Zero-Shot Architectural Style Classification with Large Language Models

Lei Li[1][0009−0006−9844−1753], Yixin Hao[2][0009−0009−2715−3035], and Runjia Tian*[3][0000−0002−5983−9754]

[1] The University of Hong Kong, HK, China
lileiaad@connect.hku.hk
[2] Central Academy Of Fine Arts, China
emilyhao1014@gmail.com
[3] Generative Game Inc., USA
runjiatian@gmail.com

Abstract. In architectural research, accurate classification of architectural styles is crucial for integrating historical and contemporary design elements as well as for educational purposes. This paper presents a novel application of existing Large Language Models (LLMs), including the GPT-4o model from OpenAI, for architectural style classification tasks. LLMs are language neural networks with billions of parameters trained on extensive datasets. GPT-4o, a state-of-the-art LLM, demonstrates significant potential for achieving Zero-Shot classification capabilities, comparable to specifically trained supervised classifiers on many tasks. We explore the model's Zero-Shot ability to identify and classify architectural styles without prior specialized training on architectural images. Various prompt engineering approaches were experimented with to augment the model's prior knowledge about architecture, thereby increasing classification accuracy. Experimental results indicate that GPT-4o can achieve classification accuracy comparable to, and even exceeding, traditional convolutional neural network (CNN) models without any training. This study not only highlights the flexibility and potential of large pre-trained models for architectural style classification but also proposes a new paradigm for resource-efficient and high-accuracy architectural analysis.

Keywords: Digital Naturalism · Scale S · Large Language Model · Architectural Style Classification · Zero-Shot Learning

1 Introduction

In architecture and urban design research, the ability to accurately classify architectural styles from images is critical. It aids architects in blending historical and modern designs and helps scholars study architectural forms and cultural connotations [8]. Moreover, it is essential for public education and understanding of architectural knowledge. The reliability and accuracy of architectural style

* Corresponding author

classification are crucial for the advancement of architectural disciplines and public learning.[4]

However, architectural styles classification is a challenging task due to the complex genres and characteristics of architectural details. Unlike generic object recognition (e.g., classifying everyday items or medical images), architectural style classification poses unique challenges. Architectural styles are defined by subtle visual cues, historical context, and overlapping design elements. [14] Identifying common features within a style and distinguishing between different styles remain significant challenges[27]. Traditional machine learning approaches, like classification based on discriminative models (DPM), often fall short in accuracy[24]. Although deep learning techniques, such as convolutional neural networks (VGG [13], Inception [15], ResNet [6], GoogLeNet [16]), have improved architectural image classification, they require large, annotated datasets and substantial computational resources[14], and they do not fully capture the intuitive and flexible nature of human architectural cognition[5].

To address these limitations, research has shifted towards new technologies like transfer learning [20] and self-supervised learning [3]. These aim to enhance the efficiency and accuracy of image classification, particularly in small or zero-shot learning scenarios[23]. However, these methods still struggle with complex architectural images[11]. The rapid development of AI, especially with LLMs like GPT-4o, Qwen-VL, and Claude3, offers new possibilities[18]. These models, with their extensive training data and exceptional generalization capabilities, show substantial potential in architectural style classification.

LLMs excel in understanding complex features, adapting, and analyzing rich details in images, enhancing classification accuracy and contextual understanding[9,18]. This paper focuses on applying the GPT-4o model to architectural style classification. Without specific training, GPT-4o uses its extensive pre-trained data to analyze architectural images through simple and precise prompts, surpassing traditional models' limitations. Our research aims to improve GPT-4o's performance in architectural style recognition by in-context learning and prompt engineering. Preliminary experiments indicate that the GPT-4o model, guided by structured prompts, can match or exceed the classification accuracy of traditional CNN models [17] under zero-shot conditions[26,21]. This demonstrates the broad potential and flexibility of LLMs in architecture, signaling a new era of resource-efficient and widely applicable architectural image analysis.

2 State-of-the-Art

2.1 Architectural Style Classification

Over the past decade, architectural style classification has evolved from handcrafted feature engineering to deep learning–based methods. Early work by Xu et al. (2014) built a large-scale dataset (scraped from Wikimedia Commons) and employed a Deformable Part Model (DPM) plus latent logistic regression to detect components (e.g., doors, windows) and infer style labels probabilistically[24].

While this approach was foundational and forward-looking, its accuracy remained limited, particularly for visually overlapping styles.

With the rise of deep learning, researchers adopted CNNs (e.g., VGG, ResNet) to automate feature extraction, significantly boosting classification performance [24] . Llamas et al.(2017) found deep features particularly effective for heritage images[8], while Yoshimura et al.(2019) categorized contemporary designs via fine-tuned CNNs [25]. However, these models typically require large labeled datasets and may struggle with styles demanding contextual knowledge (e.g., distinguishing Renaissance vs Neoclassical by construction era).

To address data and complexity issues, intermediate strategies emerged. Chen et al. (2021) proposed hierarchical multi-label classification, recognizing that buildings can exhibit multiple style traits[1]. Xia et al. (2020) classified residential styles by clustering formal elements (like roofs, windows) into fixed combinations[22]. Additionally, attention mechanisms (Wang et al. 2023) have helped localize subtle style indicators such as Gothic arches[17].

Despite these advances, architectural styles can overlap and evolve, complicating rigid classification boundaries. Recent studies[7,26] introduced few-shot and zero-shot approaches, acknowledging the impracticality of labeling every possible style. Our approach aligns with this direction by eliminating the need for dedicated training through a powerful pre-trained model.

2.2 Zero-Shot Visual Recognition and Visual Language Models

Zero-shot learning (ZSL) enables classification of categories unseen during training by relying on auxiliary semantic information (e.g., textual descriptions, word embeddings). Traditional vision models struggle with truly novel classes because they learn direct image-to-label mappings. By contrast, ZSL methods transfer knowledge from known classes to unseen ones, often via textual or ontological representations [19].

Large language models (LLMs) and vision-language models (VLMs) have significantly advanced ZSL performance. Their training in massive datasets, which span both text and images, gives them a broad world knowledge and adaptable feature representations [21]. Chen et al. (2023), for example, demonstrated how a vision-enabled LLM could interpret chest X-rays for COVID-19 diagnosis in a few-shot setting, underscoring its cross-domain inference capacity[2]. Similarly, Wu et al. (2023) benchmarked a vision-language model on tasks including scene recognition and texture classification, confirming robust zero-shot capability[21].

In architecture, Zeng et al. (2024) applied such a model to estimate building ages from façade images in a zero-shot manner. By linking visual clues (ornamental details, structural design) with historical knowledge, the model inferred approximate construction periods without style-specific pre-training[26]. Inspired by these successes, we adopt a similar paradigm for architectural style classification. Rather than training a CNN on thousands of labeled examples[24], we rely on a multi-modal large model's existing "knowledge" and prompt it to identify styles. This approach aligns with the broader trend of leveraging LLM/VLM flexibility for specialized tasks that lack exhaustive training datasets.

In summary, while architectural style classification has seen progress through CNN-based pipelines and element-based clustering[17,12,24,8], zero-shot learning with vision-language models offers a new path[26,21]. By uniting visual analysis and contextual knowledge, these models reduce dependency on labeled data and can recognize nuanced style features that might be overlooked by purely supervised methods[10]. Our work builds upon these insights and aims to demonstrate the feasibility of zero-shot style identification in a domain where styles are inherently multi-dimensional and often under-labeled.

3 Methods

3.1 Datasets and Ethical

We evaluated our approach on the architectural style image dataset (Fig. 1) originally constructed by Xu et al[24]. This dataset was collected from the "Architecture by style" entries on Wikipedia (Wikimedia Commons), using a systematic crawl to gather photographs of building exteriors for a wide range of styles[24]. It covers both well-known styles (e.g., Gothic, Baroque, Modernist) and more region-specific or era-specific styles (e.g., Achaemenid, Russian Revival). Each style in the dataset contains roughly 60 to 400 images, with a total of nearly 5,000 images across all categories. All images are in color with roughly 800×600 resolution, capturing buildings such as churches, houses, museums, and other structures, each labeled by architectural style.

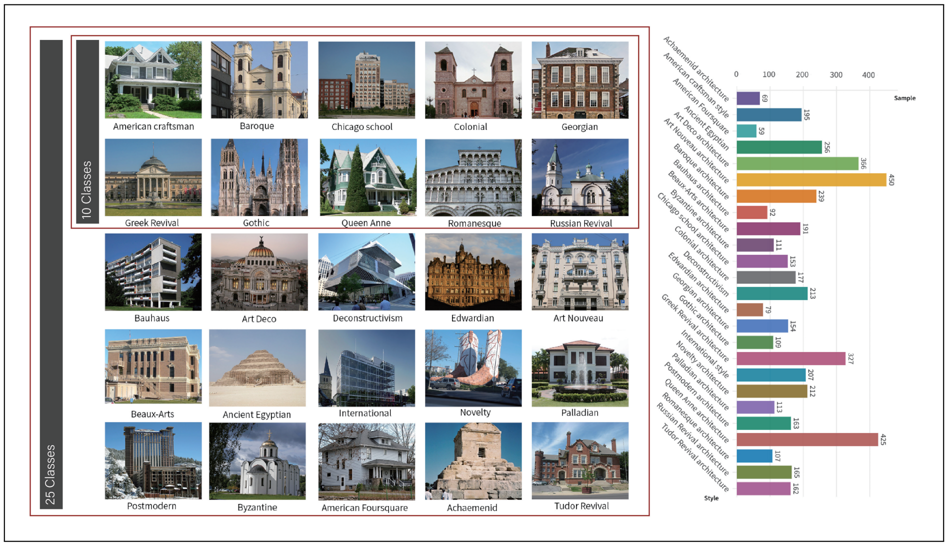


Fig. 1. Features of the dataset.

For our experiments, we considered two classification scenarios: 10-class and 25-class subsets task (see Fig. 1). In the 10-class task, we use a selection of ten styles that are relatively easier to distinguish visually (American Craftsman, Baroque, Chicago School, Colonial, Georgian, Gothic, Greek Revival, Queen Anne, Romanesque, Russian Revival). The 25-class task uses a much broader set of styles (essentially the full dataset, comprising 25 style categories including the above and others such as Art Deco, Art Nouveau, International, Tudor Revival, etc.). The 25-class scenario is significantly more challenging, as it contains stylistic nuances and several styles that have overlapping features.

We also acknowledge several ethical considerations regarding the dataset and model behavior. This dataset is sourced from Wikipedia Commons under open/fair-use licenses, used solely for research. However, it may overrepresent European and North American architecture, leaving other regions underrepresented. GPT-4o, trained mostly on Western data, could thus favor well-documented styles. While style classification is benign in nature, it can still reflect cultural narratives (e.g., defining “canonical” Gothic). We highlight where the model struggles and encourage diverse data to mitigate these biases.

3.2 Overall Experiment Design

Our experimental framework leverages GPT-4o’s multimodal (vision+language) capabilities through strategic prompting. The goal is to assess and improve the zero-shot classification of architectural styles by providing the model with additional context in various ways. We devised four methods (see Fig. 2) that progressively increase the amount of guidance given to GPT-4o when it analyzes an image. Below we describe each method and the rationale behind choosing these four:

Table 1. Comparison of four prompt engineering methods.

Method	Step 1: Image Analysis	Step 2: Classification Prompt	Outputs
1. Direct	None; directly use image	Ask GPT-4o to classify style from the image directly	Style label
2. Tags-Only	Generate descriptive tags from image (via GPT-4o)	Provide tags + style hints; ask for best matching style	Style label
3. Pic+Tags	Generate tags from image	Provide both image and tags + style descriptions; ask for classification	Style label
4. Pic+Tags +Reason	Generate tags from image	Provide image + tags + detailed style info; ask for style and explanation	Style label + Reason

– **Method 1 (GPT4o-Vision-DIRECT):** This is the most straightforward baseline. We present the architectural image to GPT-4o with a simple prompt asking for the architectural style. Essentially, GPT-4o is directly tasked with

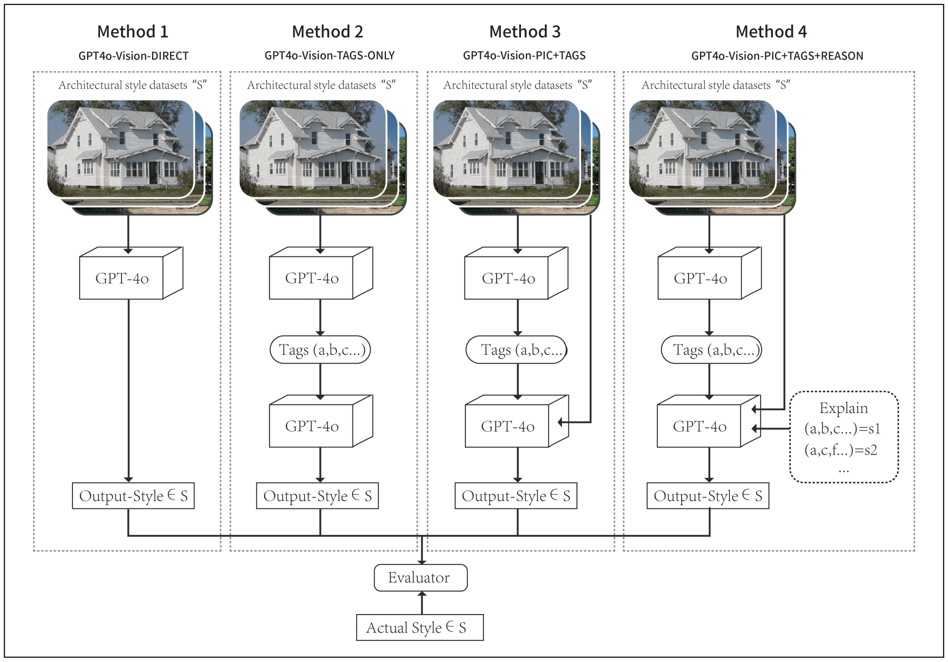
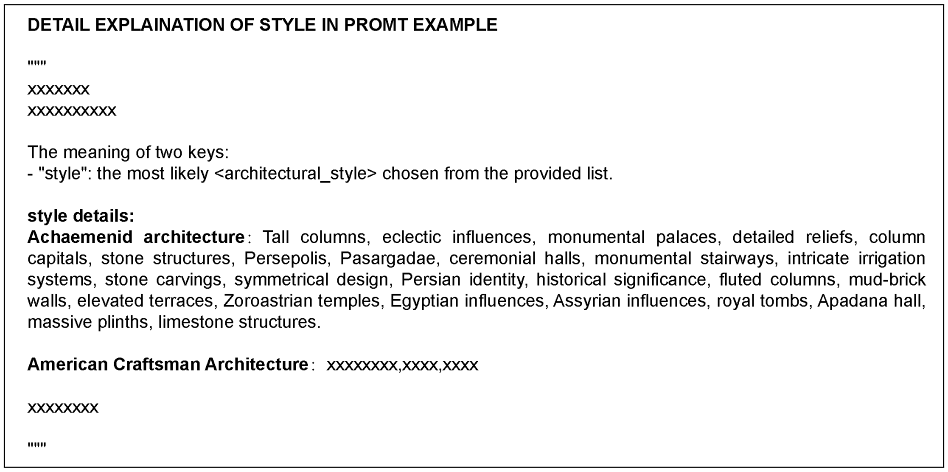


Fig. 2. Framework for zero-shot architectural style classification based on GPT-4o.

labeling the image's style in one step. No extra context or information is given beyond the image itself. This tests the model's raw zero-shot ability using its internal knowledge.

- **Method 2 (GPT4o-Vision-TAGS-ONLY):** In this method, we introduce an intermediate step of feature extraction. GPT-4o first generates concise architectural tags from the image (e.g., "pointed arch, flying buttress, stained glass"), representing key stylistic elements. These tags are then used—without the image—as input to a second GPT-4o prompt that compares them against style descriptions to infer the most likely style. This two-step approach (Image → Tags → Style) aims to enhance classification by focusing the model's reasoning on explicit architectural features.
- **Method 3 (GPT4o-Vision-PIC+TAGS):** This method combines the image and the previously generated tags as joint input to GPT-4o, along with style descriptions. The goal is to let the model leverage both raw visual input and distilled feature cues. Tags serve to direct attention toward key elements, while the image allows visual verification. We test whether this multimodal setup improves classification over using tags alone.
- **Method 4 (GPT4o-Vision-PIC+TAGS+REASON):** Building on Method 3, this approach adds a requirement for reasoning. GPT-4o receives the image, tags, and style descriptions (see Fig. 3), and is prompted to justify its classification with a brief explanation. This encourages "chain-of-thought"

reasoning and enhances interpretability. Even if accuracy gains are minimal, the added rationale offers insight and confidence in model predictions.



DETAIL EXPLAINATION OF STYLE IN PROMT EXAMPLE

"""
xxxxxxx
xxxxxxxxxx

The meaning of two keys:
- "style": the most likely <architectural_style> chosen from the provided list.

style details:
Achaemenid architecture: Tall columns, eclectic influences, monumental palaces, detailed reliefs, column capitals, stone structures, Persepolis, Pasargadae, ceremonial halls, monumental stairways, intricate irrigation systems, stone carvings, symmetrical design, Persian identity, historical significance, fluted columns, mud-brick walls, elevated terraces, Zoroastrian temples, Egyptian influences, Assyrian influences, royal tombs, Apadana hall, massive plinths, limestone structures.

American Craftsman Architecture: xxxxxxxx,xxxx,xxxx

xxxxxxxx

"""

Fig. 3. Detail explanation of style in prompt.

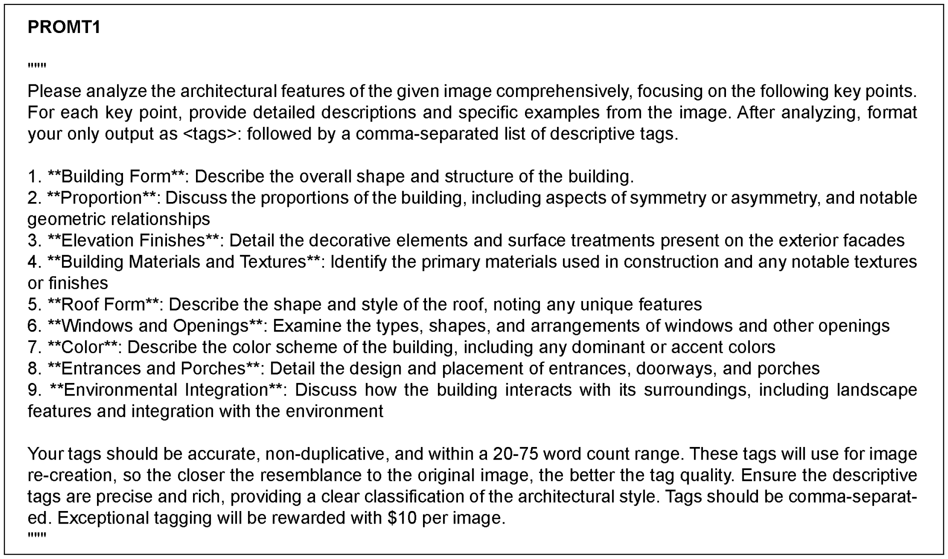
To summarize, these four methods were intentionally designed to evaluate how incremental additions of context affect GPT-4o's zero-shot classification performance. Method 1 serves as a baseline with direct querying. Method 2 adds textual context by translating visual input into tags, testing whether language-based reasoning alone improves results. Method 3 combines image and tags as multimodal input to assess whether cross-verification enhances accuracy. Method 4 introduces explanatory reasoning to test model consistency and interpretability. Together, these methods reveal the trade-offs between simplicity, predictive performance, and transparency

3.3 Prompt Engineering Workflow

To implement Methods 2–4, we designed a two-stage prompt workflow, using two query prompts Q1 and Q2 in sequence for each image (for Method 1, only one prompt is needed). We formalize the process as follows:

Stage 1 (Q1 – Tag Extraction): We input the image I to GPT-4o with an initial prompt Q1 (Fig. 4)that instructs the model to analyze the image and output a set of descriptive tags or short phrases capturing key architectural features. Formally, we denote this as:

$$\mathbf{T} = \mathbf{g}_{\textbf{GPT-4o}}(\mathbf{I};\ \mathbf{Q}_1)$$



PROMT1

"""
Please analyze the architectural features of the given image comprehensively, focusing on the following key points. For each key point, provide detailed descriptions and specific examples from the image. After analyzing, format your only output as <tags>: followed by a comma-separated list of descriptive tags.

1. **Building Form**: Describe the overall shape and structure of the building.
2. **Proportion**: Discuss the proportions of the building, including aspects of symmetry or asymmetry, and notable geometric relationships
3. **Elevation Finishes**: Detail the decorative elements and surface treatments present on the exterior facades
4. **Building Materials and Textures**: Identify the primary materials used in construction and any notable textures or finishes
5. **Roof Form**: Describe the shape and style of the roof, noting any unique features
6. **Windows and Openings**: Examine the types, shapes, and arrangements of windows and other openings
7. **Color**: Describe the color scheme of the building, including any dominant or accent colors
8. **Entrances and Porches**: Detail the design and placement of entrances, doorways, and porches
9. **Environmental Integration**: Discuss how the building interacts with its surroundings, including landscape features and integration with the environment

Your tags should be accurate, non-duplicative, and within a 20-75 word count range. These tags will use for image re-creation, so the closer the resemblance to the original image, the better the tag quality. Ensure the descriptive tags are precise and rich, providing a clear classification of the architectural style. Tags should be comma-separated. Exceptional tagging will be rewarded with $10 per image.
"""

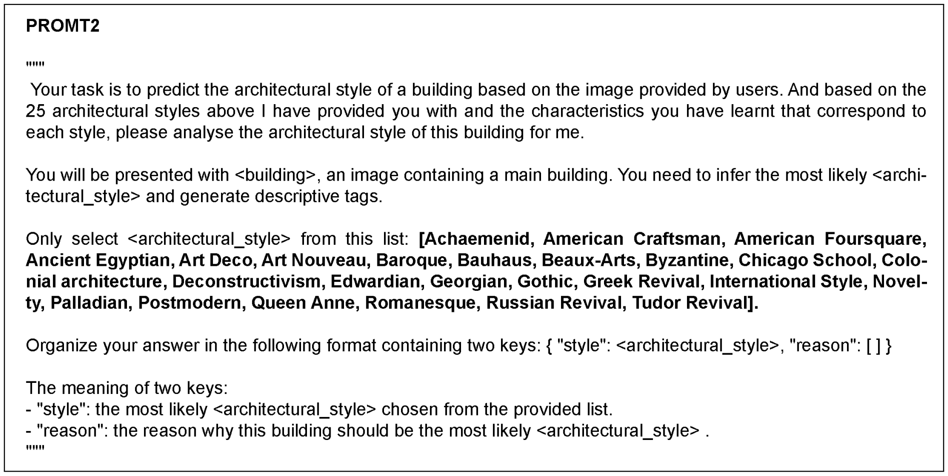
Fig. 4. Q1: Prompt of analyzing tags.

Here, $g_{\text{GPT-4o}}(I; Q1)$ represents GPT-4o's operation on image I with prompt $Q1$, producing T, the set of tags. The prompt Q1 is carefully worded to ensure the tags are relevant to architectural style (e.g., "List distinctive architectural features you observe, in a few words each."). This stage leverages GPT-4o's deep semantic understanding of the image, translating visual details into textual descriptors. (Fig. 4) shows an example of such Q1 and its tag output. In our running example, I (a cathedral image) might yield $T =$ ["pointed arch", "flying buttress", "stained glass", "tall spire"].

Stage 2 (Q2 – Style Reasoning): Next, we craft a follow-up prompt Q2 (Fig. 5) that provides context about architectural styles and asks GPT-4o to identify the style given the information from Stage 1. We feed GPT-4o with T (the tags from Stage 1) and, depending on the method, optionally the image I again, plus explanatory text about possible styles. In general, we denote this reasoning stage as:

$$(\mathbf{S}^\wedge, \mathbf{R}^\wedge) = \mathbf{F}\left(\mathbf{g}_{\mathbf{GPT-4o}}(\mathbf{I}; \mathbf{Q}_1); \mathbf{Q}_2\right)$$

Where $S^\wedge$ is the predicted style label and $R^\wedge$ is an optional detailed explanation. The function F represents GPT-4o's reasoning process under prompt $Q2$, which may incorporate both the tags T and the image I. For Method 2, F takes only T (no image) with a prompt that lists styles and asks "Given these features, which style fits best?" For Method 3, F takes (T, I) so the model can inspect the image while considering the tags. For Method 4, F takes (T, I) with an expanded prompt that says, for example, "Determine the style and explain



PROMT2

"""

Your task is to predict the architectural style of a building based on the image provided by users. And based on the 25 architectural styles above I have provided you with and the characteristics you have learnt that correspond to each style, please analyse the architectural style of this building for me.

You will be presented with <building>, an image containing a main building. You need to infer the most likely <architectural_style> and generate descriptive tags.

Only select <architectural_style> from this list: **[Achaemenid, American Craftsman, American Foursquare, Ancient Egyptian, Art Deco, Art Nouveau, Baroque, Bauhaus, Beaux-Arts, Byzantine, Chicago School, Colonial architecture, Deconstructivism, Edwardian, Georgian, Gothic, Greek Revival, International Style, Novelty, Palladian, Postmodern, Queen Anne, Romanesque, Russian Revival, Tudor Revival].**

Organize your answer in the following format containing two keys: { "style": <architectural_style>, "reason": [ ] }

The meaning of two keys:
- "style": the most likely <architectural_style> chosen from the provided list.
- "reason": the reason why this building should be the most likely <architectural_style> .
"""

Fig. 5. Q2: Prompt of analyzing style and reason.

your reasoning. Here are some known architectural styles and their characteristics: ... [list]. The image features are: [tags]." The output includes both $S^\wedge$ and $R^\wedge$ (the model's explanation). (Fig. 5) illustrates an example of Q2 for Method 3. In our example, Q2 might contain descriptions of Gothic, Romanesque, Renaissance styles, etc., and GPT-4o would output $S^\wedge$ = Gothic, with $R^\wedge$ = "The building has tall pointed arches, flying buttresses, and large stained-glass windows, which are characteristic of Gothic architecture."

This two-stage prompt engineering allows GPT-4o to flexibly perform multi-step reasoning: first as a visual analyst, then as a domain expert. Importantly, it uses GPT-4o "in context" learning abilities – we do not fine-tune the model, only guide it with carefully chosen textual context.

Having set up the methods, we next describe our experiments and results.

4 Results and Discussion

In this section, we present a comprehensive analysis of our results and discuss their implications. We begin by examining the feature tags generated by our approach to verify how they contribute to classification accuracy and interpretability. We then evaluate our zero-shot architectural style classification framework (based on GPT-4o) in comparison with a broad range of benchmark models, from classical hand-crafted feature methods to state-of-the-art deep learning and large language models. This evaluation encompasses both a 10-class task (a subset of styles) and a more challenging 25-class task (the full style set). We further analyze the outcomes through a confusion matrix to understand which architectural styles are easily recognized and which are frequently confused. Next, we provide a per-style performance breakdown with key metrics (precision, recall, F1-score, accuracy), explaining what each metric indicates in the context of architectural classification.

Finally, we discuss the practical application of the GPT-4o model in a zero-shot setting, highlighting the significance of our tag-assisted methodology and its long-term potential in architectural style recognition.

4.1 Tags Features Analysis

Understanding visual tags extracted from images is crucial for interpreting how the model distinguishes architectural styles. We visualized tag frequency and interrelationships using word clouds and network graphs (Fig. 6). The word cloud displays frequently occurring descriptors in larger font, while the network graph connects commonly co-occurring tags, showing how architectural features correlate. This visualization reveals which features the model "sees" most often and how they group into higher-level concepts.

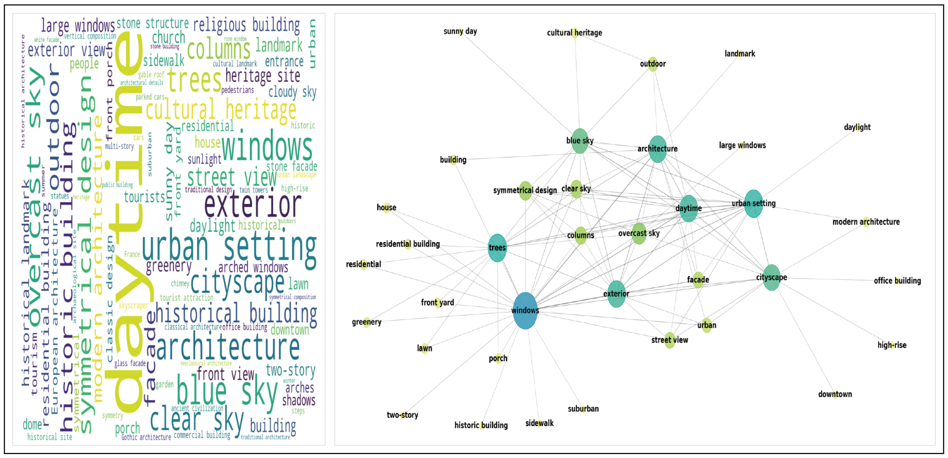


Fig. 6. Tags word clouds and network analysis.

We filtered generic environmental and photographic tags ("daytime," "blue sky," "trees," etc.) to focus on distinctive architectural characteristics. The top ten distinctive tags were: "symmetrical design," "outdoor," "historic building," "columns," "cultural heritage," "facade," "historical building," "street view," "modern architecture," and "religious building." These correspond to salient features associated with particular styles—"columns" and "symmetrical design" often signal classical architecture, while "modern architecture" implies contemporary styles. Tags like "historical building" suggest the model recognizes antiquity cues that correlate with historical styles.

The network graph reveals how architectural features co-occur. Connected tag clusters show feature groups that appear together in images. For example, "trees" links with "residential building," "lawn," and "front yard," indicating residential styles often include greenery settings. Similarly, "urban setting" connects to "modern architecture" and "office building," showing modern styles frequently appear in urban environments. These linkages demonstrate that the model captures both isolated features and contextual associations, helping explain its reasoning: when features consistently appear together, GPT-4o can leverage that information for more informed predictions.

This analysis validates our tag-generation approach. By identifying emphasized visual cues and their relationships, we confirm the model leverages meaningful architectural features rather than random patterns. Understanding feature importance and interconnectedness aids classification—the clearer we discern defining style tags, the more accurately the model interprets new images. The tags serve as an interpretable bridge between raw images and style predictions, providing insights into architectural characteristics and enabling more accurate, explainable interpretations.

4.2 Benchmark Experiments

As seen in (Fig. 7). To evaluate the performance of our zero-shot GPT-4o classifier, we compared it against a diverse set of baseline methods, ranging from traditional hand-crafted features to modern deep learning and large multimodal models. These baselines include: (1) rule-based global descriptors such as GIST and Spatial Pyramid Matching (SP), which use engineered features without learning; (2) object- and mid-level representations like Object Bank (OB) and MLLR+SR, which add semantic structure by encoding object-level patterns; (3) classical supervised classifiers, including enhanced Deformable Part Model variants (e.g., DPM-LSVM, DPM-MLLR, DPM+IEP-SVM), which integrate hand-crafted or intermediate features with training; and (4) deep learning models, such as the Nonlinear Consensus Style Centralized Autoencoder (NCSCAE) and pretrained CNNs like Inception-v3, both with and without architectural attention modules (e.g., CSAM). Finally, we evaluated GPT-4o alongside other state-of-the-art multimodal LLMs like Qwen-VL and Claude-3, both capable of zero-shot image-language classification.

Each method was tested under two classification scenarios: a 10-class task focusing on broader distinctions, and a more fine-grained 25-class task. The results reveal a consistent trend: GPT-4o achieved the highest accuracy in both settings, outperforming all baselines without any additional training. Its performance not only surpassed rule-based and supervised models—which depend heavily on labeled datasets—but also exceeded that of other advanced zero-shot LLMs. This outcome underscores GPT-4o's strong generalization capabilities, driven by large-scale image-text pretraining and its ability to reason contextually about architectural features.

4.3 Benchmark Result Discussion

To clarify performance differences, baseline methods are categorized into three types: (a) rule-based feature extractors (e.g., GIST, SP), (b) supervised classifiers

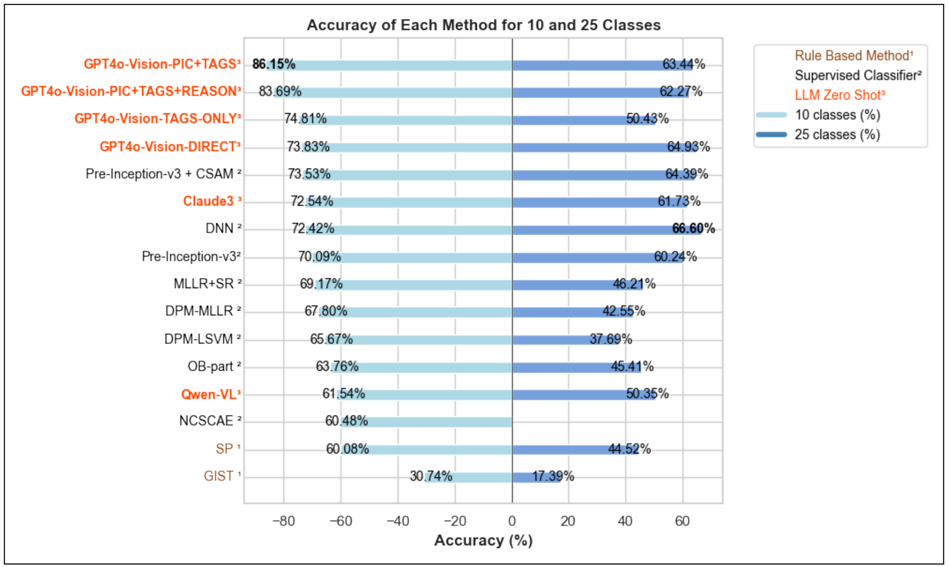


Fig. 7. Accuracy of each method for 10 and 25 classes.

(e.g., DPM variants, Inception-v3), and (c) large multimodal models (e.g., GPT-4o, Claude-3, Qwen-VL), each evaluated on both 10-class and 25-class tasks.

Rule-based methods offered minimal discriminative power, relying on low-level image descriptors that proved insufficient for capturing the nuanced features of architectural styles. Their performance dropped sharply on the 25-class task, confirming their limitations in complex, fine-grained classification.

Supervised learning models, including traditional SVM and CNN-based approaches, achieved stronger results, particularly when using pretrained models like Inception-v3. The inclusion of architectural attention mechanisms (e.g., CSAM) further improved accuracy. However, these methods require extensive labeled datasets and training, limiting scalability and adaptability to new or rare styles.

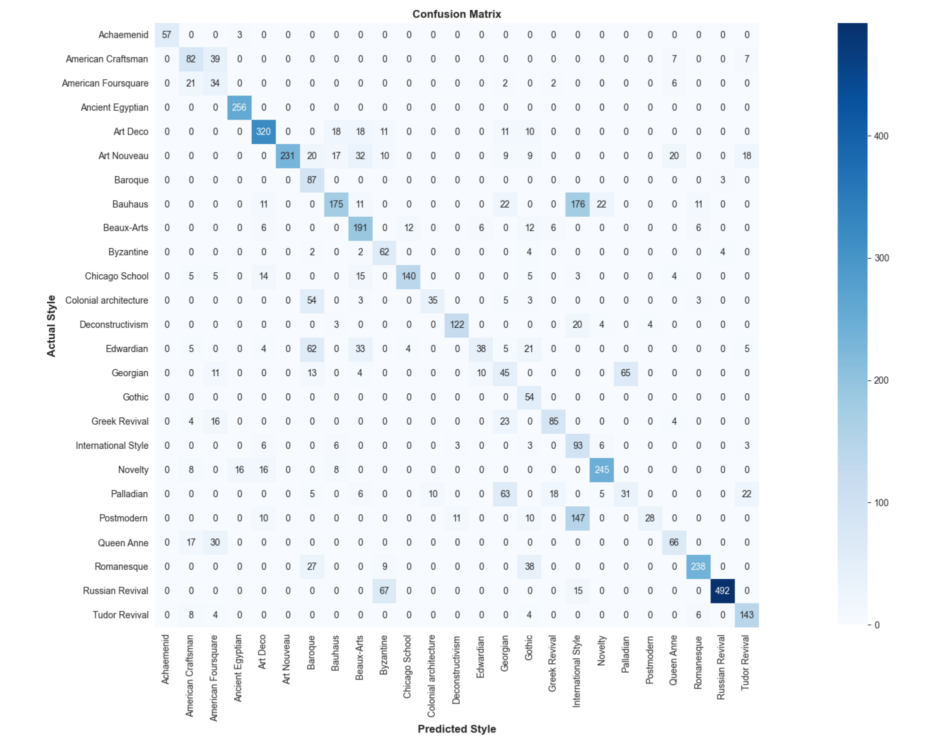
Multimodal LLMs, especially GPT-4o, outperformed all baselines. Operating in zero-shot mode, GPT-4o achieved 86.1% accuracy on the 10-class task and approximately 63.4% on the more challenging 25-class task. This was accomplished without any fine-tuning, demonstrating its strong generalization ability derived from large-scale image-text pretraining. Comparable models like Claude-3 and Qwen-VL also showed solid performance, but GPT-4o had a clear edge, likely due to its broader training corpus and superior contextual reasoning.

Among GPT-4o prompting strategies, the "PIC+TAGS" configuration (image plus feature tags) yielded the highest accuracy. Adding reasoning steps ("PIC+TAGS+REASON") did not improve results significantly, suggesting that interpretability may come at a slight trade-off in precision. Nonetheless, the model's ability to explain its predictions remains a valuable feature.

In summary, GPT-4o not only rivals but often surpasses traditional supervised models, offering a flexible, data-efficient approach to architectural style classification. Its success highlights the strength of combining visual input with language-driven context in complex recognition tasks.

4.4 Confusion Matrix Evaluation

To better understand the model's performance across categories, we analyzed the 25-class confusion matrix (Fig. 8). The matrix shows how often each architectural style was correctly identified (diagonal values) versus misclassified as another (off-diagonal values). Styles with distinct visual identities—such as Ancient Egyptian and Art Deco—were classified with particularly high accuracy. For example, Ancient Egyptian achieved a perfect score with 256/256 correct predictions, and Art Deco similarly reached 320 correct identifications, with minimal confusion. These results demonstrate that the model is highly reliable when styles exhibit clear, iconic features such as monumental stone forms or geometric ornamentation.



	Achaemenid	American Craftsman	American Foursquare	Ancient Egyptian	Art Deco	Art Nouveau	Baroque	Bauhaus	Beaux-Arts	Byzantine	Chicago School	Colonial architecture	Deconstructivism	Edwardian	Georgian	Gothic	Greek Revival	International Style	Novelty	Palladian	Postmodern	Queen Anne	Romanesque	Russian Revival	Tudor Revival
Achaemenid	57	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
American Craftsman	0	82	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	7
American Foursquare	0	21	34	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	6	0	0	0
Ancient Egyptian	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Art Deco	0	0	0	0	320	0	0	18	18	11	0	0	0	0	11	10	0	0	0	0	0	0	0	0	0
Art Nouveau	0	0	0	0	0	231	20	17	32	10	0	0	0	0	9	9	0	0	0	0	0	20	0	0	18
Baroque	0	0	0	0	0	0	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Bauhaus	0	0	0	0	11	0	0	175	11	0	0	0	0	0	22	0	0	176	22	0	0	0	11	0	0
Beaux-Arts	0	0	0	0	6	0	0	0	191	0	12	0	0	6	0	12	6	0	0	0	0	0	6	0	0
Byzantine	0	0	0	0	0	0	2	0	2	62	0	0	0	0	0	4	0	0	0	0	0	0	0	4	0
Chicago School	0	5	5	0	14	0	0	0	15	0	140	0	0	0	0	5	0	3	0	0	0	4	0	0	0
Colonial architecture	0	0	0	0	0	0	54	0	3	0	0	35	0	0	5	3	0	0	0	0	0	0	3	0	0
Deconstructivism	0	0	0	0	0	0	0	3	0	0	0	0	122	0	0	0	0	20	4	0	4	0	0	0	0
Edwardian	0	5	0	0	4	0	62	0	33	0	4	0	0	38	5	21	0	0	0	0	0	0	0	0	5
Georgian	0	0	11	0	0	0	13	0	4	0	0	0	0	10	45	0	0	0	0	65	0	0	0	0	0
Gothic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54	0	0	0	0	0	0	0	0	0
Greek Revival	0	4	16	0	0	0	0	0	0	0	0	0	0	0	23	0	85	0	0	0	0	4	0	0	0
International Style	0	0	0	0	6	0	0	6	0	0	0	0	3	0	0	3	0	93	6	0	0	0	0	0	3
Novelty	0	8	0	16	16	0	0	8	0	0	0	0	0	0	0	0	0	0	245	0	0	0	0	0	0
Palladian	0	0	0	0	0	0	5	0	6	0	0	10	0	0	63	0	18	0	5	31	0	0	0	0	22
Postmodern	0	0	0	0	10	0	0	0	0	0	0	0	11	0	0	10	0	147	0	0	28	0	0	0	0
Queen Anne	0	17	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	0	0	0
Romanesque	0	0	0	0	0	0	27	0	0	9	0	0	0	0	0	38	0	0	0	0	0	0	238	0	0
Russian Revival	0	0	0	0	0	0	0	0	0	67	0	0	0	0	0	0	0	15	0	0	0	0	0	492	0
Tudor Revival	0	8	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	6	0	143

Fig. 8. Confusion Matrix.

However, the matrix also reveals consistent confusion patterns among visually or historically related styles. A notable case is the misclassification of Romanesque buildings as Gothic: 27 Romanesque examples were predicted as Gothic, whereas only 13 Gothic buildings were misclassified in the reverse direction. This reflects their architectural continuity—Romanesque's round arches and heavy massing vs. Gothic's pointed arches and verticality—and indicates the model tends to default to the more elaborate Gothic label when overlap occurs. Similarly, Queen Anne is frequently confused with American Craftsman: out of 140 Queen Anne images, only 66 were correctly identified, while 30 were misclassified as Craftsman. The stylistic proximity of these residential types—both characterized by porches, pitched roofs, and wood detailing—makes accurate classification challenging, especially when ornamental features are subdued. Additional confusion was observed between Beaux-Arts and Edwardian, with 33 Edwardian images misclassified as Beaux-Arts, further highlighting the model's sensitivity to transitions between adjacent formal traditions.

These confusion patterns point to areas for improvement. Most misclassifications occur where styles share key visual elements, temporal overlap, or functional typologies. Enhancing the prompt with finer-grained contextual cues—such as structural systems, material palettes, or decorative richness—could help disambiguate similar styles. Alternatively, secondary prompts or rule-based checks could be introduced to verify classification in known-confusion zones. In summary, while the model performs exceptionally well with stylistically distinct categories, the confusion matrix clearly highlights boundaries where visual similarity leads to ambiguity. Addressing these edge cases is crucial for enhancing the robustness and interpretability of zero-shot architectural style classification.

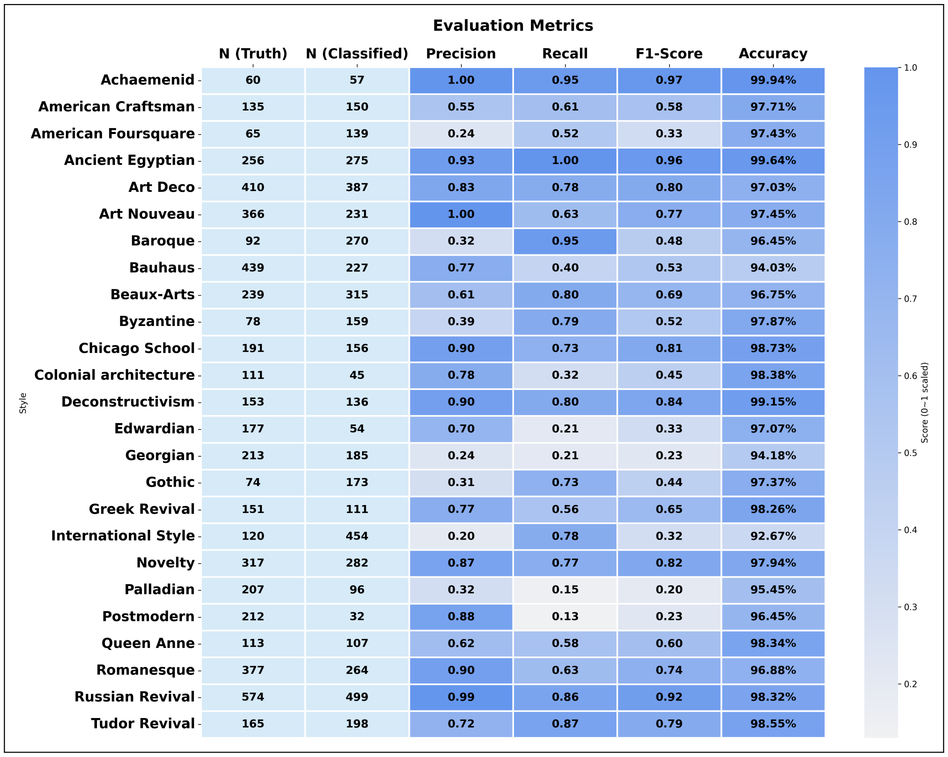
4.5 Analysis for Each Style

To assess classification performance in detail, we computed standard evaluation metrics for 25 architectural style: Precision, Recall, F1-score, and Accuracy (Fig. 9). **Precision** reflects how often a predicted label is correct—that is, among all images the model predicted as a given style, what proportion truly belong to that style. **Recall** shows how often the true class was correctly detected—among all images that actually belong to a style, how many were successfully identified. **F1** provides a balanced measure of both, serving as the harmonic mean of precision and recall.

We also report **Accuracy** as a per-class measure, defined as the percentage of all images (across all styles) that were correctly identified with respect to a specific style—this includes both true positives and true negatives. However, due to the class imbalance in the dataset (i.e., far more images not of a given style than of that style), this metric can appear deceptively high even for classes where the model performs poorly. For this reason, we emphasize precision and recall in our discussion, as they more directly reflect classification behavior for each style.

Two supporting counts further contextualize the metrics: **n(Truth)**, which indicates the number of images that truly belong to a given style, and **N(Classified)**, which represents the number of images that the model predicted as that style, regardless of correctness. These values form the denominators for recall and precision, respectively.

The overall precision of the 25-class task was 63%, a strong result for zero-shot performance in a fine-grained classification task. In the following discussion, we report precision, recall, and F1-score in that order (e.g., Achaemenid: 1.00 / 0.95 / 0.97).



Evaluation Metrics

Style	N (Truth)	N (Classified)	Precision	Recall	F1-Score	Accuracy
Achaemenid	60	57	1.00	0.95	0.97	99.94%
American Craftsman	135	150	0.55	0.61	0.58	97.71%
American Foursquare	65	139	0.24	0.52	0.33	97.43%
Ancient Egyptian	256	275	0.93	1.00	0.96	99.64%
Art Deco	410	387	0.83	0.78	0.80	97.03%
Art Nouveau	366	231	1.00	0.63	0.77	97.45%
Baroque	92	270	0.32	0.95	0.48	96.45%
Bauhaus	439	227	0.77	0.40	0.53	94.03%
Beaux-Arts	239	315	0.61	0.80	0.69	96.75%
Byzantine	78	159	0.39	0.79	0.52	97.87%
Chicago School	191	156	0.90	0.73	0.81	98.73%
Colonial architecture	111	45	0.78	0.32	0.45	98.38%
Deconstructivism	153	136	0.90	0.80	0.84	99.15%
Edwardian	177	54	0.70	0.21	0.33	97.07%
Georgian	213	185	0.24	0.21	0.23	94.18%
Gothic	74	173	0.31	0.73	0.44	97.37%
Greek Revival	151	111	0.77	0.56	0.65	98.26%
International Style	120	454	0.20	0.78	0.32	92.67%
Novelty	317	282	0.87	0.77	0.82	97.94%
Palladian	207	96	0.32	0.15	0.20	95.45%
Postmodern	212	32	0.88	0.13	0.23	96.45%
Queen Anne	113	107	0.62	0.58	0.60	98.34%
Romanesque	377	264	0.90	0.63	0.74	96.88%
Russian Revival	574	499	0.99	0.86	0.92	98.32%
Tudor Revival	165	198	0.72	0.87	0.79	98.55%

Fig. 9. Evaluation Matrix.

The model excels on styles with highly distinctive visual identities. Achaemenid (1.00, 0.95, 0.97), Ancient Egyptian (0.93 / 1.00 / 0.96), and Russian Revival (0.99 / 0.86 / 0.92) showed exceptional metrics, confirming GPT-4o's ability to recognize iconic features like monumental columns, hieroglyphs, or colorful onion domes. Chicago School and Novelty also scored F1 > 0.80, reflecting strong detection of styles with localized or consistent design traits.

By contrast, the model struggled with styles that have less distinctive features or overlap with others. American Foursquare had low precision (0.24) and moderate recall (0.52), suggesting frequent confusion with other domestic styles such as Craftsman or Colonial. Baroque (0.32 / 0.95 / 0.48) and Byzantine (0.39 / 0.79 / 0.52) showed high recall but very low precision, indicating over-prediction—likely due to mislabeling other ornate styles as Baroque or Byzantine. These findings align with the confusion matrix and highlight the challenge of distinguishing richly decorated, historically adjacent styles.

Some styles showed imbalanced performance. Art Nouveau had perfect precision (1.00) but only 0.63 recall—implying the model labeled examples conservatively and missed subtle cases. Beaux-Arts (0.61 / 0.80 / 069) and Edwardian (0.70 / 0.21 / 0.32) exhibited moderate precision-recall tradeoffs, while Georgian had uniformly poor scores (0.24 / 0.21 / 0.23), likely due to visual similarity with other classical residential forms. Postmodern is another edge case: high precision (0.88) but very low recall (0.13), meaning the model rarely predicts it, but is accurate when it does.

These style-level metrics suggest improvement opportunities lie in better distinguishing overlapping or underrepresented styles. For instance, International Style had a recall of 0.78 but only 0.20 precision—often over-predicted. Refining prompts or tags to emphasize unique visual cues (e.g., materiality, ornamentation level, structural logic) may address these gaps. Additionally, pairing GPT-4o's predictions with rule-based disambiguation for high-confusion pairs may further improve performance.

4.6 GPT Application

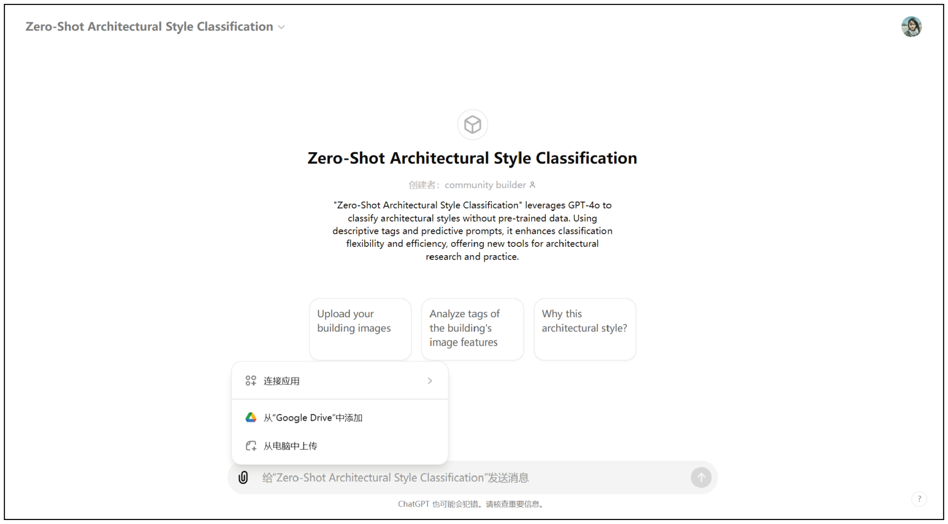
In this study, we present a zero-shot architectural style classification GPT based on our GPT-4o-TAGS+PICS methodology. This approach harnesses the advanced multimodal capabilities of GPT-4o, combining descriptive architectural feature tags with predictive prompts to classify styles without the need for any model fine-tuning or pre-training on specific architectural datasets. As illustrated in (Fig. 10), users can simply upload a building image directly into the ChatGPT interface for instant analysis.

Upon image upload, the system processes visual features, aligns them with relevant architectural tags, and returns a predicted style along with a concise explanation for the classification. This intuitive workflow delivers not only high accuracy but also valuable insights into the distinctive features that define architectural styles.

We invite researchers, historians, and design practitioners to explore our tool by visiting Zero-Shot Architectural Style Classification. This application represents a meaningful advancement in computational architectural analysis, offering a fast, accessible, and explainable method for interpreting architectural typologies across diverse image inputs.

5 Conclusion

We introduced a novel zero-shot method for architectural style classification using a large language model with vision capabilities (GPT-4o). Our experiments showed that this approach can achieve accuracy comparable to supervised baselines, all while requiring no labeled training data. By carefully engineering prompts, we


Zero-Shot Architectural Style Classification
Zero-Shot Architectural Style Classification
创建者：community builder
Upload your building images
Analyze tags of the building's image features
Why this architectural style?
连接应用
从"Google Drive"中添加
从电脑中上传
给"Zero-Shot Architectural Style Classification"发送消息
ChatGPT 也可能会犯错。请核查重要信息。

"Zero-Shot Architectural Style Classification" leverages GPT-4o to classify architectural styles without pre-trained data. Using descriptive tags and predictive prompts, it enhances classification flexibility and efficiency, offering new tools for architectural research and practice.

Fig. 10. Zero-Shot Architectural Style Classification GPT.

leveraged GPT-4o's extensive learned knowledge to analyze both visual and textual cues—effectively blending "machine vision" with "architectural intuition." This leads to a flexible, interpretable alternative to traditional pipelines, as the model can articulate its reasoning in natural language.

5.1 Contributions

Our work offers a fresh tool for architectural computing, enabling rapid yet robust style classification that could benefit research in architectural history, style evolution, or educational contexts. More broadly, it illustrates how a single, generalist large model can be guided via prompt engineering to handle a specialized domain task. We also highlight recurrent challenges such as certain stylistic confusions, emphasizing the importance of well-designed prompts to steer the model's reasoning effectively.

5.2 Longevity and Future Relevance

Some may view this zero-shot method as a short-lived approach specific to GPT-4o, but we argue that its relevance is likely to grow. Ongoing developments in multimodal models (e.g., the next generations of large vision-language systems) will likely bring even stronger zero-shot capabilities. Our framework is largely model-agnostic: any sufficiently advanced model that can process images and text could be substituted, potentially achieving higher accuracy or fewer errors with each iteration of training. As AI's visual understanding continues to improve, the need for exhaustive, task-specific datasets could diminish for a wide range of classification applications, making methods like ours increasingly appealing.

5.3 Limitations and Future Work

Despite promising results, our approach does come with certain limitations:

- **Black-box reliance:** GPT-4o heavily depends on prompts, and we have limited means of adjusting its internal representation. If specific styles are repeatedly confused, we cannot fine-tune in the usual sense but must rely on further prompt refinements or wait for future model updates.
- **Bias and subjectivity:** Although we did not detect overt bias in style classifications, large models can inherit societal or historical biases from their training data. Researchers should ensure prompts remain neutral and avoid eliciting subjective or value-laden responses about architectural aesthetics or design.
- **Multi-label classification:** Architectural styles commonly overlap (e.g., Baroque structures with Gothic remnants). GPT-4o might acknowledge multiple influences in free-form text, but our current evaluation protocol expects a single label. Future work could adopt partial-credit or multi-label schemes to better capture stylistic nuance.
- **Integrating Additional Feature Extraction Modules:** While GPT-4o excels in zero-shot classification via prompt engineering, it may overlook subtle architectural features, particularly in images with overlapping or nuanced style cues. Pre-trained networks such as GoogLeNet or IEP could be employed to produce domain-specific descriptors (e.g., localized architectural elements, textual summaries of structural details). These outputs would then serve as additional prompts or context for GPT-4o—without fine-tuning the large model itself—thus enhancing zero-shot performance by providing richer visual cues. This synergistic approach retains the flexibility of zero-shot classification while incorporating specialized architectural insights from traditional computer vision pipelines.

5.4 Outlook

Zero-shot architectural style classification using LLMs signals a new paradigm that fuses image recognition with broad contextual knowledge. Although we focused on style labels, the same principles can extend to other architectural or urban classification tasks by crafting suitable prompts. As large multimodal models evolve, their enhanced visual–textual reasoning could further reduce the need for labeled data, making zero-shot solutions both practical and adaptable. We hope this study stimulates deeper exploration of prompt-driven classification, multi-label handling, and the integration of cultural knowledge within intelligent systems. Ultimately, our findings underscore a shift toward more general, less data-intensive AI strategies in architecture and design—one that promises to reshape future workflows in both research and practice.

References

1. Chen, J., Stouffs, R., Biljecki, F.: Hierarchical (multi-label) architectural image recognition and classification. In: 26th International Conference of the Association

for Computer-Aided Architectural Design Research in Asia: Projections, CAADRIA 2021. pp. 161–170 (2021)

2. Chen, R., Xiong, T., Wu, Y., Liu, G., Hu, Z., Chen, L., Chen, Y., Liu, C., Huang, H.: Gpt-4 vision on medical image classification–a case study on covid-19 dataset. arXiv preprint arXiv:2310.18498 (2023)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. International conference on machine learning pp. 1597–1607 (2020)
4. Erdogan, E., Akalin, A., Yildirim, K., Erdogan, H.A.: Students' evaluations of different architectural styles. Procedia-Social and Behavioral Sciences **5**, 875–881 (2010)
5. Geirhos, R., Janssen, D.H., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, F.A.: Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969 (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Li, H., Dong, H.: Architectural style classification based on deep learning (2025)
8. Llamas, J., M. Lerones, P., Medina, R., Zalama, E., Gómez-García-Bermejo, J.: Classification of architectural heritage images using deep learning techniques. Applied Sciences **7**(10), 992 (2017)
9. Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022)
10. Pan, F., Jeon, S., Wang, B., Mckenna, F., Yu, S.X.: Zero-shot building attribute extraction from large-scale vision and language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8647–8656 (2024)
11. Selvaraju, R.R., Desai, K., Johnson, J., Naik, N.: Casting your model: Learning to localize improves self-supervised representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11058–11067 (2021)
12. Shalunts, G., Haxhimusa, Y., Sablatnig, R.: Architectural style classification of building facade windows. In: International Symposium on Visual Computing. pp. 280–289. Springer (2011)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Sun, M., Zhang, F., Duarte, F., Ratti, C.: Understanding architecture age and style through deep learning. Cities **128**, 103787 (2022)
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
17. Wang, B., Zhang, S., Zhang, J., Cai, Z.: Architectural style classification based on cnn and channel–spatial attention. Signal, Image and Video Processing **17**(1), 99–107 (2023)
18. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

19. Wang, W., Zheng, V.W., Yu, H., Miao, C.: A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–37 (2019)
20. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big Data **3**(1), 1–40 (2016)
21. Wu, W., Yao, H., Zhang, M., Song, Y., Ouyang, W., Wang, J.: Gpt4vis: what can gpt-4 do for zero-shot visual recognition? arXiv preprint arXiv:2311.15732 (2023)
22. Xia, B., Li, X., Shi, H., Chen, S., Chen, J.: Style classification and prediction of residential buildings based on machine learning. Journal of Asian Architecture and Building Engineering **19**(6), 714–730 (2020)
23. Xiao, Y.: Self-supervised learning in deep networks: A pathway to robust few-shot classification. arXiv preprint arXiv:2411.12151 (2024)
24. Xu, Z., Tao, D., Zhang, Y., Wu, J., Tsoi, A.C.: Architectural style classification using multinomial latent logistic regression. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 600–615. Springer (2014)
25. Yoshimura, Y., Cai, B., Wang, Z., Ratti, C.: Deep learning architect: classification for architectural design through the eye of artificial intelligence. Computational Urban Planning and Management for Smart Cities 16 pp. 249–265 (2019)
26. Zeng, Z., Goo, J.M., Wang, X., Chi, B., Wang, M., Boehm, J.: Zero-shot building age classification from facade image using gpt-4. arXiv preprint arXiv:2404.09921 (2024)
27. Zhao, P., Miao, Q., Song, J., Qi, Y., Liu, R., Ge, D.: Architectural style classification based on feature extraction module. Ieee Access **6**, 52598–52606 (2018)