Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatial relationships guide semantic component extraction from Gaussian splatting of traditional Chinese timber architecture

Lei Wang
Beijing University of Civil Engineering and Architecture
Ming Huang
huangming@bucea.edu.cn
Beijing University of Civil Engineering and Architecture
Weifeng Ma
Yunnan Normal University
Dong Li
Beijing University of Civil Engineering and Architecture
Xingxing Xiao
Beijing University of Civil Engineering and Architecture

Article

Keywords: 3D Gaussian Splatting, Spatial relationship graph, Multimodal large language model, Component-level semantic object extraction, Traditional timber-frame architecture, Architectural heritage digitization

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11518276/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Spatial relationships guide semantic component extraction from Gaussian splatting of traditional Chinese timber architecture

Lei Wang[a, b], Ming Huang[b, c,*], Weifeng Ma[d], Dong Li[a, b], Xingxing Xiao[a, b]

a. School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

b. School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China

c. Engineering Research Center for Representative and Ancient Building Database of the Ministry of Education, Beijing 100044, China

d. Faculty of Geography, Yunnan Normal University, Kunming, Yunnan 650500, China

* Corresponding author.

E-mail: huangming@bucea.edu.cn

Abstract: Three-dimensional Gaussian splatting enables photorealistic documentation of architectural heritage, but extracting semantically accurate components remains challenging. We propose a framework for traditional Chinese timber architecture that combines spatial relationship-guided annotation with semantic–appearance decoupled Gaussian extraction. A spatial relationship graph guides a multimodal large language model in distinguishing visually similar components. Semantic assignments and filtering weights are then learned while keeping the reconstructed scene parameters fixed. On a Qing-dynasty timber-frame building with 22 component categories, the framework increases image-level mean intersection over union from 69.03% to 85.72% and extracted-object mean intersection over union from 28.01% to 69.24% compared with Heritage-3DGS. Average isolated-component peak signal-to-noise ratio increases from 17.05 to 24.30 dB, while full-scene rendering quality is preserved. The framework provides independently renderable component categories for semantic access to photorealistic architectural heritage records.

Keywords: 3D Gaussian Splatting; Spatial relationship graph; Multimodal large language model; Component-level semantic object extraction; Traditional timber-frame architecture; Architectural heritage digitization

Introduction

Traditional Chinese timber-frame architecture requires conservation at the component scale, where deterioration and repeated repair affect individual members and their connections [1,2]. Repair records, damage descriptions and material information therefore need to be associated with identifiable components and their structural relationships [3,4]. A component-level 3D representation that supports identification, retrieval and information linking is consequently important for digital conservation [5].

Point clouds and meshes provide geometric records for architectural survey, deformation assessment and spatial analysis [6,7], while Heritage Building Information Modelling (HBIM) organizes component attributes and construction relationships for conservation records and repair management [3,8]. Panoramic imagery and image-based models support visual interpretation and communication [9,10]. Combining detailed appearance, interactive visualization and component-level information remains demanding, particularly for painted decoration and intricate timber details that require substantial texture processing and manual modeling. 3D Gaussian Splatting (3DGS) offers a complementary representation through explicit Gaussian primitives, high-quality rendering and real-time viewing [11,12,13]. Its use in component-based heritage workflows requires these primitives to be organized by architectural semantics [14].

Existing semantic methods introduce information into 3D Gaussian scenes mainly through feature distillation or two-dimensional mask projection [15,16,17,18,19,20]. Feature distillation embeds vision-language features into 3D representations to support natural-language queries. LERF [21] embeds features from Contrastive Language–Image Pre-training (CLIP) into neural radiance fields, LangSplat [15] transfers this approach to 3DGS, and LEGaussians [22] improves the alignment between semantic features and Gaussian geometry. Mask projection uses two-dimensional segmentation and multi-view consistency to assign image masks to Gaussian primitives. Gaussian Grouping [18], SAGA [19], SAGD [23] and Flash Splat [20] advance this route through identity encoding, cross-scale affinity, boundary Gaussian decomposition and optimization-based label solving, respectively.

The application of these methods to traditional architecture depends on reliable recognition of specialized component categories. General vision-language models learn semantic associations from large-scale image-text corpora and support recognition of common objects [24,25]. Traditional architectural terms, however, encode construction-specific form, position, function and relationships that are less well represented in these corpora. Open-vocabulary methods based on CLIP consequently have difficulty consistently distinguishing fine-grained architectural component categories [26]. Mask projection also depends on the quality of its two-dimensional annotations:

category errors, boundary adhesion and occlusion errors can propagate into the Gaussian field through multi-view association. Complex textures, dense components and frequent occlusion make accurate source annotations particularly important in heritage scenes.

Multimodal large language models (MLLMs) offer a way to introduce specialized knowledge through natural-language descriptions and a small number of annotated image exemplars [27]. Engineers can supply component terminology and visual examples without retraining the recognition model or collecting a large annotated training set. Heritage-3DGS [28] uses MLLM-based two-dimensional component annotation to introduce historical architectural semantics into Gaussian scenes. Its application to church-type buildings motivates examining this approach in traditional Chinese timber-frame architecture, where the identification of visually similar members depends on structural context as well as specialized vocabulary.

Traditional Chinese timber-frame architecture is highly hierarchical and nested, with components organized along bays, depths and roof-frame levels. Many members have similar colors, textures and local geometric forms, while their names and functions depend on their positions and relationships with surrounding components. Component identity is therefore jointly defined by appearance and spatial topology [29]. Figure 1 illustrates how similar forms can correspond to different categories at different structural levels or under different adjacency relationships. This dependence on construction context complicates recognition from local images and textual exemplars [30].

Close-range acquisition can further obscure these relationships. Courtyards, corridors and dense column grids restrict viewpoints, so images that preserve component details may exclude the surrounding structural context. Local masks can consequently lack the adjacency, vertical and hierarchical relationships needed for identification. Benchmarks also report limitations in MLLM judgments of relative position, depth order and multi-object spatial relationships [31,32]. These observations motivate supplying explicit construction relationships during semantic review, so that category decisions can draw on structural context alongside appearance and component descriptions.

Structured relationships can be introduced into vision-language models through learned graph representations or textual prompts. The first route incorporates graphs into model design or external knowledge encoding and learns representations of relations among entities. MR-MKG [33] uses a relational graph attention network to encode multimodal knowledge graphs. Structure-CLIP [34] constructs semantic negative samples from scene graphs to strengthen structured representations, while LLaVA-SG [35] organizes image entities and their relationships as scene graphs for fine-grained semantic understanding. These approaches integrate relational information through specialized modules, additional parameters or training data.

The second route expresses structured relationships as textual priors supplied during inference. CCoT [36] guides a model to generate a scene graph and use it for subsequent reasoning. BDoG [37] organizes multi-agent debate within a shared graph, storing and updating evidence along graph branches to constrain its aggregation. Prompt-based injection provides a route for using relational knowledge without modifying model parameters, which is relevant to heritage tasks with limited annotated data. Knowledge graphs also organize cultural assets and their associated information, as demonstrated by ArCo [38]. These studies provide approaches to representing and using entity relationships; the present task requires connecting construction-specific relationships to the recognition of timber components and their extraction from Gaussian scenes.

Accurate image-level annotations must then be converted into clean, independently renderable Gaussian subsets. A category can occupy a plausible semantic region in the complete scene while retaining unrelated primitives that become visible when surrounding categories are removed. Coupling semantic supervision with appearance optimization can also allow annotation errors to disturb reconstructed appearance, while inconsistent multi-view masks can introduce background contamination and floating artifacts [39,40]. Component-level semantic object extraction therefore requires reliable category assignment, suppression of unrelated primitives and preservation of rendering fidelity. Here, each extracted subset contains the primitives assigned to one architectural category throughout the scene, providing independent access to that category within the photorealistic record.

To address these connected recognition and extraction problems, we propose a component-level semantic object extraction framework for 3DGS scenes of traditional Chinese timber-frame architecture. A Spatial Relationship Graph represents component adjacency and hierarchical attachment as textual priors for MLLM-based review, while visual-textual similarity organizes confusable candidate categories into review groups. Verified labels are propagated across views through feature matching. A semantic–appearance decoupled extraction stage then learns Gaussian category assignments and filtering weights while retaining the converged scene parameters. Weight optimization preserves global transmittance to account for occlusion while suppressing background remnants in independently rendered category subsets. Evaluation on a Qing-dynasty minor-style timber-frame building distinguishes image-level recognition, full-scene semantic reconstruction and independent component extraction; two additional architectural cases examine applicability and failure conditions. This framework connects spatially informed annotation with independently renderable category subsets for component-level access to architectural heritage scenes.

Methods

The framework takes multi-view images, component descriptions and annotated exemplars as inputs and produces category-specific Gaussian subsets for independent rendering. Figure 2 links relationship-guided image annotation, cross-view label propagation and semantic–appearance decoupled Gaussian extraction.

Spatial Relationship-Guided Component Annotation

Initial annotation follows Heritage-3DGS [28], using component descriptions and manually annotated exemplars in Supplementary Information (Table S1 and Figure S1). The descriptions cover material, shape, decoration and spatial relationships. Subsequent review uses construction relationships to distinguish confusable categories before Gaussian semantic training.

Each exemplar contains a local image crop, the target outline and neighboring context, linking a specialized component name to visible architectural features.

The Segment Anything Model (SAM) [41] supplies initial masks, which can be fragmented by painted decoration and internal edges. Image dilation and erosion are combined with Simple Linear Iterative Clustering (SLIC) [42] superpixels to complete these regions before annotation, as shown in Figure 3.

Preliminary annotation follows the composite-prompt structure of Heritage-3DGS shown in Figure 4. The resulting candidate labels are passed to spatially guided grouped review.

Inner, corner and eave columns illustrate how structural position distinguishes visually similar components in Figure 5. Construction relationships provide this contextual information during semantic review.

Table 1 supplies the Spatial Relationship Graph to the MLLM as a node-centered mapping. Nodes describe category geometry, texture and color; relations encode adjacency, hierarchical attachment, structural position and view-dependent occlusion. These descriptions guide category identification, while visual-textual similarity determines which candidates are reviewed together.

Reviewing all 22 categories together requires fine-grained discrimination within a long prompt. Reported limitations in fine-grained classification [43] and long textual and multimodal contexts [44,45] motivate restricting each review to potentially confusable categories.

Component categories are represented by a weighted graph for grouping, with edge weights measuring visual-textual similarity. Representative images are encoded with CLIP [24], and descriptions with GLM [46]. Figure 6 shows the resulting cosine-similarity matrices.

In the grouping graph, categories form nodes, spatial adjacency and viewing relationships constrain candidate connections, and visual-textual similarity supplies edge weights. Figure 7 shows this representation; the construction relationships in Table 1 guide the subsequent semantic review.

Image and text similarities receive a 9:1 weighting. Figure 7 shows adjacency relations and selected similarity links above the median, with node colors identifying the expert-defined initial groups.

For the component exemplar set $V = \{v_1, \ldots, v_{22}\}$, CLIP provides image embeddings $\mathbf{e}^{img}u, \mathbf{e}^{img}v \in \mathbb{R}^d$. Their cosine similarity measures visual similarity:

$$s^{img}(u,v) = \cos\left(\mathbf{e}_u^{img}, \mathbf{e}_v^{img}\right) = \frac{\mathbf{e}_u^{img} \cdot \mathbf{e}_v^{img}}{\left\|\mathbf{e}_u^{img}\right\| \left\|\mathbf{e}_v^{img}\right\|} \tag{1}$$

GLM [46] encodes the structured descriptions as text embeddings $\mathbf{e}^{txt}u, \mathbf{e}^{txt}v$, whose cosine similarity is calculated analogously.

$$s^{txt}(u,v) = \cos(\mathbf{e}_u^{txt}, \mathbf{e}_v^{txt}) \tag{2}$$

The weighted combination of image and text similarities forms the edge-weight matrix used for candidate selection.

For each component, the other categories are ranked by similarity. A candidate edge is retained only when both components belong to each other's Top-K neighbors, excluding one-sided similarity matches.

Initial groups are defined manually from functional and structural affiliations in timber construction, independently of the similarity values. These expert-defined groups form the units for subsequent merging.

Mutual Top-K relationships between component categories are consolidated with the expert-defined initial grouping to obtain the semantic review groups. The pseudocode below summarizes category-level mutual selection followed by fusion of the grouping results.

```
# V: component categories; G: expert pre-defined groups
# S: candidate similarity scores; K: neighbor count
# G': final fused groups
Initialize an empty set of mutually selected pairs P
for each component category u ∈ V do
    N[u] ← Top-K candidate neighbors of u ranked by S
end for
for each unordered pair of distinct categories (u, v) do
    if v ∈ N[u] and u ∈ N[v] then
        Add pair (u, v) to P
    end if
end for
    G' ← FuseGroupingResults(G, P)
return G'
```

Each review prompt combines the group's component descriptions, exemplars and construction relationships, enabling comparison of neighboring candidates. The complete architectural image provides context, and numbered, color-coded crops identify regions for category assignment. Figure

8 shows the prompt, with relationship-annotated exemplars provided in Supplementary Information, Figure S2.

A mask whose category is absent from the review group may be assigned None. Such masks undergo weak-recall re-identification against the full category set using the exemplars in Supplementary Information, Figure S2.

Verified anchor-frame labels are transferred to adjacent views using MASt3R [47]. A target mask inherits a label when its high-confidence pixel matches fall uniquely within one semantic mask in the anchor frame. Masks with ambiguous or insufficient matches remain unassigned. Figure 9 documents the propagation procedure.

When overlap with the current anchor falls below the preset threshold γ, propagation stops and the current frame becomes a new anchor I_B. The MLLM labels newly visible or unmatched regions in I_B. Once adjacent anchors I_A and I_B are annotated, intermediate frames undergo bidirectional pixel and mask matching with mutual verification. The MLLM then supplements masks that remain unmatched.

Semantic–Appearance Decoupled Gaussian Extraction

The extraction stage separates semantic learning from appearance optimization [48,49]. After base-scene reconstruction, it learns category assignments and retention weights with the converged appearance parameters fixed. The weights suppress background remnants and floating primitives exposed during independent rendering. Figure 10 shows base-scene reconstruction, semantic assignment, filtering and subsequent boundary scale optimization.

3D Gaussian Splatting is parameterized by a set of anisotropic 3D Gaussian primitives. Each individual Gaussian primitive $\{G_i\}$ is jointly defined by its center position $\mu_i \in \mathbb{R}^3$, scale matrix $s_i \in \mathbb{R}^3$, rotation quaternion $q_i \in \mathbb{R}^4$, opacity $a_i \in [0,1]$, and color c_i encoded by spherical harmonic (SH) coefficients. During rendering, using the rotation W, translation V, and the Jacobian of the projective transformation J, the 2D covariance matrix in pixel coordinates can be expressed as $\Sigma' = JW\Sigma W^T J^T$. All primitives projected onto the same pixel are then sorted by depth order and accumulated through an α-blending based volume rendering mechanism, yielding the final rendered color $C(p)$ at pixel p as follows.

$$C(p) = \sum_{i \in \mathcal{N}} c_i \, \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{3}$$

where $\mathcal{N}$ denotes the set of Gaussian primitives covering pixel p.

The input to the scene reconstruction stage consists of the camera intrinsics and extrinsics estimated by COLMAP [50] together with the sparse point cloud, which serve as the geometric initialization of the Gaussian scene. All Gaussian parameters are subsequently optimized via

backpropagation by comparing the rendered Gaussian image I_{render} and the ground-truth image I_{gt}. The optimization objective of this stage is to minimize the composite photometric loss $\mathcal{L}_{\text{rgb}}$ as follows.

$$\mathcal{L}_{\text{rgb}} = (1-\lambda)\|I_{\text{render}} - I_{\text{gt}}\| + \lambda\left(1 - \text{SSIM}\left(I_{\text{render}}, I_{\text{gt}}\right)\right) \tag{4}$$

where $\|\cdot\|$denotes the mean absolute error loss, SSIM denotes the structural similarity index, and λ is a balancing hyperparameter that maintains consistency of high-frequency structures. While the parameters are being optimized, the Gaussian primitives are periodically densified and pruned based on accumulated gradients, continuing until the appearance features of the scene converge.

The objective of semantic field reconstruction is to determine a unique class assignment for each 3D Gaussian primitive in the converged scene. The decoupled architecture keeps the reconstructed scene parameters fixed and introduces a learnable semantic embedding vector $\ell_i \in \mathbb{R}^K$ for each Gaussian primitive $\{G_i\}$, where K denotes the number of semantic categories. During the semantic field reconstruction stage, differentiable volume rendering is used to project ℓ_i into image space, yielding the per-pixel semantic logit map $\hat{L}(u,v)$ as follows.

$$\hat{L}(u,v) = \sum_{i \in N} \ell_i \alpha_i \prod_{j=1}^{i-1}\left(1-\alpha_j\right) \tag{5}$$

During semantic training, the projected opacity α_i and the transmittance chain are detached from the computation graph, so that backpropagated gradients flow only to the semantic parameter ℓ_i. The K-dimensional logit vector $\hat{L}(u,v)$ represents the assignment score of that pixel with respect to the K categories, and the integer class index read from the 2D semantic mask serves as the ground-truth supervision signal $y(u,v)$. The two are connected through a multi-class cross-entropy loss function as follows.

$$\mathcal{L}_{\text{sem}} = \frac{1}{|\Omega|}\sum_{(u,v)\in\Omega} \text{CE}\left(\hat{L}(u,v),\ y(u,v)\right) \tag{6}$$

where Ω denotes the set of validly annotated pixels, including the background class. To reduce computation per training step, M_{cls} categories are randomly sampled from those present in the current frame at each training step, and Ω is further constrained to include only pixels belonging to the selected categories.

After semantic field reconstruction is complete, the argmax index is taken for each Gaussian primitive to obtain its final discrete class assignment label, and the entire set of Gaussian primitives is partitioned into K disjoint category subsets as follows.

$$\mathcal{O}_c = \left\{ i \;\middle|\; \arg\max_k\ \ell_i^{(k)} = c \right\}, c = 0, \ldots, K-1 \tag{7}$$

Following the separately learned semantic-opacity idea in OP2GS [51], retention weights are learned to suppress unrelated Gaussians within each assigned category.

An additional weight $\phi_i \in [0,1]$ is introduced for each Gaussian primitive $\{G_i\}$ to weight the semantic class assignment obtained in the earlier stage. This retention weight is mapped from an

internal parameter $\tilde{\phi}_i$ via a Sigmoid function as follows.

$$\phi_i = \sigma(\tilde{\phi}_i) = \frac{1}{1 + e^{-\tilde{\phi}_i}} \tag{8}$$

Instead of updating the original opacity α_i, the occupancy rendering reuses the perspective geometric projection and spatial transmittance relationships established during the appearance reconstruction stage. The light transmittance must be jointly determined by all Gaussians in the scene to preserve correct occlusion relationships, while the accumulation of the 2D occupancy map is restricted to Gaussians within the target component subset $\mathcal{O}_c$. Based on this, the occupancy map $S_c(u, v)$ of the target component c under the current view is defined as follows:

$$S_c(u, v) = \sum_{i \in \mathcal{O}_c \cap \mathcal{N}} \phi_i \ell_i^{(k)} \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{9}$$

Gradients update only the filtering weights; the original geometry, opacity and color parameters remain fixed.

After occupancy map rendering is complete, it is aligned with the 2D ground-truth component masks and supervised using a Binary Cross-Entropy loss function as follows:

$$\mathcal{L}_{\text{obj}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{BCE}\left(S_c(u, v),\ \mathbf{1}[y(u, v) = c]\right) \tag{10}$$

Here, C denotes the categories with nonempty assigned Gaussian subsets. After optimization, a weight threshold selects the primitives retained for independent rendering. The threshold analysis in the "Quality of Independently Extracted Components" section examines how this selection affects extraction quality.

Boundary Gaussian Scale Optimization

After the filtering loss converges, the learned weight ϕ is thresholded to remove artifacts and residual background from architectural components. Large Gaussian primitives straddling component boundaries, however, can extend beyond the assigned component. They tend to occupy the edge regions of the semantic mask, producing jagged outlines and outward-projecting spikes in independent renderings. Figure 11 illustrates the boundary scale optimization procedure. The rendering gradient of the component drives the scale to shrink inward toward the interior of the mask, without performing any splitting operation. This refinement targets Gaussian coverage that extends beyond component mask boundaries.

This work unfreezes the anisotropic scale parameter s_i after ϕ_i converges, and introduces object-level color rendering to constrain scale updates using component appearance. To render component c, the opacity of non-target component Gaussians ($i \notin \mathcal{O}c$) is forced to zero. The color volume rendering method along ray $\mathcal{N}$ through pixel p is given as follows:

$$\hat{C}_c(p) = \sum_{i \in \mathcal{O}_c \cap \mathcal{N}} c_i \alpha_i \prod_{j<i} (1 - \alpha_j) \tag{11}$$

Both color accumulation and transmittance decay are restricted to the isolated component subset. The semantic occupancy objective follows Eq. (10), with the occupancy map defined in Eq. (9).

A subset $\mathcal{S}'$ is randomly sampled from the foreground component categories present in the current view. The joint optimization objective is defined as follows:

$$\mathcal{L}_{\text{refine}} = \frac{1}{|\mathcal{S}'|} \sum_{c \in \mathcal{S}'} \left(\lambda_{\text{rgb}} \cdot \mathcal{L}_{\text{rgb}}^{(c)} + \lambda_{\text{obj}} \cdot \mathcal{L}_{\text{obj}}^{(c)} \right) \tag{12}$$

During training, gradient updates are applied only to the anisotropic scale parameter s_i and the weight parameter ϕi, driving the Gaussian primitives at component edges to shrink toward the interior of the mask. These updates target cross-boundary protrusions and detached dark artifacts in independent component renderings.

Results

The main quantitative case is a Qing-dynasty minor-style timber-frame building. Evaluation covers component recognition, independent extraction and full-scene preservation, followed by two additional architectural cases.

A handheld DJI Osmo Action 6 recorded the building from approximately 2 to 3 m, yielding 118 selected frames at 7680 × 4320 pixels. The images were downsampled to 4096 × 2304 pixels for processing on an NVIDIA RTX 4060 Ti GPU. GLM-5V-Turbo [46] was used for semantic annotation, and point-prompted SAM2 [52] was used for background removal. A random 10% of the viewpoints were selected for quantitative evaluation. Figure 12 shows representative input images.

For quantitative evaluation, Gaussian semantic maps and RGB images were rendered using the selected evaluation-camera poses. The corresponding mask images served as ground truth for intersection over union (IoU). The real images served as ground truth for peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS). Image-level annotation, full-scene semantic rendering and independently rendered category subsets were evaluated separately. Tables S2–S5 provide the underlying per-class results.

Recognition of Visually Confusable Components

Figure 13 shows differences in the recognition of neighboring components with similar textures and colors. For Minor Architrave and Embracing Tie-beam, Heritage-3DGS annotations extend into adjacent components, whereas the proposed masks more closely match the ground truth. The Inner Column example further illustrates the importance of structural position in distinguishing categories. The proposed method identifies the inner column on the right, where Heritage-3DGS provides no corresponding annotation. These examples highlight the role of surrounding structural

context in distinguishing components with similar local appearance. Additional category-wise comparisons are provided in Figure S3.

Figure 14 shows differences in overall annotation coverage. In view a, the proposed method labels additional regions of the columns and adjacent components. In view b, fewer regions remain unassigned on the door panels, beneath the bench and around the platform. These results show improvements in both category recognition and annotation coverage, providing more complete image labels for subsequent Gaussian semantic learning. The workflow combines strong-recall prompting, grouped semantic review and weak-recall re-identification of masks that remain unassigned or fall outside the review group.

Figure 15 and Table S2 show that the magnitude of improvement varies across component categories. Compared with Heritage-3DGS, IoU increases by 31.29, 36.35 and 31.35 percentage points for Minor Architrave, Embracing Tie-beam and Inner Column, respectively. These gains correspond to the recognition differences illustrated in Figure 13. Hanging Lattice Fascia and Gable Wall have slightly lower IoU than Heritage-3DGS. Overall, the proposed method achieves higher IoU in 20 of the 22 categories, increasing mean IoU (mIoU) from 69.03% to 85.72%. Its mIoU also exceeds those of Grounded-SAM [53], SAM3 [54] and MaskCLIP [55].

Figure 16 shows that several pronounced category confusions in the preliminary annotations fall within the same review groups. These include confusions between architraves and frieze boards, inner and eave columns, and among platforms, steps and inclined side-stones. These category pairs correspond to the groups in Table 2, allowing review to focus on confusable components using their construction relationships and annotated exemplars. Grouping thus concentrates observed recognition difficulties within smaller candidate sets to support targeted semantic correction.

For singleton categories such as Corbel, subsequent weak-recall re-identification provides an additional recognition pathway. Masks assigned None during preliminary annotation or grouped review are re-examined against the full category set using annotated exemplars to recover missing component labels. Grouped review and weak-recall re-identification together support the completeness and category accuracy of the final annotations.

Quality of Independently Extracted Components

Rendering a Gaussian category subset independently exposes residual primitives previously obscured by surrounding components, directly affecting the extracted object's outline and appearance. In Figure 17, Heritage-3DGS retains isolated color patches around the eave column, pronounced streak-like artifacts around the Baofu Pattern and Minor Architrave, and extensive background remnants around the platform. The proposed method reduces these artifacts, revealing clearer column outlines, painted component surfaces and the main platform structure. In the

Hanging Lattice Fascia example, fewer peripheral remnants make the internal lattice openings easier to distinguish.

Figure 18 and Table S4 show higher extraction IoU for 21 of the 22 categories, with mIoU increasing from 28.01% to 69.24%. Baofu Pattern, Minor Architrave and Hanging Lattice Fascia improve by 72.17, 66.58 and 61.18 percentage points, respectively. These gains correspond to the reduced background remnants and clearer component outlines in Figure 17. Improvements occur in both painted and openwork components, demonstrating extraction performance across different component forms. Bench-under Lattice Fascia shows a slight IoU decrease, from 51.01% to 49.04%.

Improved extraction is also reflected in the rendering metrics of isolated components. Figure 19 and Table S5 show mean PSNR increasing from 17.05 to 24.30 dB, SSIM from 0.8991 to 0.9530, and LPIPS decreasing from 0.1946 to 0.0871. All three metrics improve for Baofu Pattern, Minor Architrave and Hanging Lattice Fascia, consistent with the clearer painted surfaces and lattice structures in Figure 17. Bench-under Lattice Fascia also improves across all three rendering metrics despite its slightly lower IoU, illustrating how silhouette agreement and appearance similarity capture different aspects of extraction quality.

Figure 20 illustrates the effect of the filtering threshold on independent rendering of Hanging Lattice Fascia. At a threshold of 0.0, the selected region retains a prominent column-like remnant and peripheral artifacts. Increasing the threshold to 0.1 removes most of these prominent remnants. Further increases reduce local background interference while keeping the main lattice structure discernible. Table 3 shows that mean IoU and all three rendering metrics improve as the threshold increases from 0.1 to 0.9, with 0.9 yielding the best mean performance. Tables S6–S8 provide the detailed category-wise results.

Preservation of Scene Appearance and Semantic Structure

Figure 21 and Table S3 show a full-scene semantic mIoU of 89.95%, compared with 90.49% for Heritage-3DGS, a decrease of 0.54 percentage points. Meanwhile, independent component extraction mIoU increases from 28.01% to 69.24%, a gain of 41.23 percentage points, as shown in Figure 18 and Table S4. This comparison shows that a small reduction in full-scene semantic accuracy accompanies a substantial improvement in independent component extraction. For applications requiring independent access to and visualization of components, this trade-off offers practical value through clearer component outlines and appearance and markedly reduced interference from surrounding scene remnants.

Figure 22 compares the input image-level semantic annotations (panel a, labeled “Image Semantic”) with the Gaussian semantic fields obtained through Decoupled Semantic Assignment (panel b), Artifact and Background Filtering (panel c), and Heritage-3DGS (panel d). Panels b and

c show successive stages of the semantic–appearance decoupled extraction method. Both the proposed method and Heritage-3DGS recover some component regions omitted from the image annotations, providing more complete semantic coverage in the reconstructed scene. The white boxes in panel c highlight local regions where the proposed semantic field shows reduced quality, with less accurate component boundaries or category assignments. The local errors complement the quantitative comparison of full-scene semantic accuracy in Figure 21.

Figure 23 presents the scene renderings after semantic assignment and filtering. Decoupled training at this stage keeps the original Gaussian geometry, opacity and color parameters fixed, learning only the additional semantic parameters and filtering weights. The original 3DGS scene appearance is therefore preserved. By comparison, Heritage-3DGS exhibits local color or shape changes in the regions marked by white boxes. Table 4 shows that the proposed method matches the reported original 3DGS values of 22.95 dB PSNR, 0.83 SSIM and 0.25 LPIPS. Preserving component textures, painted details and visual form maintains the existing visual record for digital documentation and image-based assessment of historic architecture.

Effect of Boundary Gaussian Scale Optimization

Following semantic assignment and filtering, boundary Gaussian scale optimization was evaluated as an additional refinement of the extracted components. Figure 24 compares component renderings before and after this refinement. Some Gaussian primitives extend across component boundaries, producing outward-projecting spikes when the components are rendered independently. Scale optimization contracts their coverage toward the component interior, reducing the protrusions highlighted by the red dashed boxes and producing more regular component outlines. Table 5 reports an increase in mean IoU from 69.24% after semantic assignment and filtering to 75.90% after boundary optimization, a gain of 6.66 percentage points. Table 6 reports rendering quality across filtering thresholds after boundary optimization: the highest PSNR and SSIM occur at 0.1, whereas the lowest LPIPS occurs at 0.9.

Applicability to Additional Architectural Cases

To examine applicability across traditional architectural settings, we evaluated the framework on Huacheng Temple and Sanqing Pavilion from Gao's datasets [56], shown in Figure 25. The buildings differ in structural organization and component configuration, and component categories, descriptions and annotated exemplars were adapted to each case. Huacheng Temple illustrates applicability to another building, while Sanqing Pavilion reveals limitations in recognizing repeated components across multiple stories.

At Huacheng Temple, the local examples in Figure 26 show architraves at different eave levels with similar geometric forms, textures and colors. The proposed annotations distinguish these categories more clearly than Heritage-3DGS, whose highlighted regions contain category confusion. Figure 27 shows the reconstructed Gaussian semantic field and independently extracted category subsets. The renderings separate upper- and lower-eaves architraves and display other categories, including columns, the plaque and steps, with their characteristic shapes and textures. Some isolated components retain residual artifacts. Together, these results demonstrate the applicability of the annotation and extraction workflow to a building with a different component configuration.

Sanqing Pavilion provides a contrasting case that exposes the limits of this applicability. Its multistory structure contains repeated eaves and windows with similar appearance, requiring distinctions across structural levels. In Figure 28, columns are recognized more reliably, whereas windows and eaves remain confused. Providing more detailed annotated component exemplars did not fully resolve these ambiguities. This persistent confusion identifies multistory repetition as a challenging condition for the current workflow.

Discussion

The results support the use of construction relationships to guide recognition of visually similar timber components. Building on the MLLM-based architectural annotation approach of Heritage-3DGS [28], the proposed workflow supplies adjacency and hierarchical attachment information during grouped semantic review. Connecting this annotation workflow to Gaussian category assignment and filtering provides independently renderable component categories within the reconstructed scene.

Full-scene semantic accuracy and independent extraction quality capture different properties of a Gaussian representation. Removing surrounding components exposes residual background and floating primitives that are less apparent in the complete scene. The extraction gains therefore show the value of evaluating category subsets independently when the goal is component-level access. Semantic assignment and filtering preserve the original scene appearance by keeping its Gaussian parameters fixed. Boundary scale optimization subsequently refines the extracted subsets by adjusting Gaussian coverage at component edges.

The Sanqing Pavilion case reveals a limitation of the spatial relationship representation. Repeated windows and eaves across multiple stories have similar appearance but are spatially separated, while overlapping eave levels obscure their relationships in local views. The confusion observed in Figure 28 suggests that local adjacency and hierarchical attachment provide insufficient context for these repeated components.

The resulting category subsets provide semantic access to photorealistic heritage records and a basis for associating architectural information with the reconstructed scene. Linking these subsets to individual repair records or HBIM elements would require distinguishing separate instances within each category. Cross-level relationships and instance identifiers are therefore directions for extending the framework toward component-specific conservation information management.

Data Availability

The original images collected for this study are available in the "traditional-timber-building-images" dataset on Hugging Face at https://doi.org/10.57967/hf/11393. The publicly available Chinese Ancient Architecture Image Database used in this study is hosted on Science Data Bank and is available at https://doi.org/10.57760/sciencedb.ai.00008.

Code Availability

The code developed in this study is not currently publicly available because it has not yet been packaged and documented for public release. It is available from the corresponding author upon reasonable request.

Acknowledgements

The authors used AI-assisted tools for translation and language polishing. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 42671556 and 42171416] and the National Science and Technology Major Project of China [grant number 2025ZD1700705]. The funding bodies had no role in the study design, data analysis, manuscript preparation or decision to submit the manuscript for publication.

Competing Interests

The authors declare no competing financial or non-financial interests.

Author contributions

L.W.: Conceptualization, Methodology, Software, Investigation, Visualization, Writing – original draft.

M.H.: Supervision, Project administration, Funding acquisition, Resources, Writing – review and editing.
W.M.: Software, Visualization, Validation.
D.L.: Investigation, Data curation.
X.X.: Formal analysis, Validation.

References

[1] Y. Hao, Z. Yao, R. Wu, Y. Bao, Damage and restoration technology of historic buildings of brick and wood structures: a review, Heritage Sci. 12 (2024) 301. https://doi.org/10.1186/s40494-024-01422-y.

[2] M. Riggio, D. D'Ayala, M.A. Parisi, C. Tardini, Assessment of heritage timber structures: review of standards, guidelines and procedures, J. Cult. Heritage 31 (2018) 220–235. https://doi.org/10.1016/j.culher.2017.11.007.

[3] T. Penjor, S. Banihashemi, A. Hajirasouli, H. Golzad, Heritage building information modeling (HBIM) for heritage conservation: framework of challenges, gaps, and existing limitations of HBIM, Digital Appl. Archaeol. Cult. Heritage 35 (2024) e00366. https://doi.org/10.1016/j.daach.2024.e00366.

[4] M. Riggio, V. Nasir, Digital twins in structural health assessment and monitoring: applications in historical timber buildings, Int. J. Archit. Heritage 19 (2025) 2398–2421. https://doi.org/10.1080/15583058.2024.2437618.

[5] R. Mora, L.J. Sanchez-Aparicio, M.A. Mate-Gonzalez, J. Garcia-Alvarez, M. Sanchez-Aparicio, D. Gonzalez-Aguilera, An historical building information modelling approach for the preventive conservation of historical constructions: application to the historical library of salamanca, Autom. Constr. 121 (2021) 103449. https://doi.org/10.1016/j.autcon.2020.103449.

[6] S. Huang, Q. Hu, M. Ai, P. Zhao, J. Li, H. Cui, S. Wang, Weakly supervised 3D point cloud semantic segmentation for architectural heritage using teacher-guided consistency and contrast learning, Autom. Constr. 168 (2024) 105831. https://doi.org/10.1016/j.autcon.2024.105831.

[7] I. Munoz-Pandiella, C. Bosch, M. Guardia, B. Cayuela, P. Pogliani, G. Bordi, M. Paschali, C. Andujar, P. Charalambous, Digital 3D models for medieval heritage: diachronic analysis and documentation of its architecture and paintings, Pers. Ubiquitous Comput. 28 (2024) 521–547. https://doi.org/10.1007/s00779-024-01816-6.

[8] J. Moyano Campos, E. Carreno, J.E. Nieto Julian, I. Gil Arizon, S. Bruno, Systematic

approach to generate historical building information modelling (HBIM) in architectural restoration project, Autom. Constr. 143 (2022) 104551. https://doi.org/10.1016/j.autcon.2022.104551.
[9] H. Xu, Y. Lu, Y. Song, N. Wang, F. Tian, Y. Li, Transforming panoramic images into 3D experiences for online cultural heritage visualization, Npj Herit. Sci. 14 (2026) 67. https://doi.org/10.1038/s40494-025-02239-z.
[10] Y. Shinde, K. Lee, B. Kiper, M. Simpson, S. Hasanzadeh, A systematic literature review on 360° panoramic applications in architecture, engineering, and construction (AEC) industry, J. Inf. Technol. Construct. 28 (2023) 405–437. https://doi.org/10.36680/j.itcon.2023.021.
[11] B. Kerbl, G. Kopanas, T. Leimkuehler, G. Drettakis, 3D gaussian splatting for real-time radiance field rendering, ACM Trans. Graphics 42 (2023) 139:1-139:14. https://doi.org/10.1145/3592433.
[12] G. Chen, W. Wang, A survey on 3D gaussian splatting, ACM Comput. Surv. 58 (2026) 313:1-313:39. https://doi.org/10.1145/3807511.
[13] Y. Yu, E. Verbree, P. van Oosterom, U. Pottgiesser, 3D gaussian splatting for modern architectural heritage: integrating UAV-based data acquisition and advanced photorealistic 3D techniques, AGILE: GIScience Ser. 6 (2025) 51. https://doi.org/10.5194/agile-giss-6-51-2025.
[14] S. He, P. Ji, Y. Yang, C. Wang, J. Ji, Y. Wang, H. Ding, A survey on 3D gaussian splatting applications: segmentation, editing, and generation, IEEE Trans. Pattern Anal. Mach. Intell. (2026) 1–20. https://doi.org/10.1109/TPAMI.2026.3704980.
[15] M. Qin, W. Li, J. Zhou, H. Wang, H. Pfister, LangSplat: 3D language gaussian splatting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024: pp. 20051–20060. https://doi.org/10.1109/CVPR52733.2024.01895.
[16] J.-C. Shi, M. Wang, H.-B. Duan, S.-H. Guan, Language embedded 3D gaussians for open-vocabulary scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024: pp. 5333–5343. https://doi.org/10.1109/CVPR52733.2024.00510.
[17] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, A. Kadambi, Feature 3DGS: supercharging 3D gaussian splatting to enable distilled feature fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024: pp. 21676–21685. https://doi.org/10.1109/CVPR52733.2024.02048.
[18] M. Ye, M. Danelljan, F. Yu, L. Ke, Gaussian grouping: segment and edit anything in 3D scenes, in: Computer Vision - ECCV 2024, Springer Nature Switzerland, 2025: pp. 162–179. https://doi.org/10.1007/978-3-031-73397-0_10.
[19] J. Cen, J. Fang, C. Yang, L. Xie, X. Zhang, W. Shen, Q. Tian, Segment any 3D gaussians, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2025: pp. 1971–

1979. https://doi.org/10.1609/aaai.v39i2.32193.
[20] Q. Shen, X. Yang, X. Wang, FlashSplat: 2D to 3D gaussian splatting segmentation solved optimally, in: Computer Vision - ECCV 2024, Springer, Cham, 2025: pp. 456–472. https://doi.org/10.1007/978-3-031-72670-5_26.
[21] J. Kerr, C.M. Kim, K. Goldberg, A. Kanazawa, M. Tancik, LERF: language embedded radiance fields, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2023: pp. 19672–19682. https://doi.org/10.1109/ICCV51070.2023.01807.
[22] M. Wang, J.-C. Shi, S.-H. Guan, H.-B. Duan, Language embedded 3D gaussians for open-vocabulary scene querying, IEEE Trans. Pattern Anal. Mach. Intell. 48 (2026) 4928–4941. https://doi.org/10.1109/TPAMI.2025.3648837.
[23] X. Hu, Y. Wang, L. Fan, C. Luo, J. Fan, Z. Lei, Q. Li, J. Peng, Z. Zhang, SAGD: boundary-enhanced segment anything in 3D gaussian via gaussian decomposition, IEEE Trans. Image Process. 35 (2026) 4934–4946. https://doi.org/10.1109/TIP.2026.3689408.
[24] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021: pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.
[25] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2023: pp. 11941–11952. https://doi.org/10.1109/ICCV51070.2023.01100.
[26] H. Xu, S. Xie, X.E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, C. Feichtenhofer, Demystifying CLIP data, in: International Conference on Learning Representations, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/d1450d6c10c6b6cf1b80964357f5fa08-Abstract-Conference.html.
[27] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, Natl. Sci. Rev. 11 (2024) nwae403. https://doi.org/10.1093/nsr/nwae403.
[28] Z. Liang, J.C.F. Chan, J. Zhang, Z. Liang, B. Wang, M. Wang, J.C.P. Cheng, Multimodal LLM-driven language-embedded 3D gaussian splatting for semantic and realistic digitization of historical buildings, Autom. Constr. 181 (2026) 106628. https://doi.org/10.1016/j.autcon.2025.106628.
[29] R. Yang, S. Mapesela, H. Li, R. Lorenzo, Mechanical properties of dougong bracket in chinese traditional timber structure under vertical loads: a systematic review, J. Build. Eng. 68 (2023) 106125. https://doi.org/10.1016/j.jobe.2023.106125.

[30] J. Zhang, M. Hou, J. Chen, X. Zhou, D. Hou, J. Ren, S. Yin, H. Zhao, Z. Zhang, A CityGML ADE for modeling ancient chinese timber architecture in 3D with semantic information, Npj Herit. Sci. 14 (2026) 271. https://doi.org/10.1038/s40494-026-02535-2.

[31] X. Wang, W. Ma, T. Zhang, C.M. de Melo, J. Chen, A.L. Yuille, Spatial457: a diagnostic benchmark for 6D spatial reasoning of large Mutimodal models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025: pp. 24669–24679. https://doi.org/10.1109/CVPR52734.2025.02297.

[32] W. Ma, H. Chen, G. Zhang, Y.-C. Chou, J. Chen, C. de Melo, A.L. Yuille, 3DSRBench: a comprehensive 3D spatial reasoning benchmark, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2025: pp. 6924–6934. https://doi.org/10.1109/ICCV51701.2025.00651.

[33] J. Lee, Y. Wang, J. Li, M. Zhang, Multimodal reasoning with multimodal knowledge graph, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024: pp. 10767–10782. https://doi.org/10.18653/v1/2024.acl-long.579.

[34] Y. Huang, J. Tang, Z. Chen, R. Zhang, X. Zhang, W. Chen, Z. Zhao, Z. Zhao, T. Lv, Z. Hu, W. Zhang, Structure-CLIP: towards scene graph knowledge to enhance multi-modal structured representations, Proc. AAAI Conf. Artif. Intell. 38 (2024) 2417–2425. https://doi.org/10.1609/aaai.v38i3.28017.

[35] J. Wang, J. Ju, J. Luan, Z. Deng, LLaVA-SG: leveraging scene graphs as visual semantic expression in vision-language models, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025: pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10887586.

[36] C. Mitra, B. Huang, T. Darrell, R. Herzig, Compositional chain-of-thought prompting for large multimodal models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2024: pp. 14420–14431. https://doi.org/10.1109/CVPR52733.2024.01367.

[37] C. Zheng, D. Liang, W. Zhang, X.-Y. Wei, T.-S. Chua, Q. Li, A picture is worth a graph: a blueprint debate paradigm for multimodal reasoning, in: Proceedings of the 32nd ACM International Conference on Multimedia, ACM, 2024: pp. 419–428. https://doi.org/10.1145/3664647.3681102.

[38] V.A. Carriero, A. Gangemi, M.L. Mancinelli, L. Marinucci, A.G. Nuzzolese, V. Presutti, C. Veninata, ArCo: the italian cultural heritage knowledge graph, in: The Semantic Web - ISWC 2019, Springer, Auckland, New Zealand, 2019: pp. 36–52. https://doi.org/10.1007/978-3-030-30796-7_3.

[39] J. Zhang, J. Jiang, Y. Chen, K. Jiang, X. Liu, COB-GS: clear object boundaries in 3DGS segmentation based on boundary-adaptive gaussian splitting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025: pp. 19335–19344. https://doi.org/10.1109/CVPR52734.2025.01801.

[40] H. Zhai, H. Li, Z. Li, X. Pan, Y. He, G. Zhang, PanoGS: gaussian-based panoptic segmentation for 3D open vocabulary scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2025: pp. 14114–14124. https://doi.org/10.1109/CVPR52734.2025.01317.

[41] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollar, R. Girshick, Segment anything, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2023: pp. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371.

[42] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Suesstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2274–2282. https://doi.org/10.1109/TPAMI.2012.120.

[43] G. Geigle, R. Timofte, G. Glavas, African or European swallow? Benchmarking large vision-language models for fine-grained object classification, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024: pp. 2653–2669. https://doi.org/10.18653/v1/2024.emnlp-main.154.

[44] N.F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: how language models use long contexts, Trans. Assoc. Comput. Linguist. 12 (2024) 157–173. https://doi.org/10.1162/tacl_a_00638.

[45] W. Wang, S. Zhang, Y. Ren, Y. Duan, T. Li, S. Liu, M. Hu, Z. Chen, K. Zhang, L. Lu, X. Zhu, P. Luo, Y. Qiao, J. Dai, W. Shao, W. Wang, Needle in a multimodal haystack, in: Advances in Neural Information Processing Systems 37, Neural Information Processing Systems Foundation, Inc., Vancouver, Canada, 2024: pp. 20540–20565. https://doi.org/10.52202/079017-0649.

[46] Glm-V Team, W. Hong, X. Gu, Z. Pan, Z. Yang, Y. Wang, Y. Wang, Y. Yue, Y. Wang, Y. Wang, Y. Wang, X. Liu, W. Yu, W. Wang, W. Li, S. Duan, S. Yang, R. Lv, M. Liu, L. Pan, K. Ning, J. Ji, J. Wang, J. Chen, J. Xu, J. Zhu, J. Cheng, J. Qi, G. Gan, G. Wang, C. Yao, Z. Dou, Z. Zhou, Z. Wang, Z. Ge, Z. Li, Z. Hou, Z. Xue, Z. Wang, Z. Qi, Z. He, Y. Zhang, Y. Liu, Y. Cen, Y. Li, Y. Wang, Y. Yang, Y. Liu, Y. Lu, Y. Xu, Y. Wang, Y. Zhao, Y. Wang, Y. Xue, Y. Xu, X. Zhang, X. Liu, X. Liu, W. Zhao, W. Li, T. Tong, T. Zhang, S. Zhang, S. Yan, Q. Zheng, M. Xu, L. Bao, Lat Long Long, J. Xu, J. Fan, J. Qian, J. Chen, J. Lin, J. Sun, H. Zheng, H. Wang, H. Li, H. Lai, H. Xu, F. Yang, D. Zhang, D. Yin, C. Zhao, C. Wu, B. Shi, B. Lv, B. Jia, B. Li, B. Chen, B. Wang, P. Zhang, D. Liu, B. Xu, J. Li, M. Huang, Y. Dong, J. Tang, GLM-5V-turbo: toward a native foundation

model for multimodal agents, arXiv Prepr. arXiv:2604.26752 (2026).
https://doi.org/10.48550/arXiv.2604.26752.
[47] V. Leroy, Y. Cabon, J. Revaud, Grounding image matching in 3D with MASt3R, in: Computer Vision - ECCV 2024, Springer Nature Switzerland, 2025: pp. 71-91. https://doi.org/10.1007/978-3-031-73220-1_5.
[48] S. Katragadda, C.-Y. Wu, Y. Guo, X. Huang, G. Huang, L. Ren, Online language splatting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2025: pp. 25882–25892. https://doi.org/10.1109/ICCV51701.2025.02401.
[49] Q. Peng, B. Planche, Z. Gao, M. Zheng, A. Choudhuri, T. Chen, C. Chen, Z. Wu, 3D vision-language gaussian splatting, in: International Conference on Learning Representations, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/98ed250b203d1ac6b24bbcf263e3d4a7-Abstract-Conference.html.
[50] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016: pp. 4104–4113. https://doi.org/10.1109/CVPR.2016.445.
[51] G. Liu, N. Vaara, J. Mustaniemi, J. Kannala, J. Heikkila, OP2GS: object-aware 3D gaussian splatting with dual-opacity primitives, arXiv Prepr. arXiv:2605.20044 (2026). https://doi.org/10.48550/arXiv.2605.20044.
[52] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Radle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K.V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, C. Feichtenhofer, SAM 2: segment anything in images and videos, in: International Conference on Learning Representations, 2025. https://mlanthology.org/iclr/2025/ravi2025iclr-sam/.
[53] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, L. Zhang, Grounded SAM: assembling open-world models for diverse visual tasks, arXiv Prepr. arXiv:2401.14159 (2024). https://doi.org/10.48550/arXiv.2401.14159.
[54] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris Coll-Vinent, C. Ryali, K.V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Radle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H.K. Cheng, P. Dollar, N. Ravi, K. Saenko, P. Zhang, C. Feichtenhofer, SAM 3: segment anything with concepts, in: International Conference on Learning Representations, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.
[55] C. Zhou, C.C. Loy, B. Dai, Extract free dense labels from CLIP, in: Computer Vision - ECCV

674 2022, Springer, 2022: pp. 696–712. https://doi.org/10.1007/978-3-031-19815-1_40.
675 [56] W. Gao, Chinese ancient architecture image database, (2024).
676 https://doi.org/10.57760/sciencedb.ai.00008.

Figures and legends

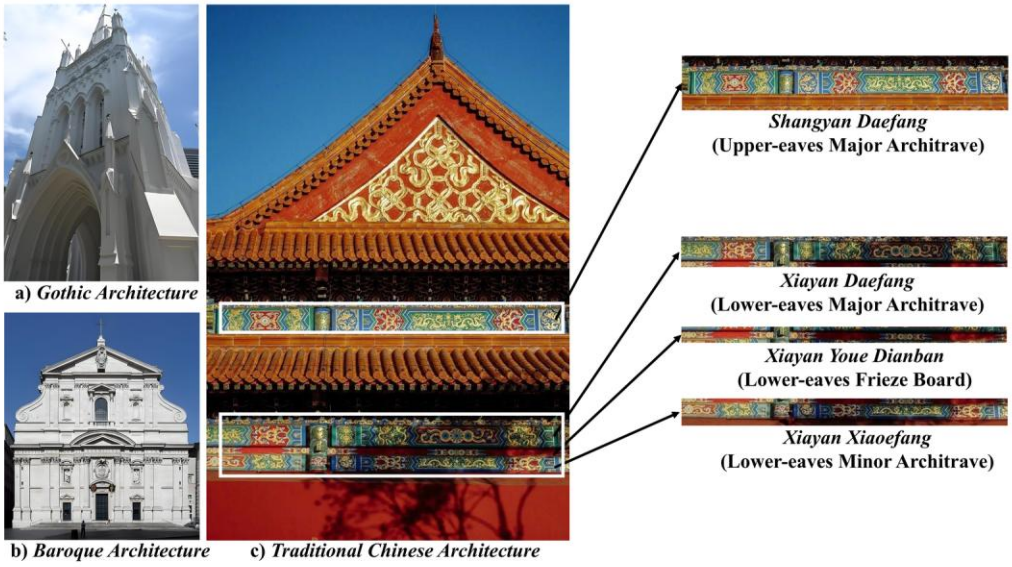


Figure 1. Architectural types and visually similar components in traditional Chinese timber architecture. a) Gothic architecture. b) Baroque architecture. c) Traditional Chinese architecture with enlarged component views. From top to bottom, the enlarged views show the upper-eaves major architrave, lower-eaves major architrave, lower-eaves frieze board and lower-eaves minor architrave. White rectangles identify the component regions, and black arrows connect them to the corresponding enlarged views.

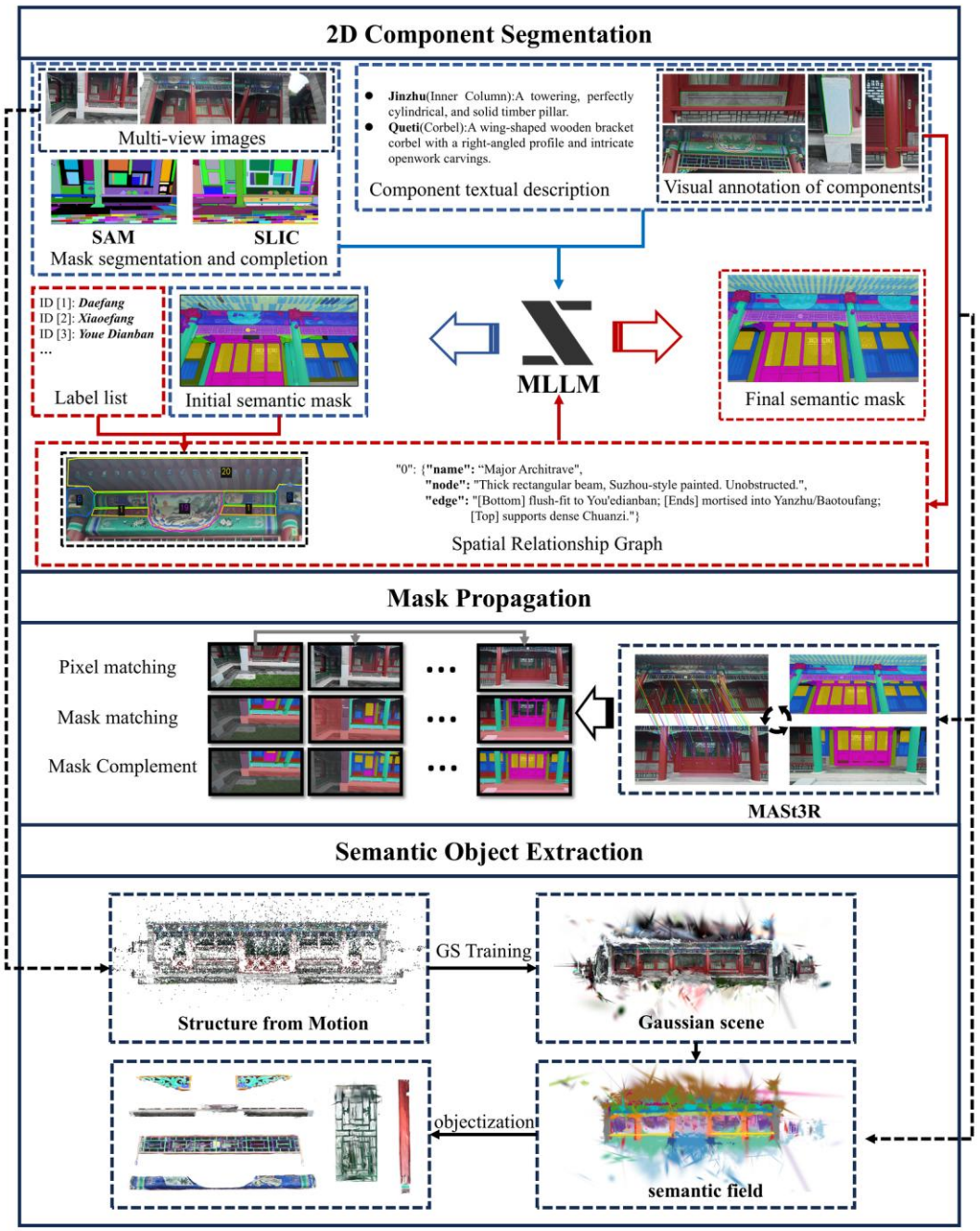


Figure 2. Framework for spatial relationship-guided component annotation and Gaussian object extraction. The workflow comprises two-dimensional component annotation, cross-view mask propagation and Gaussian semantic object extraction. The Segment Anything Model (SAM) and Simple Linear Iterative Clustering (SLIC) generate and refine candidate masks. A multimodal large language model (MLLM) identifies components using textual descriptions, image exemplars and spatial relationships. Cross-view matching propagates semantic labels, which subsequently guide Gaussian semantic assignment and component extraction. Colored regions indicate semantic categories, and arrows show processing order and information flow.

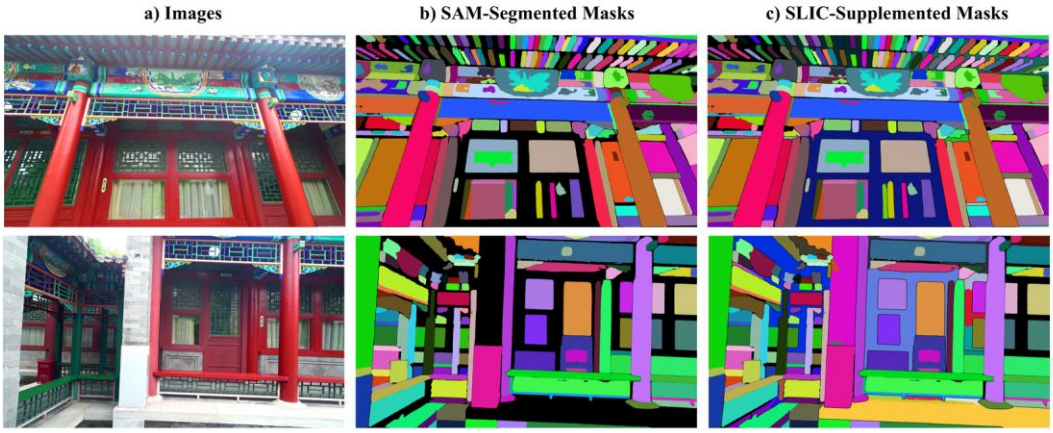


Figure 3. Initial SAM masks and SLIC-assisted mask completion. a) Input images. b) Initial masks generated by SAM. c) Masks after SLIC-assisted completion. The two rows show different architectural views, and colors distinguish segmented regions. SAM denotes the Segment Anything Model, and SLIC denotes Simple Linear Iterative Clustering.

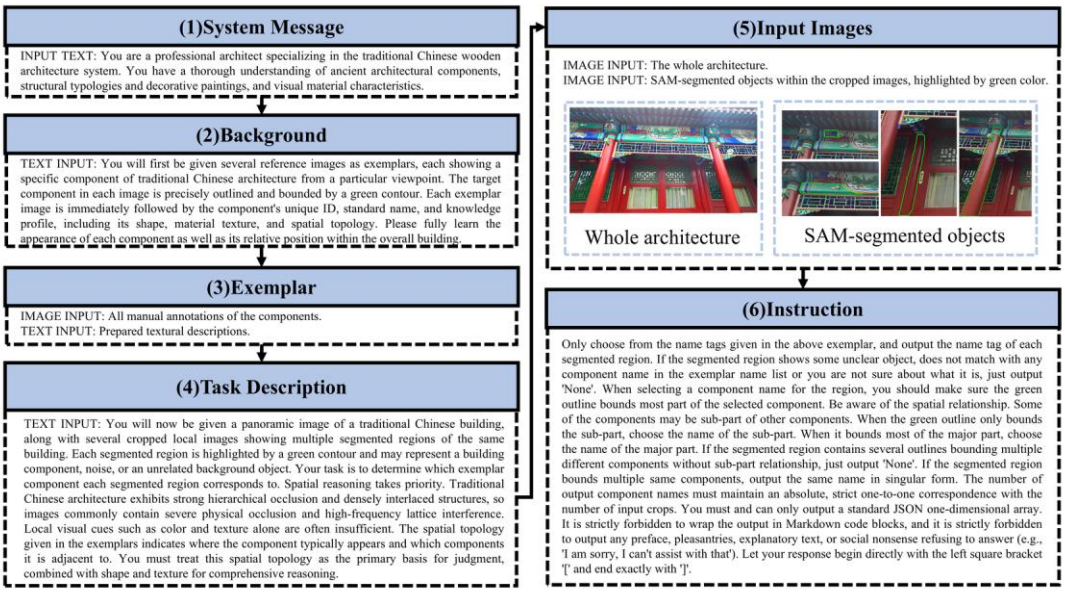


Figure 4. Prompt structure for preliminary semantic annotation. The prompt comprises a system message, background knowledge, component exemplars, a task description, input images and output instructions. MLLM denotes a multimodal large language model, and SAM denotes the Segment Anything Model. Colored outlines in the input images identify candidate regions.

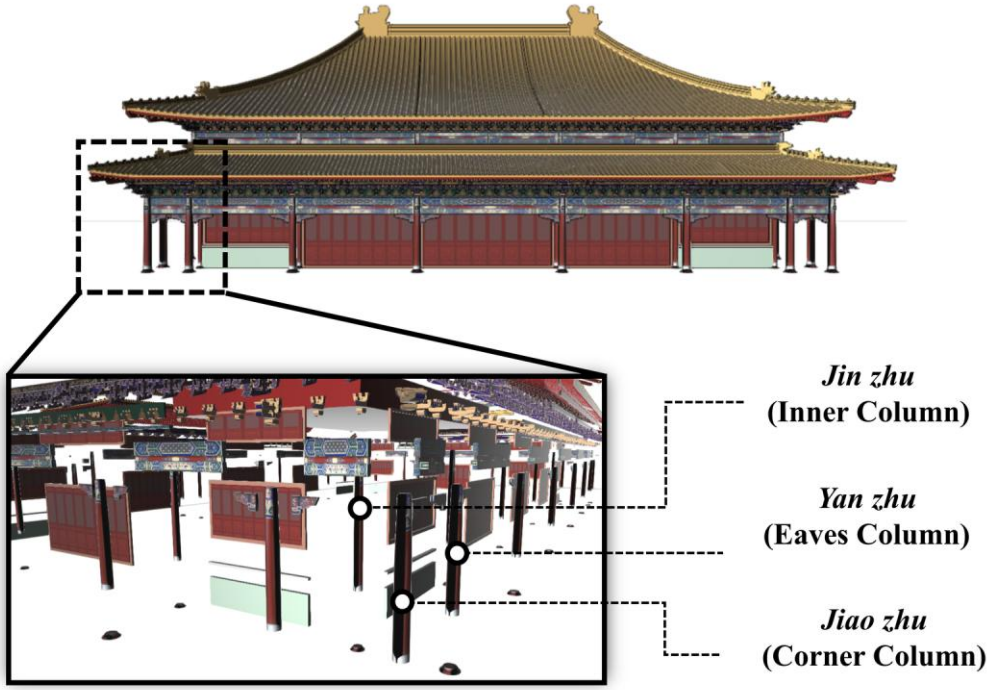


Figure 5. Relationships between structural position and column names. The upper image shows the building, and the lower enlargement shows the column arrangement. Dashed outlines and connecting lines identify corresponding regions. The labels indicate inner columns, eave columns and corner columns.

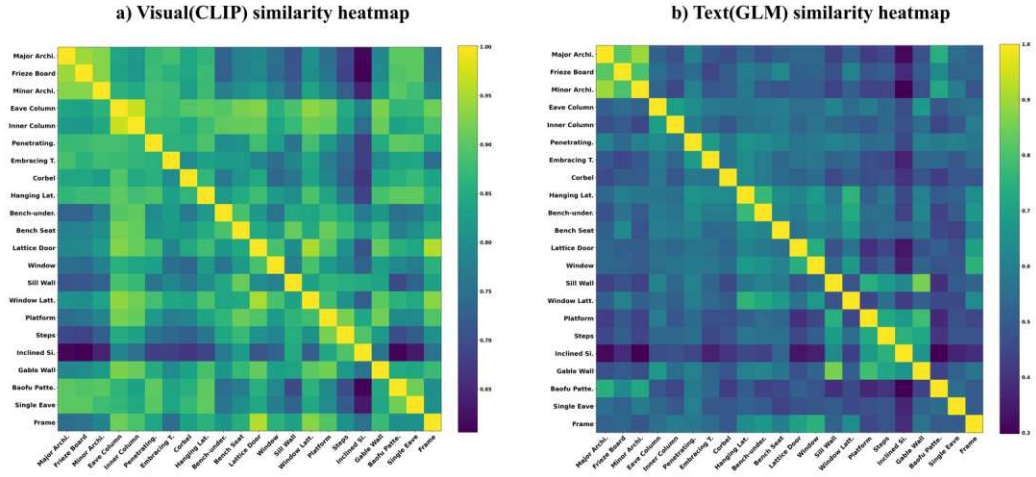


Figure 6. Visual and textual similarity matrices of component categories. a) Visual cosine similarity calculated from CLIP image features. b) Textual cosine similarity calculated from GLM text features. Both axes list component categories. Colors progress from purple to yellow as similarity increases, with numerical ranges shown by the respective color bars. CLIP denotes Contrastive Language–Image Pre-training, and GLM denotes General Language Model.

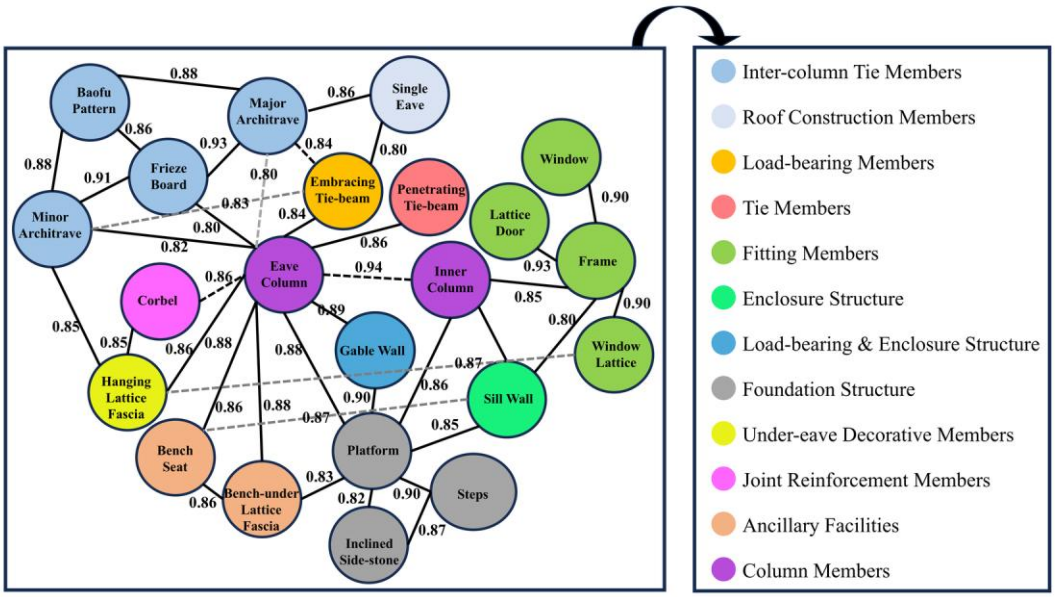


Figure 7. Spatial relationships and similarity links for component grouping. Nodes represent component categories, and node colors correspond to the initial functional or construction groups listed in the legend. Edges indicate retained category relationships, with adjacent numbers giving the combined visual–textual similarity scores. Node labels identify the components.

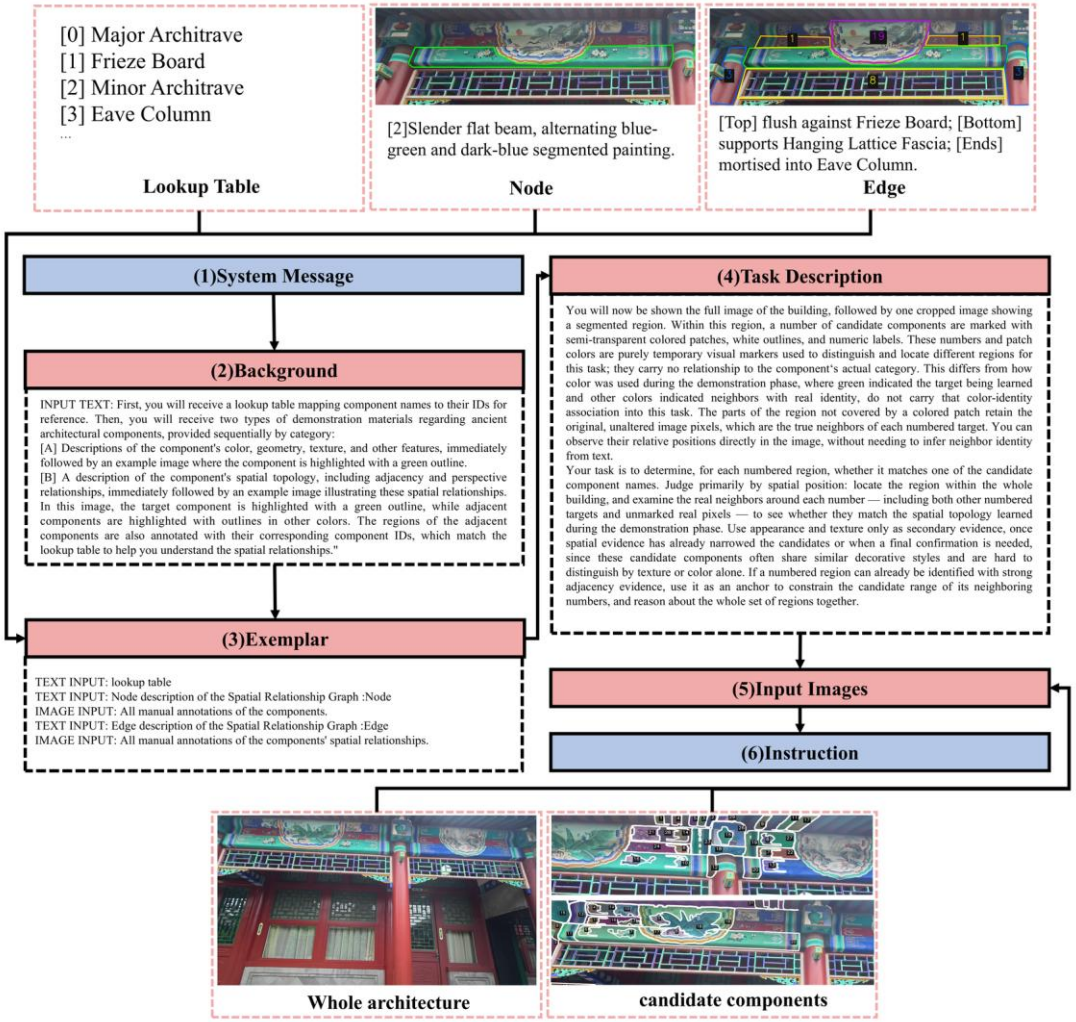


Figure 8. Prompt structure for spatial relationship-guided semantic re-verification within groups. The prompt comprises a system message, background knowledge, component exemplars, a task description, input images and output instructions, incorporating component categories and spatial relationship descriptions. MLLM denotes a multimodal large language model.

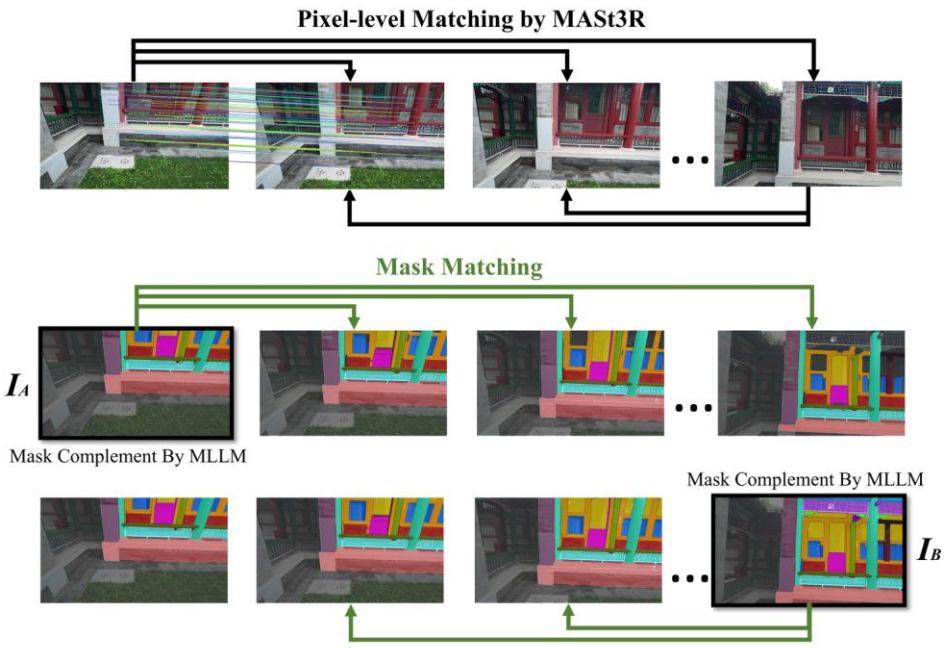


Figure 9. Cross-view semantic mask propagation using MASt3R matching. The upper part shows cross-view pixel matching, and the lower part shows semantic mask matching, bidirectional propagation and supplementary annotation. I_A and I_B denote two anchor frames. Black arrows connect images used for pixel matching, green arrows indicate mask matching and propagation directions, and ellipses represent intermediate views. Colored regions indicate semantic masks. MLLM denotes a multimodal large language model.

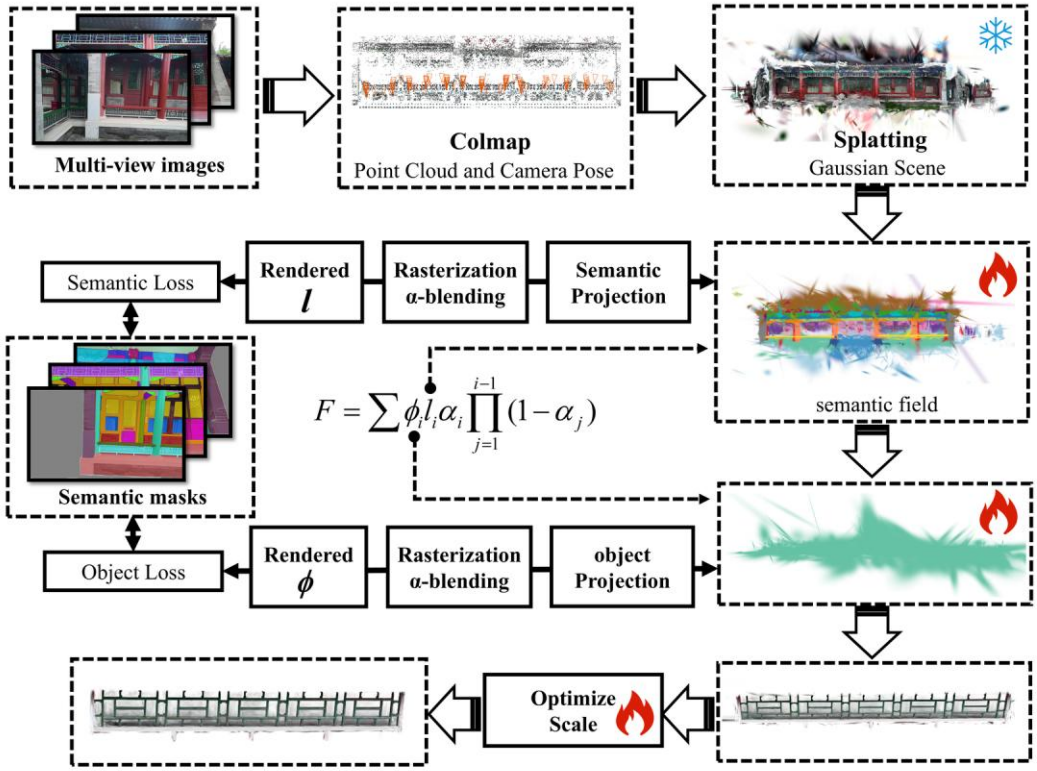


Figure 10. Training workflow for Gaussian semantic object extraction and boundary refinement. COLMAP recovers camera parameters and a sparse point cloud from multi-view images, followed by three-dimensional Gaussian splatting (3DGS) reconstruction, semantic assignment, component filtering and boundary scale optimization. Snowflake symbols indicate frozen parameters, and flame symbols indicate trainable parameters.

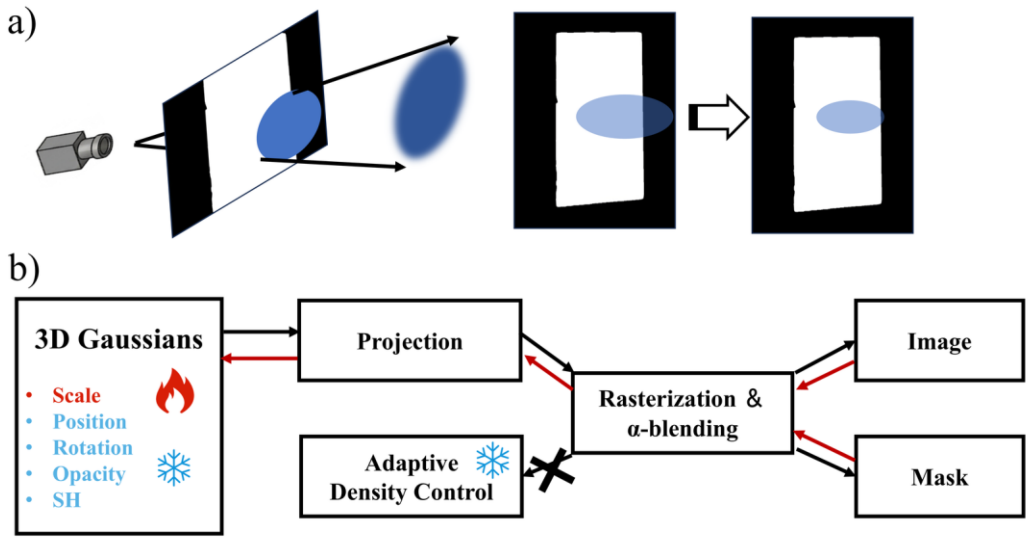


Figure 11. Principles and training workflow of boundary Gaussian scale optimization. a) Gaussian projection and scale reduction. b) Parameter optimization workflow. Blue ellipses represent Gaussian coverage, and black-and-white images represent target masks. Flame symbols indicate trainable parameters, snowflake symbols indicate frozen parameters, and the black cross indicates that adaptive density control is disabled. SH denotes spherical harmonic coefficients, and alpha-blending denotes opacity-based color compositing. Arrows indicate the flow of rendering and supervision information.



Figure 12. Acquired images before and after background filtering. The upper row shows the acquired images before background filtering, and the lower row shows the corresponding filtered images. Each column presents the same viewpoint, illustrating the retained building regions and removed surroundings.

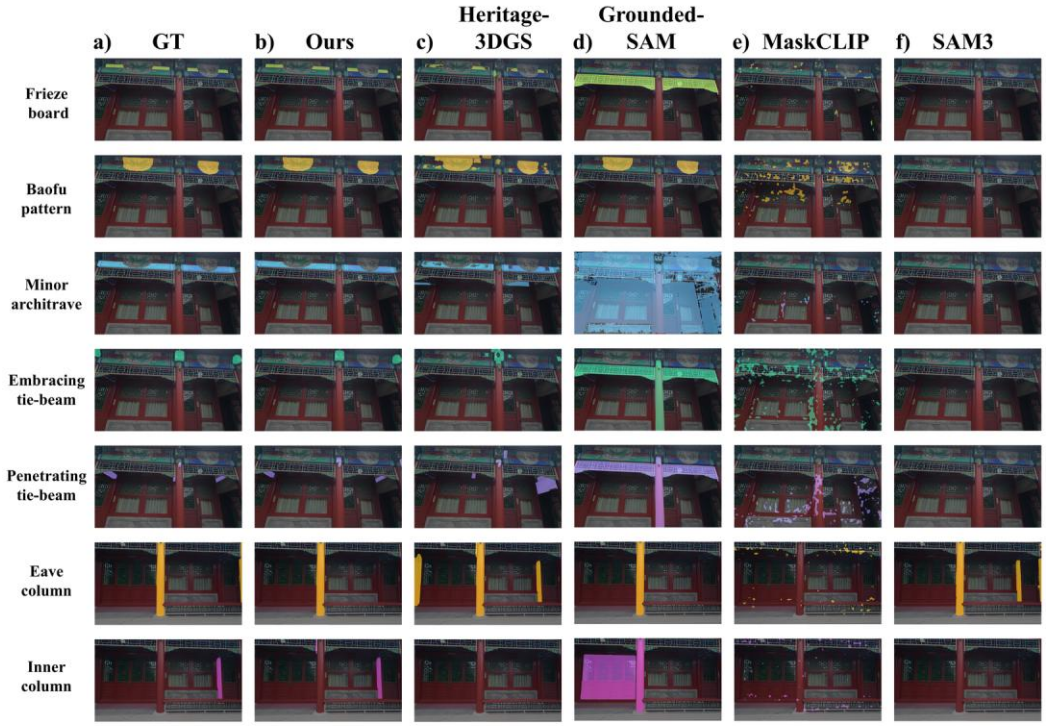


Figure 13. Image-level semantic segmentation of visually similar components. a) Ground truth. b) Our method. c) Heritage-3DGS. d) Grounded-SAM. e) MaskCLIP. f) SAM3. Rows correspond to the component categories labeled on the left. Colored masks overlaid on the photographs show the segmented target regions. GT denotes ground truth, SAM denotes Segment Anything Model, and CLIP denotes Contrastive Language–Image Pre-training.

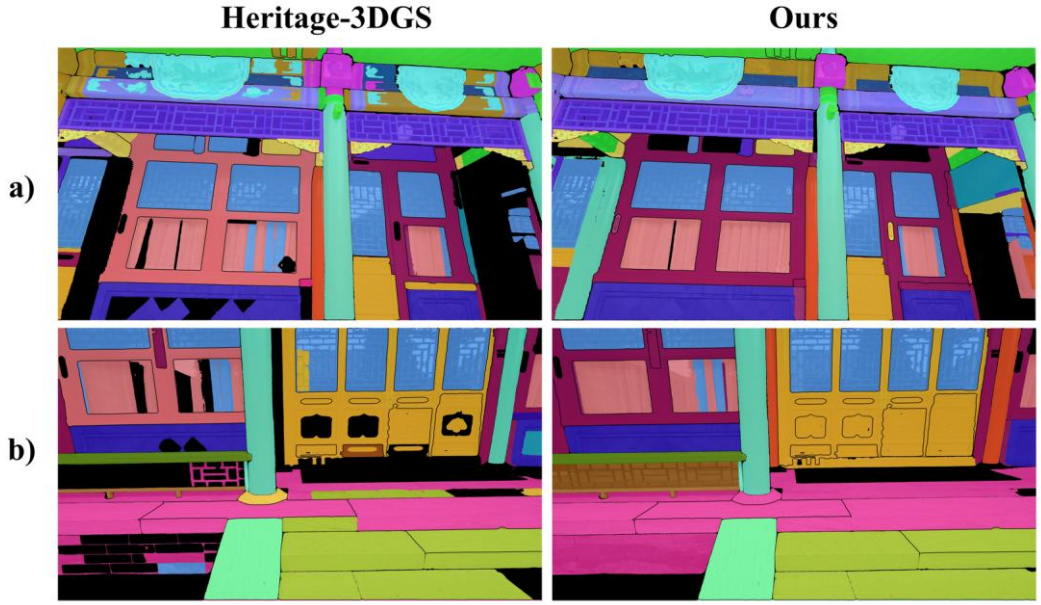


Figure 14. Whole-image semantic annotations from two architectural viewpoints. a) and b) show two viewpoints, with Heritage-3DGS in the left column and our method in the right column. Colored regions indicate component categories, while uncovered regions retain their original photographic appearance.

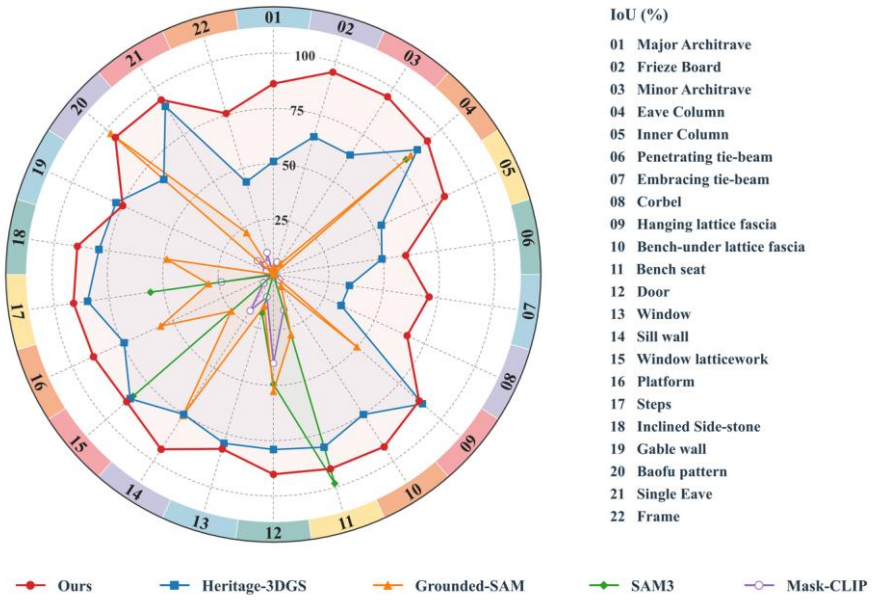


Figure 15. Image-level semantic segmentation accuracy across 22 component categories. Each radial axis represents a component category, with the outer numbers corresponding to the category list on the right. Radial values indicate intersection over union (IoU), with higher values farther from the center. Red, blue, orange, green and purple curves represent our method, Heritage-3DGS, Grounded-SAM, SAM3 and MaskCLIP, respectively. SAM denotes Segment Anything Model, and CLIP denotes Contrastive Language–Image Pre-training.

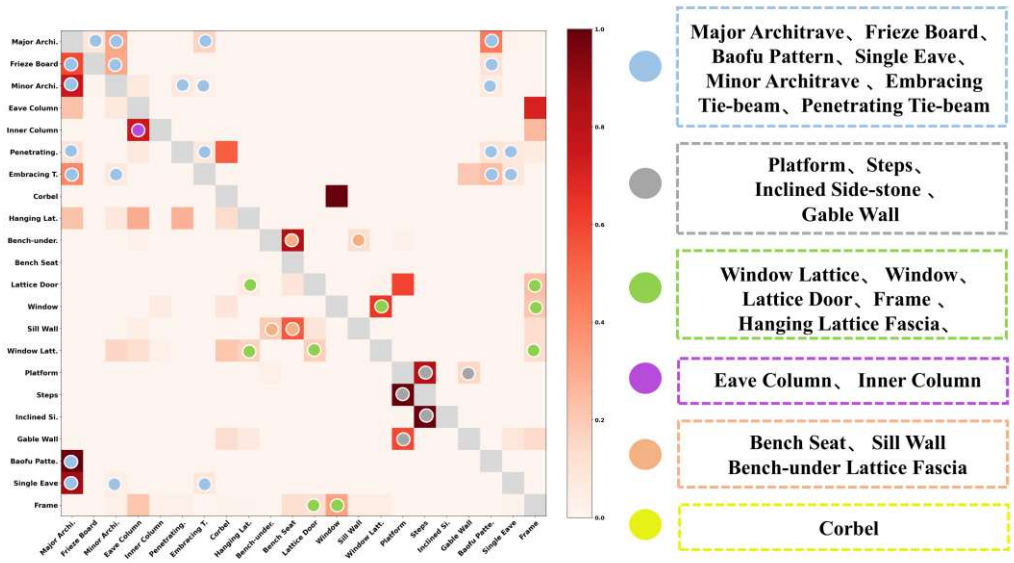


Figure 16. Overlap between preliminary category confusions and semantic review groups. Both axes list component categories. Heatmap colors progress from light shades to dark red as confusion values increase, with the numerical range shown by the color bar. Colored circles mark overlaps between confused category pairs and the review groups. Light blue, gray, green, purple, light orange and yellow circles and matching dashed boxes identify the six review groups, whose component categories are listed on the right.

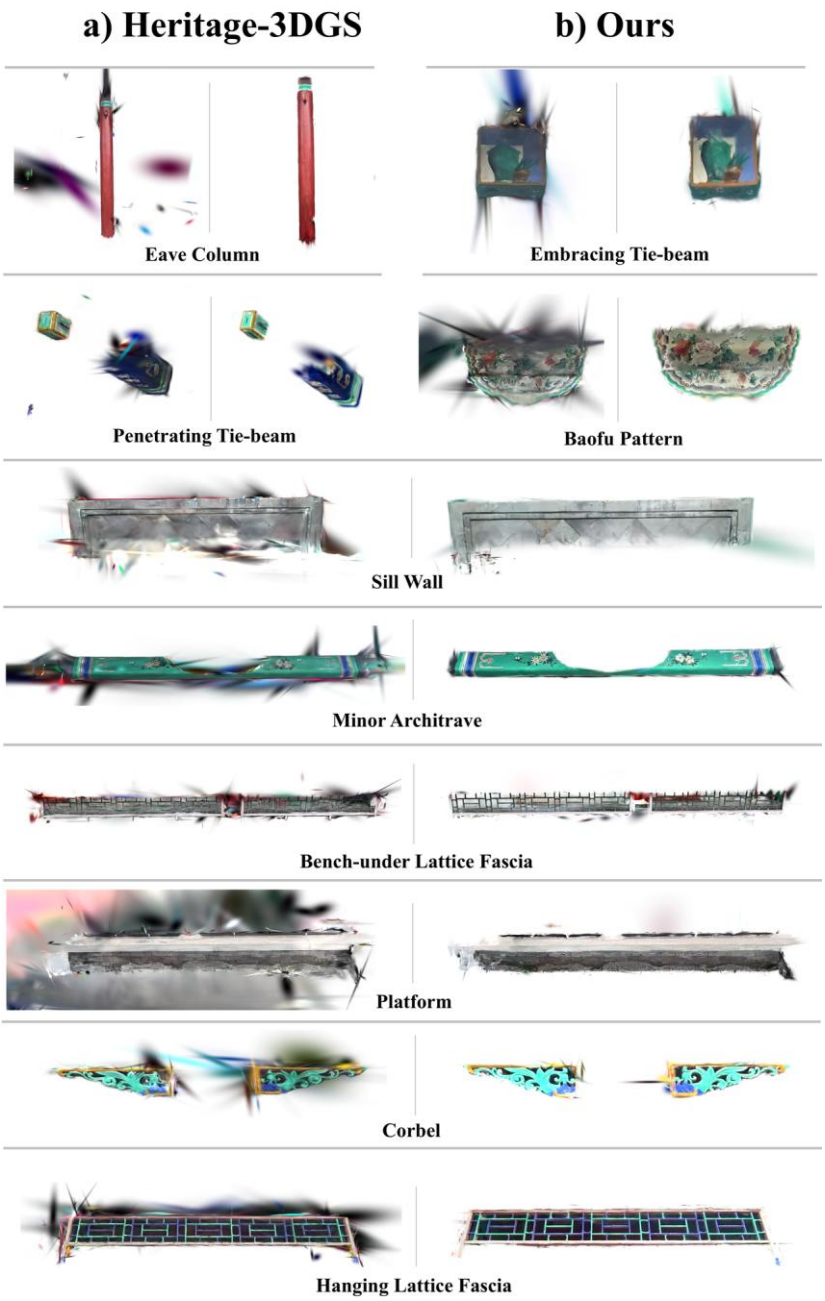


Figure 17. Independent renderings of extracted component categories. a) Heritage-3DGS. b) Our method. Each group shows the component category identified by its label, including columns, beams, painted elements, walls, lattice elements and platforms. Paired results correspond to the same category and are displayed against a white background.

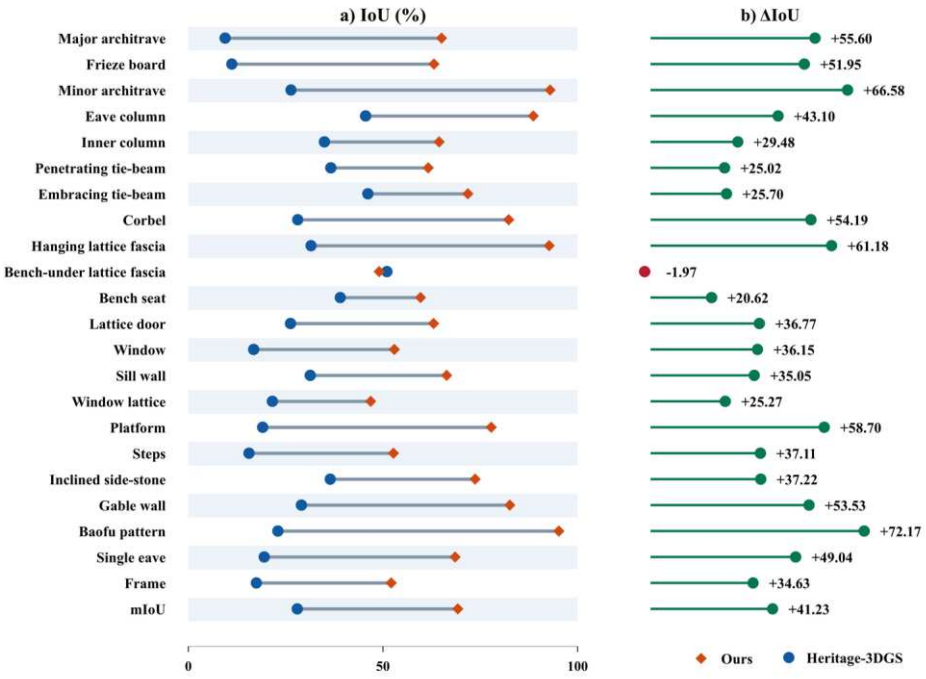


Figure 18. Independent component extraction accuracy and relative IoU gains. a) Category-wise intersection over union (IoU) for our method and Heritage-3DGS. b) IoU differences (ΔIoU) between our method and Heritage-3DGS, expressed in percentage points. Orange-red diamonds represent our method, blue circles represent Heritage-3DGS, and connecting lines link results for the same category. Green markers indicate positive gains, and red markers indicate negative gains. mIoU denotes mean intersection over union across categories.

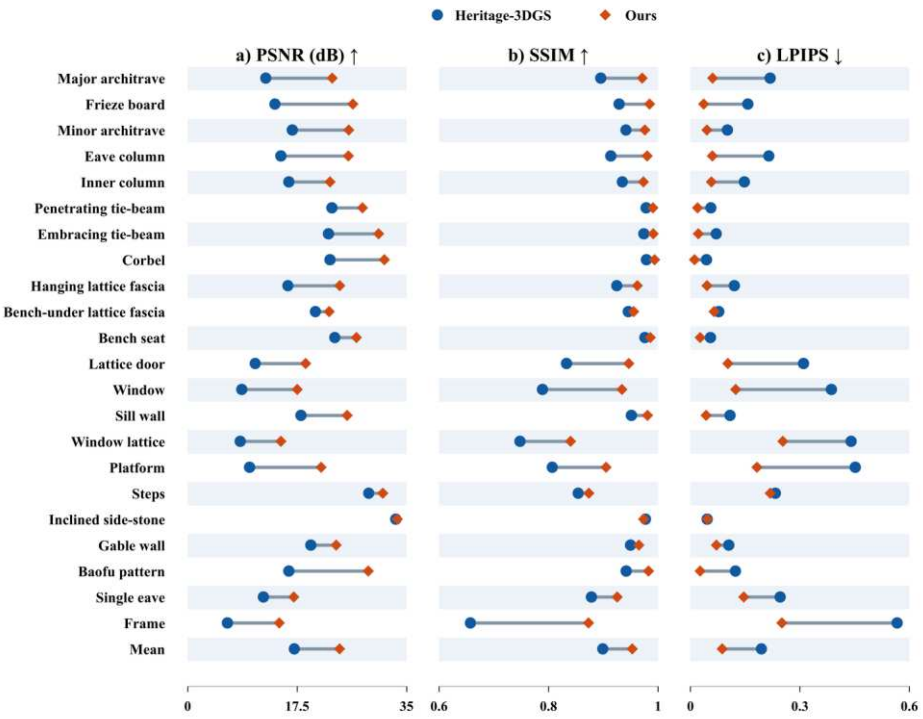


Figure 19. Rendering quality of independently extracted components. a) Peak signal-to-noise ratio (PSNR), measured in decibels (dB). b) Structural similarity index (SSIM). c) Learned Perceptual Image Patch Similarity (LPIPS). Blue circles represent Heritage-3DGS, orange-red diamonds represent our method, and connecting lines link results for the same category. Upward arrows indicate that higher values are better, and downward arrows indicate that lower values are better. Mean denotes the average across categories.

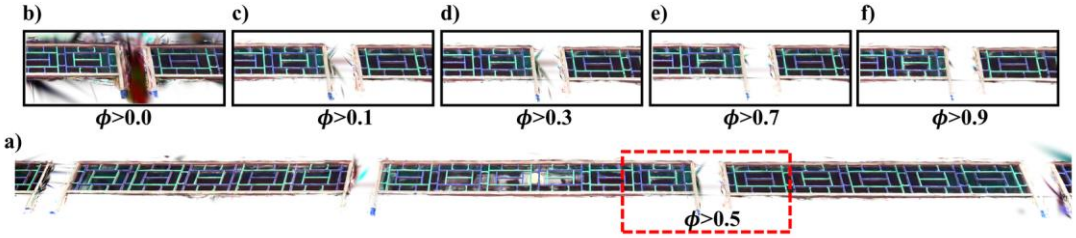


Figure 20. Independent renderings under different component retention thresholds. a) The overall rendering at a threshold of 0.5. b)–f) Local results at thresholds of 0.0, 0.1, 0.3, 0.7 and 0.9, respectively. The red dashed rectangle identifies the region selected for local comparison.

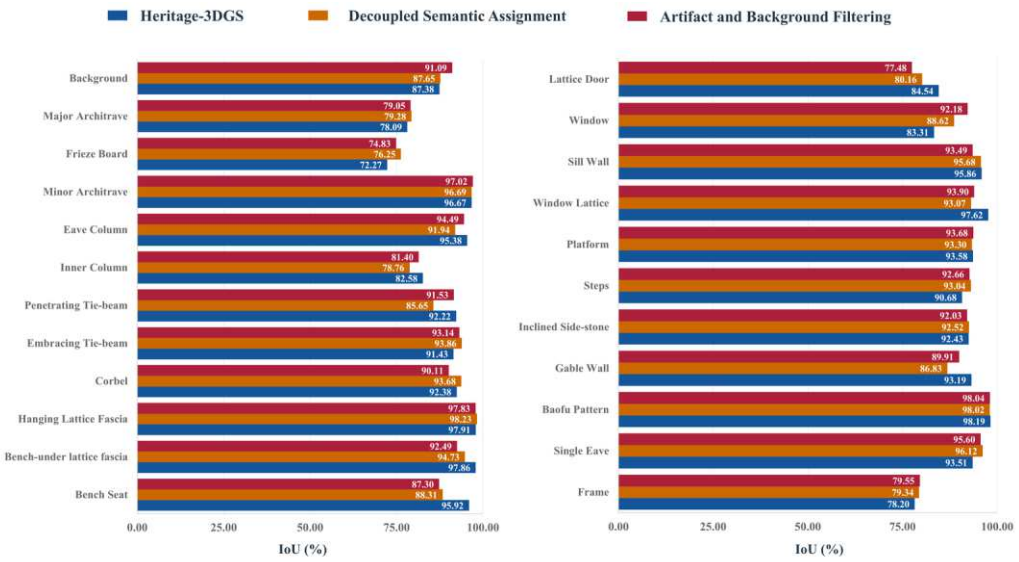


Figure 21. Full-scene category IoU after Gaussian semantic assignment and filtering. Blue, orange and red bars represent Heritage-3DGS, decoupled semantic assignment, and artifact and background filtering, respectively. The left and right parts display different categories. The horizontal axis gives intersection over union (IoU), and the numbers at the ends of the bars give the corresponding percentages.

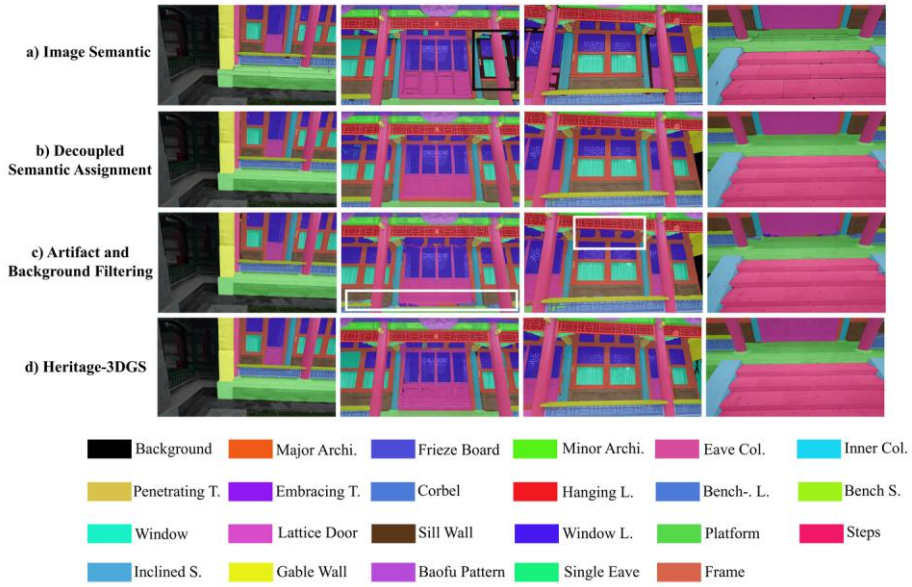


Figure 22. Image-level semantic annotations and reconstructed Gaussian semantic fields. a) Input image-level semantic annotations, labeled "Image Semantic" in the figure. b) Decoupled Semantic Assignment. c) Artifact and Background Filtering. d) Heritage-3DGS. Panels b and c show successive stages of the proposed extraction method. Columns show corresponding architectural views, and colors identify component categories according to the legend. White boxes in panel c highlight local regions of reduced semantic quality.



Figure 23. Full-scene appearance renderings produced by different methods. a) Ground-truth photographs. b) Original three-dimensional Gaussian splatting (3DGS). c) Heritage-3DGS. d) Our method. Columns show corresponding scene regions. White rectangles identify locations for comparing local color and shape. GT denotes ground truth.

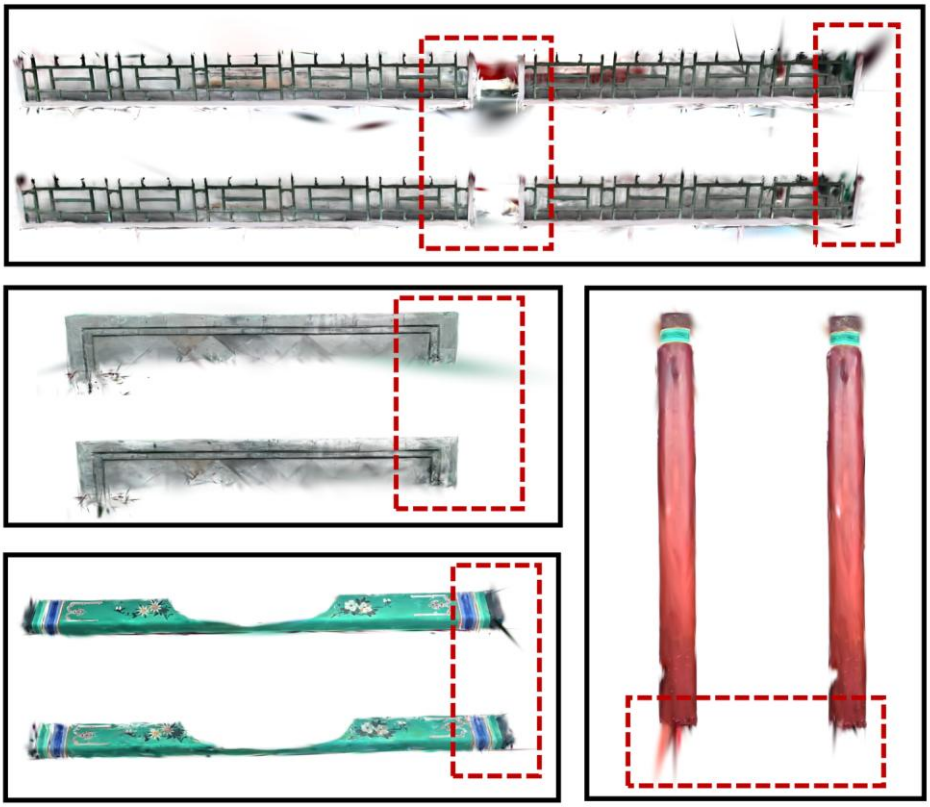


Figure 24. Component renderings before and after boundary scale optimization. Each black-bordered region contains paired renderings of the same component before and after optimization. Red dashed rectangles identify comparison regions containing component boundaries, spikes and residual artifacts.

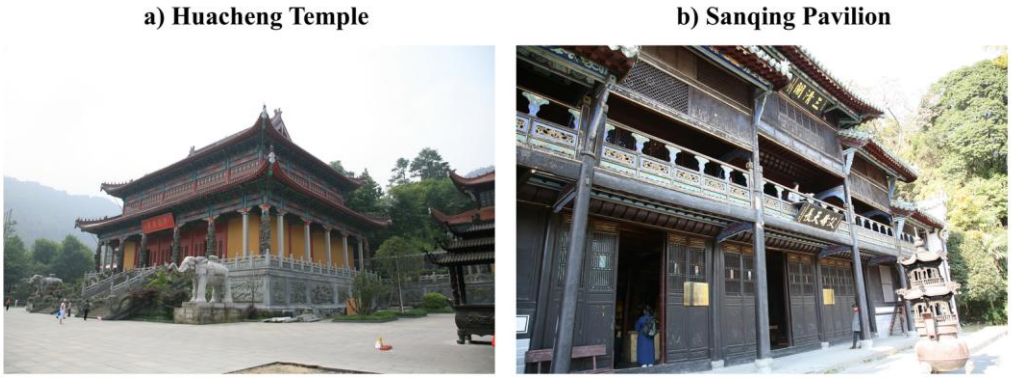


Figure 25. Buildings used for the additional case studies. a) Huacheng Temple. b) Sanqing Pavilion. The photographs show the overall architectural forms, eave levels and component arrangements. Source images: Wei Gao, Chinese Ancient Architecture Image Database [56].

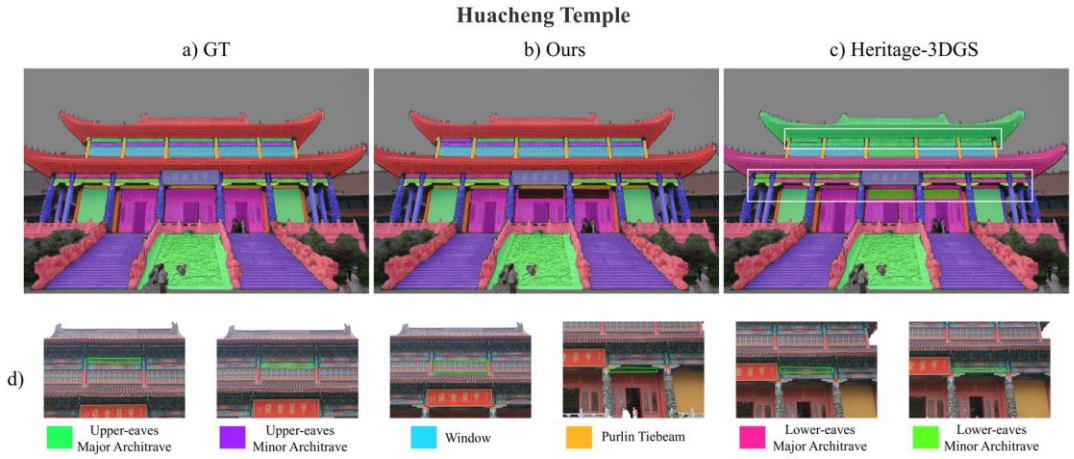


Figure 26. Image-level component semantic segmentation for Huacheng Temple. a) Ground truth. b) Our method. c) Heritage-3DGS. d) Local component examples. Colored masks indicate component categories, with the color swatches and labels identifying architraves at different eave levels, windows and the other illustrated components. GT denotes ground truth. Source images: Wei Gao, Chinese Ancient Architecture Image Database [56].

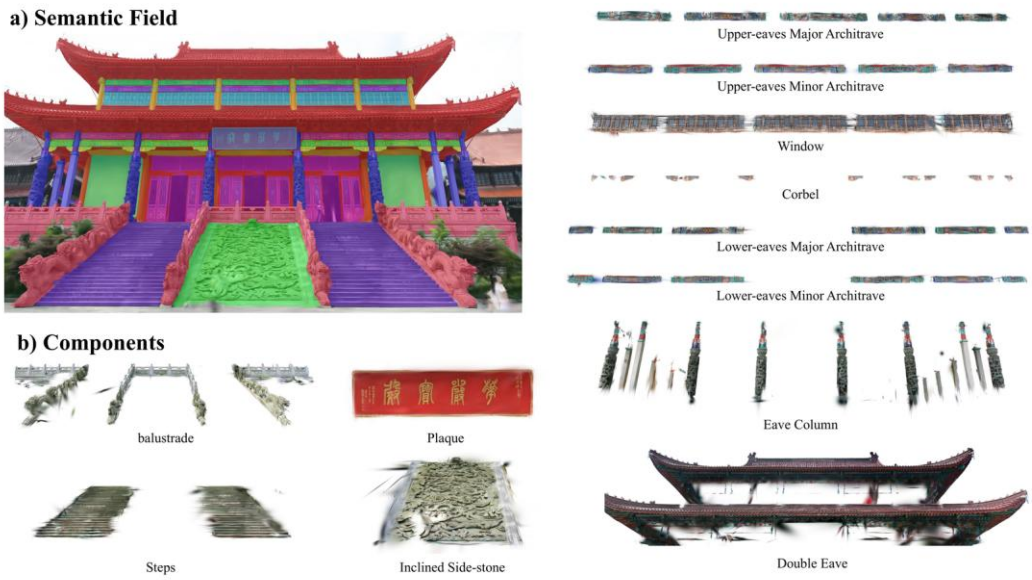


Figure 27. Gaussian semantic field and extracted components of Huacheng Temple. a) The full-building semantic field. b) Independently extracted components, including the labeled categories shown on the right and below. Colors distinguish categories in the semantic field, while independent renderings display the shapes and textures of the extracted components.



Figure 28. Component semantic recognition for Sanqing Pavilion. a) Our method. b) Heritage-3DGS. c) Local component examples. Color swatches and labels identify components at different eave levels and the illustrated window and wall categories. Matching colors in the images indicate the corresponding semantic categories.

Tables

Table 1 Structured Representation of the Spatial Relationship Graph

ID	component	Node	Edge
0	Major Architrave	Thick rectangular beam, Suzhou-style polychrome painting. Geometric border lines and floral-beast motifs, confined within its own tier.	[Bottom] flush against Frieze Board; [Ends] mortised into Eave Column, or adjoins Embracing Tie-beam or Baofu Pattern; [Top] supports Single Eave.
1	Frieze Board	Thin flat panel, uniform vermilion paint. Geometric border lines, no relief.	[Between] Major Architrave and Minor Architrave, flush against both; [Ends] adjoins Eave Column, or Embracing Tie-beam or Baofu Pattern.
2	Minor Architrave	Slender flat beam, alternating blue-green and dark-blue segmented painting.	[Top] flush against Frieze Board; [Bottom] supports Hanging Lattice Fascia; [Ends] mortised into Eave Column.
3	Eave Column	Tall regular cylindrical column, vermilion lacquer with blue-green ring bands at the head.	[Top] mortised into Minor Architrave or Embracing Tie-beam; [Upper shaft] may expose Penetrating Tie-beam end; [Base] stands on Platform; [Sides] adjoins Corbel, Hanging Lattice Fascia, or Bench-under Lattice Fascia and Bench Seat; [Occlusion] in 2D images vertically occludes Inner Column, cutting continuity of Lattice Door, Window, or Sill Wall.
4	Inner Column	Tall regular cylindrical column, smooth vermilion lacquer without painted decoration.	[Top] adjoins Penetrating Tie-beam; [Base] stands on Platform; [Sides] mortised orthogonally against Lattice Door, Window, or Sill Wall; [Occlusion] often hidden behind foreground Eave Column, Corbel, or Hanging Lattice Fascia.
5	Penetrating Tie-beam	Horizontal rectangular tie beam extending in depth, blue-green or dark-blue Suzhou-style painting, exposed end framed by regular border lines.	[Mid-upper shaft] penetrates Eave Column, extends inward to Inner Column; [Below] sits directly beneath Embracing Tie-beam; [Occlusion] main body often occluded by Eave Column, Corbel, Hanging Lattice Fascia, and the architrave system Major Architrave, Frieze Board, Minor Architrave, or Baofu Pattern.
6	Embracing Tie-beam	Square-sectioned projecting beam, end face bears a white-ground Baogu cartouche painting, blue-green and dark-blue background.	[Bottom] rests on Eave Column head and Penetrating Tie-beam end; [Top] supports Single Eave; [Sides] connects to Major Architrave, Frieze Board, or Gable Wall.
7	Corbel	Right-triangular winged wooden bracket, curved lower edge, openwork carving, blue-green or dark-blue lacquer with gilded edges.	[Top] flush against Hanging Lattice Fascia; [Inner side] fixed to Eave Column shaft; [Occlusion] in 2D images often occludes Inner Column and Penetrating Tie-beam behind it.
8	Hanging Lattice Fascia	Slender flat openwork lattice screen, orthogonal timber strips forming a repeating high-frequency grid.	[Top] hung beneath Minor Architrave; [Ends] flanked by Eave Column; [Lower corners] adjoins Corbel; [Occlusion] in 2D images often fragments Inner Column, Lattice Door, or Window in the background.
9	Bench-under Lattice Fascia	Low elongated rectangular openwork guard panel, orthogonal timber strips, lotus-leaf blocks at the base.	[Top] flush against Bench Seat; [Base] lotus-leaf blocks stand on Platform; [Ends] mortised into Eave Column base; [Occlusion] in 2D images often occludes the base of Sill Wall behind it.
10	Bench Seat	Horizontal elongated solid timber board, uniform red or green lacquer, opaque.	[Bottom] rests on Bench-under Lattice Fascia; [Ends] slotted into Eave Column; [Occlusion] in 2D images often horizontally cuts across Inner Column, Lattice Door, Window, or Sill Wall behind it.

ID	component	Node	Edge
11	Lattice Door	Multiple vertical timber door leaves in a row, openwork lattice grid above, relief-carved enclosed panels below.	[Base] stands on Platform; [Sides] flanked by Inner Column; [Upper half] fitted with Window Lattice; [Occlusion] frequently cut by foreground Eave Column, Corbel, or Bench Seat.
12	Window	Large rectangular timber window, heavy vermilion outer frame, openwork strip infill.	[Bottom] flush against Sill Wall; [Sides] flanked by Inner Column or Frame; [Upper half] fitted with Window Lattice; [Occlusion] frequently cut by foreground Eave Column, Hanging Lattice Fascia, or Bench Seat.
13	Sill Wall	Low vertical rectangular masonry wall, blue-grey brick coursing.	[Top] supports Window or Frame; [Ends] mortised into Inner Column; [Base] stands on Platform; [Occlusion] often occluded by Bench Seat and Bench-under Lattice Fascia in front.
14	Window Lattice	Regular translucent grid of orthogonal timber strips, red or blue-green lacquer, decorative joints at intersections.	[Edges] nested within the frame of Lattice Door, Window, or Frame; [Occlusion] shares occlusion with its parent component; upper regions especially prone to fragmentation by Hanging Lattice Fascia or Corbel in front.
15	Platform	Stepped stone base at the bottom of the building, coarse blue-grey natural stone texture.	[Top] supports Eave Column, Inner Column, Lattice Door, Sill Wall, and Bench-under Lattice Fascia; [Front edge] connects horizontally to Steps; [Side edges] connects to Inclined Side-stone.
16	Steps	Stepped stone stairway receding upward in tiers, pale grey stone texture with anti-slip striations.	[Top] connects horizontally to Platform surface; [Sides] flush against Inclined Side-stone.
17	Inclined Side-stone	Diagonally laid solid stone ramp slope, coarse pale grey stone with sanded texture.	[Top] connects to Platform; [Inner face] flush against Steps section; [Bottom] extends down to ground level.
18	Gable Wall	Tall solid enclosure wall, uniform blue-grey or dark-grey brick and stone, staggered horizontal joints.	[Base] stands on Platform; [Inner side] runs parallel and flush against Eave Column; [Occlusion] edges often occlude Major Architrave and Embracing Tie-beam in 2D images.
19	Baofu Pattern	Suzhou-style polychrome decorative painting, richly layered color, sole distinguishing trait is a continuous framed cartouche composition spanning three tiers as one unbroken image.	[Spans] Major Architrave, Frieze Board, and Minor Architrave simultaneously, positioned between two Eave Column heads; [Bottom] adjoins Hanging Lattice Fascia; [Top] adjoins Single Eave.
20	Single Eave	Single-tier projecting roof-eave system of rafters, eave tiles, and drip tiles, blue-green or dark-blue geometric painted rafter ends.	[Bottom] covers Major Architrave, Embracing Tie-beam, and Baofu Pattern; positioned at the top of the building with no component above it.
21	Frame	Rectangular solid timber frame, uniform vermilion lacquer, smooth without openwork.	[Outer edge] adjoins Inner Column; [Inner edge] encloses Window Lattice or Window; [Bottom sill] adjoins Sill Wall; [Occlusion] shares occlusion with Lattice Door or Window as a whole.

Table 2 Component Grouping Results

Bidirectional Grouping Result	Merged Result
1.Major Architrave、Minor Architrave、Frieze Board、Baofu Pattern、Single Eave. 2.Sill Wall、Bench Seat、Bench-under Lattice Fascia. 3.Platform、Inclined Side-stone、Steps、Gable Wall. 4.Window、Window Lattice、Frame、Lattice Door. 5.Inner Column、Eave Column. 6.Embracing Tie-beam. 7.Penetrating Tie-beam. 8.Hanging Lattice Fascia. 9.Corbel.	1.Major Architrave、Minor Architrave、Frieze Board、Baofu Pattern、Single Eave、Embracing Tie-beam、Penetrating Tie-beam. 2.Sill Wall、Bench Seat、Bench-under Lattice Fascia. 3.Platform、Inclined Side-stone、Steps、Gable Wall. 4.Window、Window Lattice、Frame、Lattice Door、Hanging Lattice Fascia. 5.Eave Column、Inner Column. 6.Corbel

Table 3 Sensitivity of Component Extraction to the Filtering Threshold

ϕ	mIoU↑	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
0.1	62.10	23.03	0.9467	0.1013
0.3	65.89	23.57	0.9498	0.0945
0.5	66.71	23.77	0.9509	0.0917
0.7	67.81	24.15	0.9518	0.0897
0.9	**69.24**	**24.30**	**0.9530**	**0.0871**

Table 4 Overall Scene Rendering Quality Comparison

Method	PSNR↑	SSIM↑	LPIPS↓
3DGS	22.95	0.83	0.25
Heritage-3DGS	22.98	0.79	0.27
Ours (semantic assignment and filtering)	22.95	0.83	0.25

Table 5 Rendering and Segmentation Quality Comparison after Boundary Optimization

Component	PSNR ↑	SSIM ↑	LPIPS ↓	IoU ↑ After filtering	IoU ↑ After boundary optimization
Major Arch.	23.36	0.9712	0.0619	65.06	65.89
Frieze Bo.	26.89	0.9847	0.0378	63.12	64.16
Minor Arch.	26.79	0.9760	0.0437	92.96	92.95
Eave Col.	27.42	0.9808	0.0601	88.64	91.95
Inner Col.	24.01	0.9768	0.0507	64.45	68.18
Penetrating.	28.24	0.9910	0.0180	61.64	61.38
Embracing.	31.99	0.9925	0.0166	71.83	88.43
Corbel	31.94	0.9942	0.0112	82.30	81.61
Hanging La.	24.81	0.9631	0.0420	92.71	92.74
Bench-un.	23.31	0.9602	0.0547	49.04	75.89
Bench Seat	26.12	0.9848	0.0284	59.67	55.42
Lattice Door	19.34	0.9447	0.0979	63.03	63.90
Window	19.86	0.9488	0.0845	52.94	65.97
Sill Wall	25.48	0.9781	0.0432	66.38	66.80
Window La.	18.02	0.8790	0.1637	46.86	66.11
Platform	21.23	0.9003	0.1717	77.82	89.54
Steps	33.51	0.8723	0.2118	52.72	71.25
Inclined Si.	35.18	0.9744	0.0417	73.69	77.01
Gable Wall	25.16	0.9672	0.0620	82.59	88.10
Baofu Pat.	29.55	0.9822	0.0244	95.20	96.67
Single Eave	22.79	0.9435	0.1008	68.56	93.15
Frame	15.46	0.8824	0.2222	52.12	52.72
Mean	25.48	0.9567	0.0750	69.24	75.90

Table 6 Sensitivity to the Filtering Threshold after Boundary Gaussian Scale Optimization

ϕ_i	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
0.0	22.86	0.9490	0.0987
0.1	**25.48**	**0.9567**	0.0750
0.3	25.35	0.9564	0.0743
0.5	25.32	0.9564	0.0738
0.7	25.28	0.9562	0.0736
0.9	25.20	0.9558	**0.0735**

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf