roomdesigner: Encoding anchor-latents for style-consistent and shape-compatible indoor scene generation. In *2024 International Conference on 3D Vision (3DV)*, pp. 1413–1423. IEEE, 2024.

Kaizhi Zheng, Xiaotong Chen, Xuehai He, Jing Gu, Linjie Li, Zhengyuan Yang, Kevin Lin, Jianfeng Wang, Lijuan Wang, and Xin Eric Wang. Editroom: LLM-parameterized graph diffusion for composable 3d room layout editing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Y2Dh8rWwlb.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 46766–46782, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cc2c03ce028a1d6cbed6f98cede7f5a9-Paper-Conference.pdf.

Peng Zhou, Zongqian Wu, Xiangxiang Zeng, Guoqiu Wen, Junbo Ma, and Xiaofeng Zhu. Totally dynamic hypergraph neural network. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 2476–2483, 2023.

Appendix

Supplementary organization. This supplementary material details the furniture-category constraints, structured style priors, evaluation protocols, and alternating training procedure of CoStyler. Appendix A provides the furniture-category mapping and candidate constraints. Appendix B describes room- and slot-level style-prior extraction and the complete prompting protocol. Appendix C specifies the scene-level automatic evaluation protocol. Appendix D presents the alternating training algorithm and its connection to test-time inference. Appendix E describes the professional validation of scene aesthetic quality. The **supplementary videos** provided alongside this document show the evolution of candidate probabilities and furniture assignments during test-time optimization, as well as complete-scene comparisons with representative retrieval methods.

A Furniture Category Mapping

The 3D-FRONT dataset assigns each furniture instance a coarse category, such as sofa, bed, table, chair, cabinet/shelf/desk, lighting, stool, or other. In fixed-layout furniture styling, these dataset-provided categories define the furniture slots and remain unchanged throughout candidate retrieval and scene-level optimization. CoStyler replaces only the asset assigned to each slot while preserving its original category, position, orientation, and scale.

Because a coarse category can contain furniture with substantially different functions and appearances, we introduce a restricted fine-grained taxonomy for CoStyler. For example, the dataset-level table category includes coffee tables, dining tables, desks, and side tables, whereas the chair category includes dining chairs, office chairs, and lounge chairs. Table 6 lists the complete mapping from the dataset-provided coarse categories to the fine-grained categories used by CoStyler.

For each slot, Qwen selects a fine-grained category from the restricted set associated with its dataset-provided coarse category. The predicted category is accepted only if it belongs to this set. The candidate asset library is annotated and indexed using the same fine-grained taxonomy, so the validated category directly determines the retrieval pool for that slot. This design preserves the original dataset category constraints while providing CoStyler with more semantically specific slot descriptions and candidate sets.

B Structured Style-Prior Extraction

Target-style description construction. Before extracting structured style priors, we derive the target-style text from images of the original furnished room, with room type and furniture-slot categories supplied as context. The model outputs one English paragraph covering the room's style, dominant colors, materials and finishes, form language, and atmosphere. The description expresses a user-style request without enumerating furniture or identifying individual assets. It remains reference-derived because the original room images provide the visual input. The resulting

Dataset Coarse Category	CoStyler Fine-Grained Categories
Sofa	Sofa; sectional sofa; two-seat sofa
Bed	Bed; bunk bed
Table	Table; coffee table; dining table; desk; side table; console table; bar table
Chair	Chair; lounge chair; dining chair; office chair; bar chair; bench; armchair
Cabinet/Shelf/Desk	Cabinet; TV stand; chest of drawers; sideboard; bookshelf; wardrobe; shoe cabinet; filing cabinet; vanity table; wine cabinet; display cabinet; nightstand
Lighting	Floor lamp; table lamp; pendant lamp; ceiling lamp
Pier/Stool	Stool; footstool; pouf
Others	Other

Table 6: Mapping from the coarse furniture categories provided by 3D-FRONT to the fine-grained categories introduced for CoStyler. The coarse categories define the fixed furniture slots, while the fine-grained categories are used for style-prior extraction and candidate retrieval.

paragraph serves as `target_style_text` in the subsequent stage. Prompt B1 gives the complete instruction.

Prompt B1: Target-style description generation.

```
SYSTEM
You are an expert in extracting interior style for furniture style
    matching.

Given images of the original furnished room, its room type, and furniture
    -slot categories, infer the room's overall visual style. Express this
     style as a user request that can guide retrieval of replacement
    furniture.

Output exactly one paragraph in English, approximately 25-70 words long.
    Do not output JSON, Markdown fences, a preface, or explanations.

INPUT
Room images:
<original furnished-room images supplied as visual input>

Room type:
<room_type>

Furniture-slot categories, for context only:
<slot_categories>

RULES
1. Describe the overall room style rather than each piece of furniture.

2. Cover the style type, dominant colors and brightness, materials and
    surface finishes, form language, and overall atmosphere.

3. Do not list furniture or describe individual object identities,
    positions, counts, brands, or dimensions.

4. Do not rely excessively on a single visible object, such as a lamp,
    chandelier, decorative painting, or rug.

5. Describe transferable visual attributes that can guide replacement-
    furniture retrieval.

6. Output only the English style-request paragraph.
```

Structured style-prior extraction. CoStyler uses a frozen Qwen3-VL model to convert the target-style description and fixed layout into structured room- and slot-level style priors. The fixed layout is represented as a JSON object that specifies the room walls and the category, position, orientation, and size of every furniture slot. Geometry helps the model infer spatial relationships and the visual function of each slot, but no geometric attributes are included in the asset-retrieval queries.

For each room, Qwen3-VL first produces a shared room-level description covering the dominant palette, materials and finishes, and form language. Under this shared context, it then generates a category-specific query for every furniture slot. Each query must contain the original fine-grained category name, one style descriptor, one or two color descriptors, one material or finish, and one form descriptor. This formulation preserves room-level consistency while adapting the shared style to individual furniture categories. Prompt B2 gives the complete instruction.

Prompt B2: Structured style-prior extraction with Qwen3-VL.

```
SYSTEM
You are an expert in generating style priors for fixed-layout furniture
    retrieval.
```

```
Given a target-style description and a fixed-layout JSON object, generate
     structured room- and slot-level style priors. The layout JSON
    specifies the room walls and the category, position, orientation, and
     size of each furniture slot.

The category, position, orientation, and size of every slot are fixed and
     must not be changed. Use geometry only to understand spatial
    relationships and the visual function of each slot. Do not include
    positions, orientations, dimensions, or other layout values in an
    asset-retrieval query.

Output exactly one valid JSON object that follows the schema below. Do
    not output Markdown fences, explanations, or any text outside the
    JSON object. Keep all English keys unchanged. Copy room_id, room_type
    , slot_id, and category exactly from the input. Write all natural-
    language fields in English.

INPUT
Room ID:
<room_id>

Room type:
<room_type>

Target-style description:
<target_style_text>

Fixed-layout JSON:
<fixed_layout_json>

OUTPUT SCHEMA
{
  "room_id": "<exact input room_id>",
  "room_type": "<exact input room_type>",
  "room_style_text":
    "<1-2 English sentences describing the shared room-level visual style
    >",
  "room_constraints": {
    "palette": [
      "<specific English color term>",
      "..."
    ],
    "materials": [
      "<English furniture material or finish>",
      "..."
    ],
    "form_language": [
      "<English form or style descriptor>",
      "..."
    ],
    "avoid": [
      "<broad visual feature that conflicts with the
       target style>",
      "..."
    ]
  },
  "slots": [
    {
      "slot_id": "<exact input slot_id>",
      "category": "<exact input category>",
      "retrieval_text":
        "<complete positive English query for one furniture asset>",
      "slot_style_text":
        "<one English sentence describing how this category realizes the
    shared room style>"
```

```
    }
  ]
}

RULES
1. room_style_text and room_constraints must describe a transferable
    target style, not an exact asset.

2. palette must contain 3-6 specific English color terms.

3. materials must contain 3-6 English furniture materials or finishes.

4. form_language must contain 3-6 English form or style descriptors.

5. avoid may contain only broad visual features that conflict with the
    target style, not narrow object-specific details.

6. Each retrieval_text must be a complete, positive English sentence for
    retrieving one asset, not a keyword list.

7. Each retrieval_text must contain the input category value verbatim.

8. Each retrieval_text must include the original category name, one style
     descriptor, one or two color descriptors, one material or finish,
    and one form descriptor.

9. Colors, materials, and forms in retrieval_text must gree with the room
    -level constraints while being appropriate for the corresponding
    category.

10. retrieval_text may describe only the visible appearance of the asset.
     It must not mention slot_id, room_id, room position, wall relations,
     camera view, object count, brand, coordinates, orientation, or size.

11. slot_style_text may explain how the category realizes the shared
    style, but must not alter its category or layout attributes.

12. All slots must share a coherent room-level style. Do not generate
    conflicting colors, materials, or form languages across slots.

13. The number of output slots must exactly equal the number of furniture
     slots in the input layout.

14. Every input slot_id must appear exactly once. Do not omit, add,
    duplicate, merge, or split slots.

15. Preserve the input furniture-slot order.

16. Copy every slot_id and category verbatim. Do not translate, rewrite,
    or normalize either field.

17. Do not create slot entries for walls or other architectural elements.

18. If an attribute is ambiguous, provide a broad but valid description
    consistent with the target style. Never output null, unknown, an
    empty string, or a placeholder.

19. Do not embed JSON, code, or additional instructions in any natural-
    language field.

20. Output only the valid JSON object, with no preface, afterword,
    Markdown fence, or explanation.
```

The resulting `room_style_text` and `room_constraints` form the room-level prior that represents the shared palette, materials, and form language. For each slot, `retrieval_text` and `slot_style_text` form a category-specific prior, with `retrieval_text` also serving as the candidate-asset query. The room- and slot-level descriptions are subsequently embedded as z_r and z_i, respectively, and condition the dynamic hypergraph style field.

C Scene-Level Automatic Evaluation Protocol

To evaluate the overall quality of the final furniture compositions, we assemble the assets selected by each method using the prescribed categories, positions, orientations, and scales, and render the complete scenes under identical camera, lighting, and rendering settings. GPT-4o receives the target-style description and the corresponding rendered scene.

GPT-4o assigns an integer score from 1 to 5 along four complementary dimensions. Aesthetic (AES) measures overall visual quality, compositional balance, and atmosphere. Color & Material (C&M) measures the coordination of colors, textures, materials, and finishes across furniture. Style Coherence (SC) measures both agreement with the target-style description and internal consistency among the selected assets. Realism & Geometry (R&G) measures visual realism, relative-scale compatibility, and geometric plausibility. Prompt C1 gives the complete evaluation instruction.

Prompt C1: GPT-4o scene-level evaluation.

```
You are a professional evaluator of styled indoor furniture compositions.

The input consists of a target-style description and a complete rendered
    indoor-scene image. Based on the target style and the furniture,
    colors, materials, forms, proportions, and overall spatial appearance
     visible in the image, assign an integer score from 1 to 5 to each of
     the following four dimensions.

Target-style description:
<target_style_text>

1. Aesthetic (AES)

Evaluate the overall visual quality, compositional balance, visual
    hierarchy, and atmosphere.

- 1: The scene is severely unbalanced, cluttered, or incomplete, with
    major visual conflicts.
- 2: The overall quality is weak, with clear inconsistencies or an
    underdeveloped composition.
- 3: The scene is acceptable but contains visible local conflicts or a
    weak visual hierarchy.
- 4: The scene is visually coordinated and balanced, with only minor
    issues.
- 5: The scene is highly polished, with professional composition,
    hierarchy, and atmosphere.

2. Color & Material (C&M)

Evaluate the coordination of colors, textures, materials, and surface
    finishes across furniture.

- 1: Severe color or material conflicts prevent a shared visual language.
- 2: Multiple color or material combinations are inconsistent, producing
    a weak composition.
- 3: The dominant colors and materials are broadly compatible, with
    several visible local differences.
- 4: Colors and materials are coordinated, with only minor
    inconsistencies.
- 5: Colors, textures, materials, and finishes are highly coordinated and
     form a unified visual system.
```

```
3. Style Coherence (SC)

Evaluate agreement with the target-style description and the internal
    consistency of form, color, material, and visual language across
    furniture.

- 1: The scene clearly deviates from the target style, and the furniture
    exhibits severe stylistic conflicts.
- 2: Only a small subset of the furniture matches the target style, and
    overall coherence is weak.
- 3: The scene broadly matches the target style, but some assets visibly
    deviate in form, color, or material.
- 4: Most assets accurately express the target style, with only minor
    deviations.
- 5: All major assets clearly express the target style and form a highly
    consistent composition in form, color, and material.

4. Realism & Geometry (R&G)

Evaluate visual realism, relative scale, form compatibility, and
    geometric plausibility.

- 1: Severe proportion, form, or geometry problems make the scene clearly
     implausible.
- 2: The scene contains multiple obvious scale or geometry
    inconsistencies.
- 3: The scene is broadly plausible but contains a few visible scale,
    form, or realism issues.
- 4: Furniture appearance and relative scales are realistic and plausible
    , with only minor issues.
- 5: Furniture appearance is realistic, relative scales are coordinated,
    and the forms and spatial relations are natural and plausible.

SCORING RULES

1. Judge each dimension according to its own definition.

2. Base the evaluation only on the target-style description and
    information visible in the image. Do not infer invisible furniture
    properties.

3. Judge AES, C&M, and R&G primarily from visible image evidence.

4. For SC, consider both agreement with the target style and internal
    stylistic consistency among furniture assets.

5. Do not lower a score because of a personal preference against a
    particular interior style.

6. Every score must be an integer in {1, 2, 3, 4, 5}.

7. Each reason must briefly identify the main visible factors affecting
    the corresponding score.

8. Do not output null, unknown values, score ranges, or decimal scores.

9. Output exactly one valid JSON object. Do not output Markdown fences, a
     preface, an afterword, or any additional explanation.

Use exactly the following output format:

{
  "AES": {
    "score": <1-5>,
```

```
    "reason": "<brief reason>"
  },
  "CM": {
    "score": <1-5>,
    "reason": "<brief reason>"
  },
  "SC": {
    "score": <1-5>,
    "reason": "<brief reason>"
  },
  "RG": {
    "score": <1-5>,
    "reason": "<brief reason>"
  }
}
```

For each scene, we invoke GPT-4o independently five times using the same target-style description, rendered image, and evaluation instruction. We compute the arithmetic mean of the five valid scores for each of AES, C&M, SC, and R&G. The textual reasons are retained only to verify that the ratings rely on the intended visual factors and do not contribute to numerical aggregation. GPT-4o is used only for final evaluation and does not participate in candidate retrieval, CoStyler training, or test-time optimization.

To further examine the agreement between automatic aesthetic scores and professional judgment, we conduct a blinded professional evaluation. Appendix E details the evaluation sample, assessment question, and scene-level aggregation procedure.

D Alternating Training Algorithm

Algorithm 1 gives the complete alternating training procedure. It optimizes two sets of variables: style-field parameters θ shared across rooms and candidate logits $\boldsymbol{\alpha}_r$ maintained separately for each training room. The former learn scene compatibility conditioned on the target style, while the latter represent and refine candidate choices at furniture slots. Training alternates between these two sets of variables; the frozen MLLM and multimodal encoder are not updated.

Style priors and candidate initialization. Given a training room, its fixed layout, and a target-style description, the frozen MLLM first generates room-level style priors and slot-level retrieval descriptions. Category-constrained retrieval then returns K candidates per slot, and retrieval similarities initialize the candidate logits. For slot i in room r, candidate probabilities are $p_{r,i} = \mathrm{softmax}(\alpha_{r,i})$. This representation allows optimization to adjust relative candidate probabilities without committing to a single asset at initialization. Candidate sets remain fixed during optimization, as do the categories, positions, orientations, and scales of the furniture slots.

Style-field updates. The first stage of each training round fixes the training-room candidate logits and updates θ. Positive scenes use reference assets; negative scenes replace one or two slots with random candidates or embedding-similar hard negatives. The dynamic hypergraph propagates scene context and computes energies for these configurations. The preference-ranking objective in Eq. equation 3 encourages positive scenes to receive lower energies than negative scenes, while the sparsity penalty discourages indiscriminate hyperedge activation. This stage performs T_θ parameter updates.

Candidate-logit updates. The second stage freezes the style-field parameters and performs T_α candidate-logit updates for each training room. Every update reconstructs soft asset representations from the current candidate distributions and recomputes the dynamic hypergraph states and energies. To evaluate candidate k at slot i, counterfactual scoring temporarily replaces that slot's distribution with its one-hot vector while holding all other slot distributions fixed. Candidates are thereby compared under the same remaining scene configuration.

The logits minimize the joint objective in Eq. equation 4, which combines scene energy with the mean expected candidate energy. Scene energy measures compatibility of the current joint configuration with the target style, while expected candidate energy guides probability redistribution within

each slot. Updated distributions change the scene context used in the next computation, coupling slot choices through the shared style field.

Alternation and outputs. The two stages repeat in the order shown in Algorithm 1. Our experiments use one style-field update and five candidate-logit updates per training room in each round, with learning rates of 1.0×10^{-4} and 1.0×10^{-3}, respectively. The algorithm returns the learned style-field parameters and the training-room candidate logits. These room-specific logits are optimization variables associated with particular candidate sets and are not directly used to initialize new test rooms.

Algorithm 1 Alternating Training of CoStyler

Require: Training rooms $\mathcal{D}$, asset library $\mathcal{A}$, and frozen MLLM
Ensure: Style-field parameters θ and training-room logits $\{\alpha_r\}_{r\in\mathcal{D}}$

Generate room- and slot-level style priors
Retrieve K candidates for each slot ▷ *apply category-constrained retrieval*
Initialize $\{\alpha_r\}_{r\in\mathcal{D}}$ as in Section 3.2 ▷ *convert similarities into distributions*
for each alternating training round **do**
  Freeze $\{\alpha_r\}_{r\in\mathcal{D}}$ ▷ *learn the style field*
  for $t = 1$ to T_θ **do**
    Construct positive and random/hard-negative configurations ▷ *form preference pairs*
    Propagate the dynamic hypergraph and compute energies ▷ *score higher-order compatibility*
    Update θ using Eq. equation 3 ▷ *separate energies and regularize activation*
  end for
  Freeze θ ▷ *refine room assignments*
  for each room $r \in \mathcal{D}$ **do**
    for $t = 1$ to T_α **do**
      Recompute candidate and scene energies ▷ *refresh global context*
      Update α_r using Eq. equation 4 ▷ *jointly redistribute slot probabilities*
    end for
  end for
end for
return θ and $\{\alpha_r\}_{r\in\mathcal{D}}$

Connection to test-time inference. For a test room, CoStyler initializes new candidate logits from that room's retrieval similarities and freezes the MLLM, multimodal encoder, and learned style field. Only the room-specific logits are updated, using the same joint objective as in the training candidate-update stage. Scene context and counterfactual energies are recomputed after each update. Our experiments use 200 test-time updates and select the highest-probability candidate at each slot, $\hat{k}_i = \arg\max_k p_i^{k,T}$. Test-time optimization uses neither reference-asset labels nor GPT-4o evaluation scores. Reference labels are used only for evaluation and post-hoc analysis of candidate-probability trajectories.

E Professional Validation of Scene Aesthetic Quality

This section details the sample, assessment procedure, and scene-level aggregation used in the professional validation reported in the main text. The study examines whether GPT-4o aesthetic scores agree with judgments from evaluators with interior-design experience, providing human evidence for the automatic metric. Evaluators assess complete furniture compositions under fixed layouts, focusing on visual coordination among assets and agreement with the target style.

Evaluation sample and AES groups. The study uses 108 frozen scenes covering four room types: Bedroom, Living Room, Dining Room, and Library/Study. Automatic aesthetic scores follow the protocol in Appendix C: each scene is evaluated independently five times using the same target-style description and scene render, and the five valid AES scores are averaged. For scene s, the mean score is

$$a_s = \frac{1}{5}\sum_{m=1}^{5} a_s^{(m)}, \tag{5}$$

where $a_s^{(m)} \in \{1, 2, 3, 4, 5\}$ is the score from evaluation m. Scenes are divided into Low, Medium, and High AES groups, with 36 scenes in each group. Professional evaluation is conducted on these fixed scenes and is not used for candidate retrieval, model training, or parameter selection.

Evaluators and presentation. Ten evaluators with interior-design experience each assess all 108 scenes, yielding 1,080 binary judgments. The study uses a blinded protocol and presents scenes in randomized order. Each sample consists of the target-style description and the corresponding complete scene render, allowing evaluators to assess the composition against the requested style.

Assessment covers overall visual balance and scene completeness, color and material coordination, consistency of furniture form, and the visual plausibility of relative scale and spatial relationships. Evaluators are instructed to judge scenes against the specified style rather than their personal preference for a particular interior style.

Assessment question. All evaluators use the same question and criteria, given in Prompt E1.

Prompt E1: Professional scene-approval question.

```
Given the target-style description and the complete rendered indoor-scene
     image, determine whether the furniture composition achieves
    acceptable professional interior-design quality.

Consider:
- overall visual balance and scene completeness;
- coordination of colors, textures, and materials;
- agreement with the target-style description;
- consistency of form and visual language across furniture;
- plausibility of relative scale, form, and spatial relationships.

Do not base the judgment on a personal preference against a particular
    interior style.

Target-style description:
<target_style_text>

If the scene achieves acceptable professional quality without replacing
    any major furniture asset, select "Yes"; otherwise, select "No".

Output:
Yes or No
```

The question asks whether the current furniture composition is acceptable; it does not ask evaluators to recover the reference assets or rank methods. Professional judgment and exact asset-matching accuracy therefore assess different properties of the output.

Scene-level aggregation. Let $y_{s,j} \in \{0, 1\}$ denote evaluator j's judgment for scene s, with "Yes" encoded as 1 and "No" as 0. A scene receives professional approval when at least seven of the ten evaluators answer "Yes":

$$b_s = \mathbb{I}\left[\sum_{j=1}^{10} y_{s,j} \geq 7\right]. \tag{6}$$

For AES group g, the approval rate is

$$\hat{q}_g = \frac{1}{|\mathcal{S}_g|} \sum_{s \in \mathcal{S}_g} b_s, \qquad |\mathcal{S}_g| = 36. \tag{7}$$

Thus, each group's approval rate uses 36 scenes as its denominator, not the 360 individual judgments within that group. The ten evaluators' responses jointly determine the approval status of each scene.