Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70113

Prompt-Grounded Architectural Style Retrieval: Matching Building Images with LLM-Generated Design Descriptions

Sarah Tan[1], *Derek Zhou*[2]
[1]*Computer Science, University of Washington, Seattle, WA, USA*
[2]*Architecture, University of Washington, Seattle, WA, USA*
stan.dev.2025@gmail.com

Keywords

architectural style;
image-text retrieval;
large language models;
prompt engineering;
building recognition;
cross-modal embedding

Abstract

Architectural style retrieval requires a model to connect visible design evidence with language that is more specific than a class name. This study evaluates prompt-grounded retrieval on the BuildingStyles corpus, whose labeled manifest contains 456 square RGB images distributed across 24 architectural styles. GPT-5.5 Pro generated three concise, photograph-grounded descriptions for each style, yielding a frozen bank of 72 design descriptions. The Prompt-Grounded Architectural Retrieval (PGAR) pipeline represents every image with a 1,998-dimensional descriptor covering gradient structure, color, local texture, and edge profiles; reduces the descriptor to 128 principal components; and learns a ridge projection into a 64-dimensional text space constructed from word and character TF-IDF features. A training-fold-calibrated hybrid combines the projected prompt score with shrinkage linear discriminant scores. All 456 images were evaluated once as held-out cases through stratified five-fold cross-validation. For image-to-text retrieval, the hybrid reached Recall@1 of 28.73%, Recall@5 of 62.50%, and mean reciprocal rank of 0.4484. Relative to label-only prompts, Recall@5 increased by 7.68 percentage points and mean reciprocal rank by 0.0340; paired bootstrap intervals excluded zero for both differences. For text-to-image retrieval, the hybrid achieved the highest mean average precision, 0.2894, compared with 0.2447 for label-only queries. The experiments also quantify prompt composition, visual-feature contributions, per-style behavior, cluster overlap, dominant confusions, execution time, and robustness to six image perturbations. The results establish that LLM-generated visual descriptions improve deeper retrieval and bidirectional ranking, while architectural families with shared massing and classical vocabularies remain the principal source of error.

INTRODUCTION

Architectural style is communicated through combinations of massing, roof geometry, structural rhythm, openings, materials, and ornament. A single photograph rarely reveals that system completely: viewpoint, occlusion, renovation, and revival design can hide or mix the cues that distinguish neighboring traditions. Studies of urban appearance and architectural classification nevertheless show that facade parts, streetscape details, and relations among visible components carry recoverable stylistic information [1], [2].

Retrieval is a useful formulation because it does not require the model to compress an ambiguous building into one unqualified label. In the image-to-text direction, the system ranks the descriptions that best account for a photograph; in the text-to-image direction, it ranks photographs that express a design description. Scene-centered recognition and

building-level mapping provide important spatial context [3], [4], but architectural photographs require language that can name observable evidence at facade scale.

Natural-language supervision makes that interface possible. Rather than querying the collection with a class name alone, a user can specify a stepped profile, a hipped roof, round arches, exposed rafters, a curtain wall, or another visible combination. The central challenge is to determine whether such descriptions improve ranking beyond label-only prompts and whether the gain remains after visual baselines, prompt ablations, and score calibration are considered.

Two evaluation directions are therefore necessary. Image-to-text retrieval tests whether the correct style description moves toward the top of a 24-class list. Text-to-image retrieval tests whether a description gathers relevant photographs near the top of a 456-image collection. The second direction is especially important for archive browsing, precedent search, and design-reference discovery, where users commonly inspect several candidates rather than accept one predicted label.

This study addresses three questions: how much visually grounded descriptions improve retrieval over style names and generic templates; which visual cues and prompt compositions account for the improvement; and where the method remains vulnerable to shared architectural vocabularies, limited viewpoints, and image perturbations. The analysis combines bidirectional ranking metrics, paired uncertainty estimates, per-style diagnostics, clustering, confusion analysis, and robustness tests.

This study develops Prompt-Grounded Architectural Retrieval (PGAR) and makes four contributions. It audits the complete 456-record labeled manifest of the BuildingStyles corpus, covering 24 architectural styles; constructs a fixed bank of 72 concise design descriptions; learns a lightweight projection from interpretable appearance features into the description space and calibrates it with a shrinkage discriminant model; and evaluates prompt composition, bidirectional retrieval, feature contributions, statistical significance, per-style behavior, cluster overlap, runtime, and robustness.

LITERATURE REVIEW

Architectural recognition and visual evidence

Architectural image analysis has moved from handcrafted facade cues to latent part models and scene-level supervision. Urban appearance studies identified local elements such as windows, balconies, signs, and facade details as discriminative geographic signals [1], while multinomial latent logistic regression modeled relations among architectural parts for style classification [2]. Places expanded scene-centric visual supervision [3], and the Urban Building Classification dataset formalized building detection and functional classification in overhead imagery [4]. Together, this work establishes that architecture is recognizable from visual structure, but it also highlights a scale mismatch: the evidence needed for street-level style retrieval is often finer-grained than scene labels and less stable than overhead geometry.

Vision-language models and prompt construction

Large-scale vision-language pre-training changed how visual categories can be specified. CLIP and ALIGN aligned image and text representations from large image-text collections [5], [6]; SigLIP replaced global softmax normalization with a pairwise sigmoid objective [7]; and BLIP and BLIP-2 connected retrieval and captioning with generative language components [8], [9]. Vision Transformers supplied a high-capacity image encoder for many later systems [10]. These models support zero-shot recognition, but their behavior depends on the wording used to represent a class.

Prompt research has consequently shifted from fixed templates to descriptive and adaptive language. Description-based classification uses a large language model to enumerate recognizable attributes [17], and CuPL generates class-specific prompts rather than relying on a uniform template [18]. CoOp and CoCoOp learn context tokens from labeled images [19], [20], while prompt weighting and automatic prompt optimization seek better combinations of multiple descriptions [21], [22]. These methods build on the few-shot and instruction-following capabilities of large language

models [23], [24]. For architectural styles, the important distinction is between historical or cultural background and features a photograph can actually show.

Grounded design language, explanation, and calibrated fusion

Recent design-oriented studies provide a complementary perspective on grounded language. A vision-language workflow for converting hand-drawn sketches into web prototypes evaluated both structural and visual consistency, showing the value of separating semantic correspondence from surface similarity [25]. LLM-based design critique has also been evaluated against human graphic-design judgment [26], while text-grounded rationale cards and structured visual briefs organize intentions, visible elements, hierarchy, and review cues into inspectable representations [27], [28]. Although these studies address interfaces rather than architectural retrieval, they reinforce a common principle: generated language is most useful when it is tied to observable evidence and organized for human inspection.

Retrieval systems likewise benefit when ranking and explanation are treated as connected but distinct functions. Retrieval-summary integration has been used to make code search results both findable and interpretable [29]. In multimodal estimation, uncertainty-aware late fusion combines separately calibrated evidence streams rather than assuming that raw scores are directly comparable [30]. PGAR adopts these ideas in a compact form: descriptions define the retrieval target, the projected prompt score supplies semantic evidence, the visual classifier supplies discriminative evidence, and training-fold calibration combines them without using held-out images.

Research gap and study position

Existing work leaves a specific gap at the intersection of architectural recognition, descriptive prompting, and bidirectional ranking. Most prompt studies emphasize top-1 classification on object categories, whereas architectural styles overlap in massing, ornament, and revival vocabulary. Most architectural classifiers return labels rather than ranked, human-readable descriptions. The present study therefore treats style retrieval as an evidence-matching problem and evaluates not only whether the correct style wins, but also how far it moves through the ranked list, how well a text query retrieves relevant photographs, and which visual families remain inseparable.

MATERIALS AND METHODS

The PGAR workflow has two coordinated branches. The visual branch extracts and reduces image descriptors, while the language branch embeds the fixed description bank; a training-fold calibration stage combines their scores for bidirectional retrieval. Figure 1 summarizes the complete sequence.

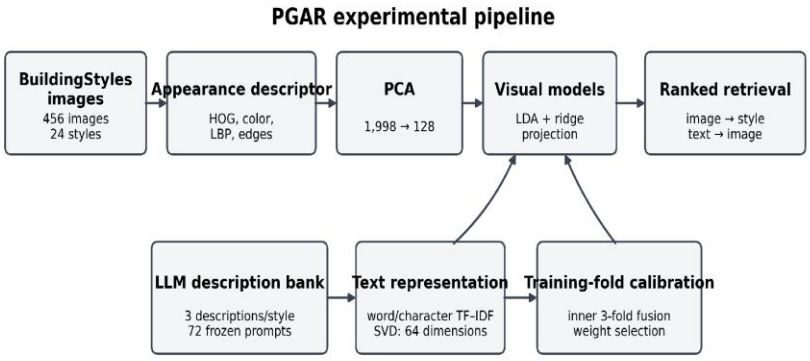

Figure 1. PGAR experimental pipeline. Image transforms were fitted inside each outer training fold; the language bank was generated once and then held fixed.

Table 1. Corpus integrity audit

Audit item	Result
JPEG files in package	512

Audit item	Result
Labeled Arrow records	456
Evaluated architectural styles	24
Manifest images matched to distinct JPEGs	456
Nonmanifest JPEG files	56
Image dimensions	512 × 512 px
Image mode after decoding	RGB
Corrupt labeled images	0
Exact duplicate groups	0
Identical 64-bit dHash groups	0

Corpus construction and integrity audit

The BuildingStyles package contains 512 JPEG files, an Arrow table, and dataset metadata. The Arrow table serves as the labeled manifest and contains 456 records, each pairing an image with a sentence that names its architectural style. Every manifest image was decoded, converted to RGB, measured, hashed, and matched to a JPEG file. All 456 labeled images were 512 × 512 pixels, and every record matched one distinct JPEG. The remaining 56 JPEG files fell outside the labeled manifest and were excluded. Cryptographic hashes found no exact duplicates, 64-bit difference hashes found no visually identical groups, and all labeled files decoded successfully. Table 1 summarizes the audit.

The evaluation corpus therefore contains 24 styles and 456 images. Class support ranges from 14 images for Achaemenid and Queen Anne to 31 for American Foursquare; the median class contains 19 images, and 18 of the 24 classes contain between 18 and 21 images. Table 2 and Figure 3 report the class distribution. Figure 2 presents the first manifest image from each class and illustrates the variation in viewpoint, scale, background, lighting, and facade completeness.

Figure 2. Representative manifest image for each of the 24 architectural styles. The montage preserves variation in viewpoint, background, scale, and facade completeness.

Table 2. Class composition of the 456-image evaluation corpus

Style	N	Style	N
Achaemenid	14	Deconstructivism	21

Style	N	Style	N
American Foursquare	31	Edwardian	15
American craftsman	21	Georgian	17
Ancient Egyptian	19	Greek Revival	20
Art Deco	20	International	18
Art Nouveau	18	Novelty	21
Baroque	19	Palladian	20
Bauhaus	20	Postmodern	19
Beaux-Arts	18	Queen Anne	14
Byzantine	20	Romanesque	16
Chicago school	19	Russian Revival	21
Colonial	19	Tudor Revival	16

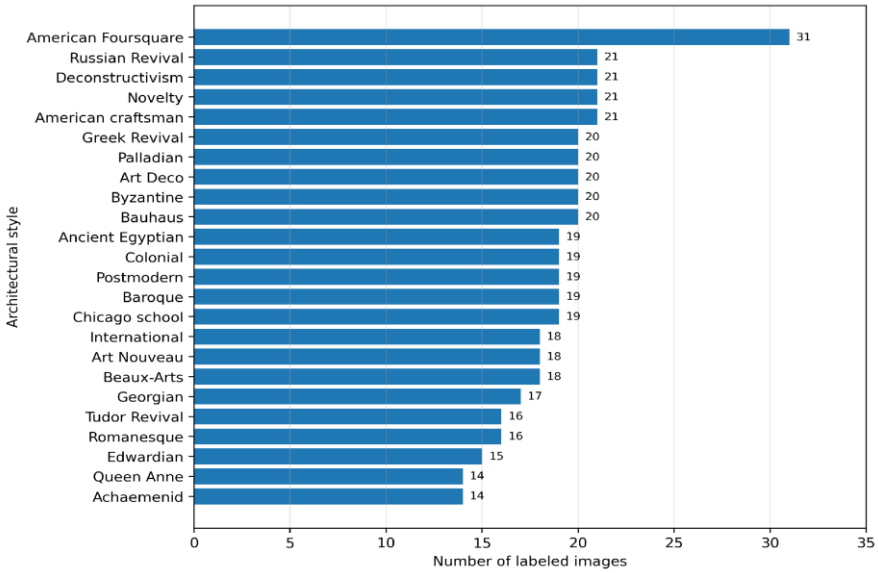

Figure 3. Class support in the 456-image labeled manifest. Counts range from 14 to 31 images.

LLM-generated design descriptions

GPT-5.5 Pro was used to generate three descriptions for each of the 24 style names. The generation instruction asked for concise, photograph-grounded language and prioritized visible massing, roof form, openings, materials, structure, and ornament when these cues were diagnostic. Dates, locations, architects, and cultural commentary were excluded. This attribute-centered construction follows description-based vision-language classification [17], [18] and instruction-following language modeling [23], [24]. The resulting bank contained 72 sentences, averaging approximately 20 words each. It was created before model fitting and held constant across all folds.

Six prompt sets were evaluated. Label only used the style name. Generic template used “a photograph of a [style] building.” LLM single used the first generated sentence. LLM two-description averaged the first two descriptions for each style. LLM three-description ensemble averaged all three. LLM visual-only ensemble replaced every occurrence of the style name with “the building,” retaining only attribute language. Many third sentences were contrastive: Art

Deco was separated from Art Nouveau through angular rather than flowing botanical geometry, while Greek Revival was separated from Palladian design through a dominant temple front rather than a horizontally extended central block and wings.

Prompt separability was measured after text embedding by the cosine similarity between class prototypes. Generic templates produced the highest inter-style similarity because the shared template words dominated every sentence. The LLM descriptions were longer but substantially more distinctive. Table 3 reports the prompt inventory and embedded similarity statistics.

Table 3. Prompt inventory and embedded inter-style similarity

Prompt set	Sent.	Mean words	Mean cosine	Max cosine
Label only	24	1.42	-0.0150	0.8196
Generic template	24	6.42	0.6234	0.9270
LLM single	24	20.25	0.0620	0.3086
LLM two-description	48	19.83	0.1101	0.3317
LLM three-description ensemble	72	20.36	0.1607	0.6735
LLM visual-only ensemble	72	20.64	0.1888	0.7283

Architectural appearance descriptor

Every image was resized to 128 × 128 pixels after EXIF orientation correction. The fixed 1,998-dimensional appearance descriptor combined five feature groups. A 1,764-dimensional histogram of oriented gradients used nine orientations, 16 × 16-pixel cells, 2 × 2-cell blocks, and L2-Hys normalization, following the structural edge representation introduced by Dalal and Triggs [11]. Color was represented by 16-bin histograms for each HSV channel (48 dimensions) and by RGB means and standard deviations in a 4 × 4 spatial grid (96 dimensions). Uniform local binary patterns with 24 sampling points and radius 3 contributed 26 texture dimensions [12]. Mean horizontal and vertical gradient profiles, each resampled to 32 positions, contributed 64 edge-layout dimensions.

Within each outer training fold, every feature dimension was standardized from training-fold statistics. Principal component analysis then reduced the descriptor to 128 whitened components; both the scaler and PCA transform were applied unchanged to the held-out fold. PCA was used because it supplies an explicit low-dimensional linear basis and stable covariance representation [13]. No held-out image contributed to standardization, component estimation, model fitting, or fusion-weight selection.

Text representation and prompt projection

All frozen prompt sentences were represented with a union of word and character TF-IDF features. The word channel used unigrams and bigrams with English stop words removed and sublinear term frequency. The character channel used within-word 3- to 5-grams, sublinear term frequency, and a 4,000-feature ceiling. A 64-component truncated singular-value decomposition was fitted to the complete frozen prompt bank. Because prompt generation and text embedding used no image or fold label information, the same fixed text basis was used across folds. Each class prototype was the normalized mean of its selected description vectors.

For each prompt condition, ridge regression mapped the 128-dimensional visual vector z_i to the prototype p_{yi} of the training image's class. The fitted mapping minimized the sum of squared projection errors plus an L2 penalty with $\alpha = 1.0$, following the regularized least-squares principle of ridge regression [14]. The prompt score for image i and style c was the cosine similarity between the normalized projected vector and p_c. Ranking the 24 class prototypes produced image-to-text retrieval; ranking the 456 calibrated image scores for a fixed p_c produced text-to-image retrieval.

Visual baselines and calibrated fusion

Five image-only baselines were fitted on the same fold-specific 128-dimensional representation: nearest visual centroid, ridge classifier, shrinkage linear discriminant analysis, radial-basis-function support vector machine, and multinomial logistic regression. Shrinkage LDA used the least-squares solver with automatically estimated covariance shrinkage, building on Fisher's discriminant criterion [15]. The RBF SVM used C = 3 and probability-free decision scores, following the maximum-margin formulation of support vector machines [16]. Logistic regression used C = 0.1; the ridge classifier used $\alpha = 10$.

PGAR calibrated hybrid combined shrinkage-LDA scores with three-description prompt-projection scores. For every image, both 24-element score vectors were standardized across classes. The fused score was $s\_ic = (1 - \lambda)\, z(s\hat{}V\_ic) + \lambda\, z(s\hat{}P\_ic)$, where $s\hat{}V$ is the LDA score, $s\hat{}P$ is the prompt score, and z denotes row-wise standardization. Within each outer training fold, a stratified three-fold inner loop selected λ from $\{0.0, 0.1, \ldots, 1.0\}$ by mean reciprocal rank. The five selected weights were 0.0, 0.2, 0.4, 0.3, and 0.1. A second text-calibrated fusion combined label-only and three-description projection scores by the same procedure.

Text-to-image queries require comparable score scales across classes. For that direction, each class score was standardized with the mean and standard deviation measured on the corresponding outer training fold, and held-out scores were transformed with those training statistics. This calibration does not alter the within-image ordering used for image-to-text retrieval; it corrects class-specific score offsets before images are compared for a fixed text query.

Evaluation protocol and statistical analysis

Stratified five-fold cross-validation used shuffle seed 2025. Every image appeared in exactly one test fold, and the five held-out score matrices were concatenated in manifest order. Depending on class support, each fold contained two to seven test images per style. Cross-validation allowed every labeled image to serve as a held-out case while keeping standardization, PCA, model fitting, and calibration confined to the training data [34]. All methods used the same folds, image order, features, and class prototypes.

Image-to-text retrieval was evaluated with Recall@K for K = 1 through 10, mean reciprocal rank (MRR), mean and median rank, and NDCG@5 and NDCG@10. Text-to-image retrieval was evaluated across the 24 text queries with macro-averaged precision, recall, and NDCG at 5, 10, and 20, first relevant rank, and mean average precision (mAP), following standard ranked-retrieval practice [36]. Wilson 95% intervals were calculated for Recall@1 and Recall@5. Paired bootstrap resampling with 3,000 image-level samples estimated confidence intervals for hybrid-minus-label differences in Recall@1, Recall@5, and MRR [35], and an exact McNemar test compared top-ranked correctness.

Feature ablations re-fitted the prompt projection with HOG alone, color alone, texture plus edge profiles, HOG plus color, HOG plus texture/edge, and all cues. Robustness tests fitted every model on original training images and evaluated the same held-out identities after grayscale conversion, an 80% center crop resized to the original dimensions, Gaussian blur with radius 1.5, brightness scaling to 0.65, JPEG recompression at quality 25, and horizontal reflection.

The 64-dimensional held-out prompt projections were visualized with t-distributed stochastic neighbor embedding using perplexity 30, PCA initialization, and seed 2025 [31]. Style centroids were clustered by average linkage over cosine distance. Cluster quality was quantified with the cosine silhouette coefficient [32], Davies-Bouldin index [33], and Calinski-Harabasz index. Negative silhouette values were retained because they directly measure cross-style overlap in the learned score spaces.

Experiments ran in Python 3.13.5 with NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, SciPy 1.17.0, scikit-image 0.26.0, Pillow 12.2.0, and PyArrow 24.0.0. Eight CPU threads were assigned to linear algebra, and no GPU was used. Table 4 lists the fixed configuration.

Table 4. Fixed experimental configuration

Component	Setting
Evaluation	Stratified 5-fold outer CV; seed 2025

Component	Setting
Inner selection	Stratified 3-fold; $\lambda \in \{0.0, 0.1, \ldots, 1.0\}$
Images/classes	456 / 24
Image descriptor	1,998 dimensions
PCA	128 whitened components
Text representation	Word 1–2 grams + character 3–5 grams
Text SVD	64 components
Projection	Ridge regression, $\alpha = 1.0$
Hybrid visual model	Shrinkage LDA
Image size for features	128 × 128
Software	Python 3.13.5; scikit-learn 1.8.0
Compute	8 CPU threads; no GPU

RESULTS

Prompt properties and image-to-text retrieval

Table 3 shows that the generic template collapsed the text space: its mean inter-style cosine similarity was 0.6234 and its closest style pair reached 0.9270. Its image-to-text Recall@1 was consequently 19.08%, 7.90 percentage points below label-only retrieval. The single LLM description reduced mean inter-style cosine similarity to 0.0620 and raised Recall@1 to 27.19%. Two descriptions produced the strongest prompt-only top-rank result, 28.95%, and MRR of 0.4469. The three-description average retained high deeper recall, 61.18% at K = 5, but its Recall@1 returned to 26.97%. Removing class names did not eliminate the gain: the visual-only ensemble reached 28.51% Recall@1 and 62.50% Recall@5. The attribute language therefore carried retrieval signal independently of the literal style token.

The complete comparison appears in Table 5 and Figure 4. RBF SVM achieved the highest Recall@1, 30.26%, but its deeper recall and MRR trailed several linear and prompt-based methods. Shrinkage LDA produced the highest MRR, 0.4531, with 29.61% Recall@1 and 61.18% Recall@5. PGAR calibrated hybrid reached 28.73% Recall@1, tied the highest Recall@5 at 62.50%, and obtained MRR of 0.4484. LLM two-description reached the highest Recall@10, 79.17%. Thus, winning at the first rank did not guarantee the strongest candidate list.

Relative to label-only prompts, the hybrid increased Recall@1 from 26.97% to 28.73%, Recall@5 from 54.82% to 62.50%, and MRR from 0.4144 to 0.4484. The mean correct-style rank improved from 6.81 to 5.92. The Recall@5 gain of 7.68 percentage points had a paired 95% bootstrap interval of 3.95 to 11.40 points. The MRR gain of 0.0340 had an interval of 0.0111 to 0.0566. The Recall@1 interval crossed zero, so the main statistically supported benefit was better ranking depth rather than a decisive change in first-position accuracy.

Table 5. Image-to-text retrieval over 24 style descriptions

Method	R@1 %	R@3 %	R@5 %	R@10 %	MRR	Mean rank
Label only	26.97	45.18	54.82	75.00	0.4144	6.81
Generic template	19.08	35.53	46.05	65.79	0.3338	8.20

Method	R@1 %	R@3 %	R@5 %	R@10 %	MRR	Mean rank
LLM single	27.19	48.46	58.99	77.85	0.4298	6.26
LLM two-description	28.95	50.00	61.62	79.17	0.4469	6.00
LLM three-description ensemble	26.97	51.10	61.18	78.51	0.4328	6.04
LLM visual-only ensemble	28.51	50.22	62.50	78.29	0.4383	6.09
Nearest visual centroid	28.07	50.00	59.43	77.85	0.4372	6.29
Ridge classifier	29.39	49.78	60.31	77.85	0.4415	6.25
Shrinkage LDA	29.61	51.54	61.18	78.29	0.4531	5.92
RBF SVM	30.26	46.05	56.14	75.00	0.4358	6.78
Logistic regression	27.19	47.37	60.31	78.95	0.4297	6.27
PGAR text-calibrated fusion	27.63	50.66	60.96	76.97	0.4354	6.15
PGAR calibrated hybrid	28.73	51.10	62.50	78.29	0.4484	5.92

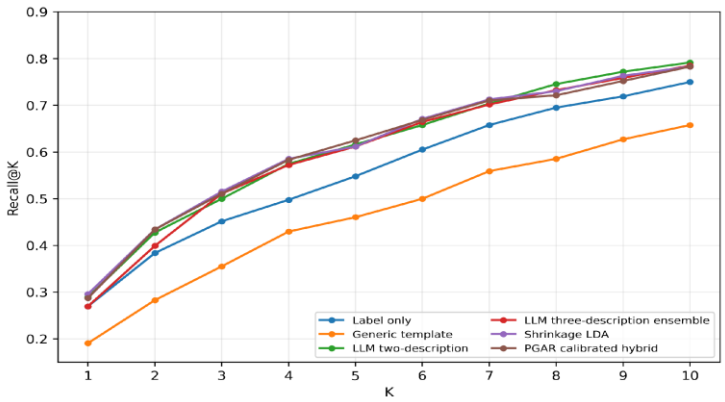

Figure 4. Recall@K curves for representative prompt, visual, and fused methods. Every point is calculated from the same 456 held-out predictions.

Text-to-image retrieval

The ranking direction changed the method order. PGAR calibrated hybrid achieved the highest text-to-image mAP, 0.2894, followed by shrinkage LDA at 0.2869 and LLM single at 0.2841 (Table 6). The hybrid also obtained the

highest P@5, 0.4333, highest NDCG@5, 0.4639, and highest NDCG@10, 0.4098. Label-only queries reached mAP 0.2447; the hybrid therefore added 0.0446 absolute mAP, an 18.2% relative increase. At 20 images per query, the hybrid retrieved 31.96% of each class on average, compared with 26.66% for label-only prompts.

Figure 5 translates the metrics into ranked photographs. Ancient Egyptian placed five relevant pyramid or temple views in the first five positions, while Art Nouveau returned two relevant facades before ornate Novelty and Baroque alternatives. Deconstructivism retrieved fragmented angular buildings at ranks one and three; Romanesque placed an arched masonry facade first. Diagnostic silhouettes therefore concentrated the rankings, whereas styles built from shared facade elements produced mixed lists.

Table 6. Text-to-image retrieval, macro-averaged across 24 style queries

Method	P@5	R@5	NDCG@5	P@10	NDCG@10	mAP
Label only	0.4000	0.1059	0.4182	0.3208	0.3566	0.2447
Generic template	0.3750	0.1000	0.3791	0.3000	0.3261	0.2258
LLM single	0.4250	0.1144	0.4567	0.3708	0.4074	0.2841
LLM two-description	0.4083	0.1096	0.4392	0.3542	0.3904	0.2822
LLM three-description ensemble	0.4250	0.1149	0.4494	0.3417	0.3820	0.2725
LLM visual-only ensemble	0.4083	0.1105	0.4455	0.3375	0.3831	0.2733
Nearest visual centroid	0.4167	0.1132	0.4219	0.3583	0.3798	0.2742
Ridge classifier	0.4000	0.1072	0.4436	0.3375	0.3859	0.2775
Shrinkage LDA	0.4083	0.1100	0.4460	0.3625	0.4022	0.2869
RBF SVM	0.2917	0.0803	0.3012	0.2875	0.2959	0.2340
Logistic regression	0.3917	0.1064	0.4166	0.3250	0.3617	0.2569
PGAR text-calibrated fusion	0.4167	0.1116	0.4455	0.3333	0.3781	0.2714
PGAR calibrated hybrid	0.4333	0.1151	0.4639	0.3708	0.4098	0.2894

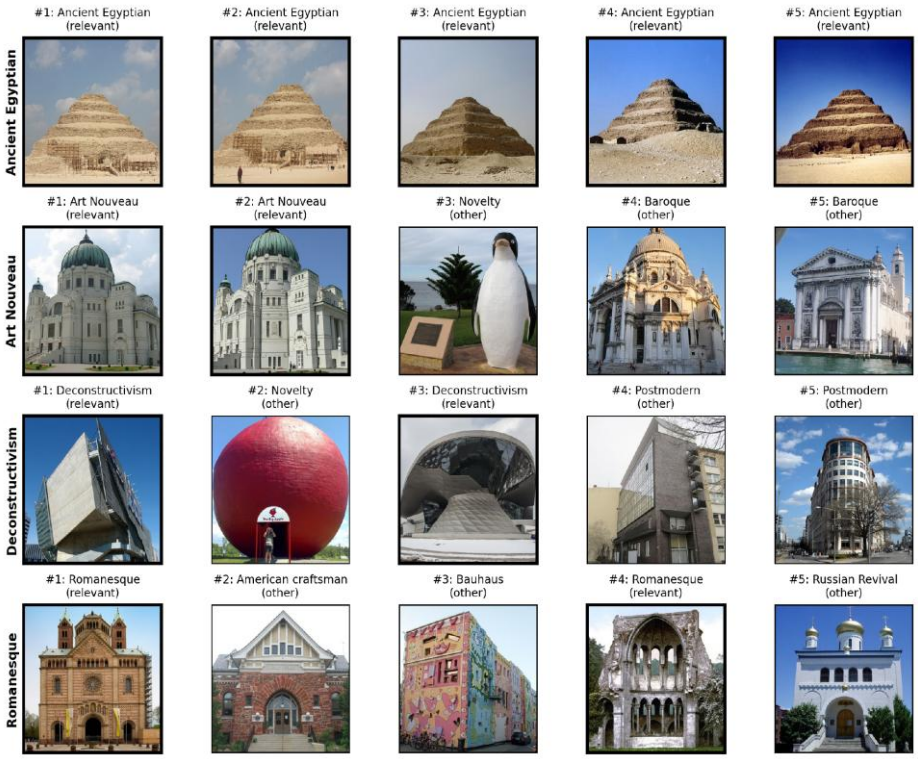

Figure 5. Top-five text-to-image results for four style queries using calibrated hybrid scores. Heavy borders mark relevant images, and titles state the manifest class.

Baseline and ablation results

Prompt ablation in Table 7 confirms that description count was not monotonic. The second description added complementary evidence and improved Recall@1 by 1.75 points over the first description. A third contrastive sentence increased the shared semantic content of class prototypes and reduced first-rank separation, although Recall@3 and Recall@5 remained high. The visual-only prompt set recovered 1.53 points at Recall@1 over the three-description set, showing that repeated class-name tokens were not necessary for the observed gain.

Table 8 compares the image-only baselines with the PGAR hybrid. Shrinkage LDA was the strongest visual method by MRR and text-to-image mAP. Ridge classification and nearest-centroid retrieval were competitive at shallow ranks, while logistic regression was close to the label projection at Recall@1 but stronger at Recall@5. The RBF SVM reached the highest image-to-text Recall@1, yet its text-to-image mAP fell below every linear visual baseline. The directional imbalance reflects class-specific decision-score scales and supports the training-fold calibration used for text queries.

Feature ablation results in Table 9 show that structural gradients carried most of the signal. HOG alone reached 25.66% Recall@1 and MRR 0.4144. Color alone reached 14.69%, and texture plus edge profiles reached 16.01%. Adding color to HOG raised Recall@5 from 58.99% to 61.40% but reduced Recall@1 slightly. The complete descriptor produced the best Recall@1, 26.97%, and the best MRR, 0.4328. Color and texture therefore complemented the candidate list but were insufficient without facade geometry.

Table 7. Prompt ablation for image-to-text retrieval

Prompt set	R@1 %	R@3 %	R@5 %	MRR	Mean rank
Label only	26.97	45.18	54.82	0.4144	6.81
Generic template	19.08	35.53	46.05	0.3338	8.20

Prompt set	R@1 %	R@3 %	R@5 %	MRR	Mean rank
LLM single	27.19	48.46	58.99	0.4298	6.26
LLM two-description	28.95	50.00	61.62	0.4469	6.00
LLM three-description ensemble	26.97	51.10	61.18	0.4328	6.04
LLM visual-only ensemble	28.51	50.22	62.50	0.4383	6.09

Table 8. Visual baselines and PGAR hybrid

Method	I→T R@1 %	I→T R@5 %	I→T MRR	T→I mAP	T→I NDCG@10
Nearest visual centroid	28.07	59.43	0.4372	0.2742	0.3798
Ridge classifier	29.39	60.31	0.4415	0.2775	0.3859
Shrinkage LDA	29.61	61.18	0.4531	0.2869	0.4022
RBF SVM	30.26	56.14	0.4358	0.2340	0.2959
Logistic regression	27.19	60.31	0.4297	0.2569	0.3617
PGAR calibrated hybrid	28.73	62.50	0.4484	0.2894	0.4098

Table 9. Visual feature ablation with three-description prompt projection

Feature set	Dims.	R@1 %	R@3 %	R@5 %	MRR
HOG only	1764	25.66	47.81	58.99	0.4144
Color only	144	14.69	30.92	41.01	0.2900
Texture and edge only	90	16.01	30.92	40.79	0.3006
HOG + color	1908	25.00	48.90	61.40	0.4189
HOG + texture/edge	1854	24.34	48.03	59.87	0.4089
All appearance cues	1998	26.97	51.10	61.18	0.4328

Per-style retrieval, confusions, and embedding structure

Per-style results appear in Table 10. Ancient Egyptian was the clearest category, with Recall@1 of 84.21%, Recall@5 of 94.74%, and MRR of 0.8837. Byzantine reached 55.00% Recall@1; Russian Revival reached 47.62%; Edwardian reached 46.67%; and American Foursquare reached 38.71%. The lowest Recall@1 values occurred for Palladian (5.00%), Achaemenid (7.14%), Deconstructivism (9.52%), Beaux-Arts (11.11%), and Tudor Revival (12.50%). Low

first-rank accuracy did not always imply weak deeper retrieval: Palladian reached 60.00% Recall@5, and Achaemenid reached 64.29%.

The t-SNE map in Figure 6 shows partial grouping rather than 24 isolated islands: Ancient Egyptian is concentrated, while domestic revival and classical categories intermingle. The dendrogram in Figure 7 joins Greek Revival with Palladian and places American Foursquare, craftsman, Georgian, Tudor Revival, and Queen Anne in a broader domestic branch. Table 11 confirms the overlap. All cosine silhouette scores were negative; PGAR had the least negative value, −0.0815, LDA the lowest Davies-Bouldin index, 5.1213, and the LLM ensemble the highest Calinski-Harabasz index, 4.6807.

The dominant confusion pairs in Table 12 and Figure 8 follow visible family resemblance. Seven of 17 Georgian images ranked American Foursquare first. American craftsman and American Foursquare were confused in both directions, accounting for six errors each. Chicago school ranked Bauhaus first for five images, while International ranked American Foursquare for five images and Postmodern for three. Greek Revival and Palladian each produced four reciprocal errors. These pairs share symmetry, porticoes, repetitive bays, or rectilinear massing, and a single facade view often suppresses the feature named in a contrastive prompt.

Table 10. Per-style image-to-text retrieval for the PGAR calibrated hybrid

Style	N	R@1 %	R@5 %	MRR	Mean rank
Achaemenid	14	7.14	64.29	0.3310	6.29
American Foursquare	31	38.71	80.65	0.5875	3.48
American craftsman	21	19.05	66.67	0.4098	5.57
Ancient Egyptian	19	84.21	94.74	0.8837	1.79
Art Deco	20	30.00	75.00	0.4911	5.35
Art Nouveau	18	33.33	66.67	0.5038	4.83
Baroque	19	21.05	52.63	0.3752	6.00
Bauhaus	20	35.00	55.00	0.4407	8.10
Beaux-Arts	18	11.11	50.00	0.3205	6.67
Byzantine	20	55.00	70.00	0.6364	5.65
Chicago school	19	31.58	73.68	0.5163	4.21
Colonial	19	21.05	42.11	0.3448	8.58
Deconstructivism	21	9.52	42.86	0.2849	8.86
Edwardian	15	46.67	73.33	0.5928	4.67
Georgian	17	17.65	47.06	0.3403	6.29
Greek Revival	20	35.00	75.00	0.5203	4.90
International	18	16.67	38.89	0.2786	10.22
Novelty	21	23.81	57.14	0.3944	5.81

Style	N	R@1 %	R@5 %	MRR	Mean rank
Palladian	20	5.00	60.00	0.2617	6.25
Postmodern	19	15.79	57.89	0.3233	6.42
Queen Anne	14	21.43	42.86	0.3384	7.21
Romanesque	16	37.50	81.25	0.5866	3.75
Russian Revival	21	47.62	76.19	0.5950	4.29
Tudor Revival	16	12.50	37.50	0.2605	8.94

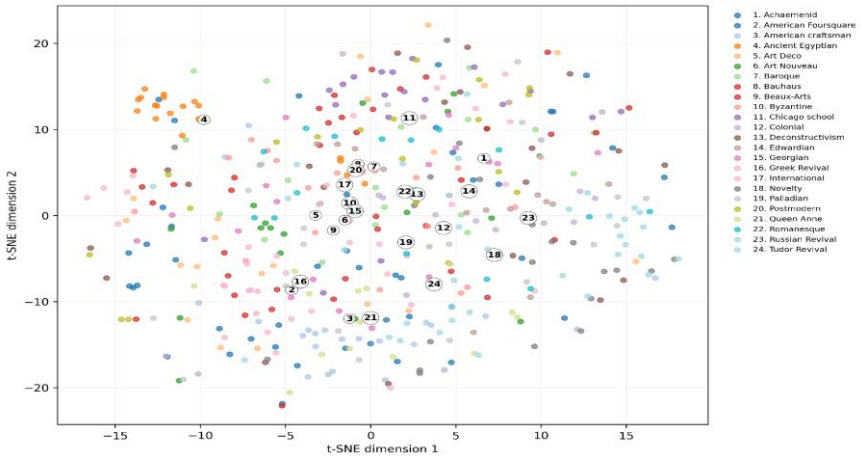

Figure 6. t-SNE visualization of the 64-dimensional held-out prompt projections. Numeric labels mark class centroids and correspond to the legend.

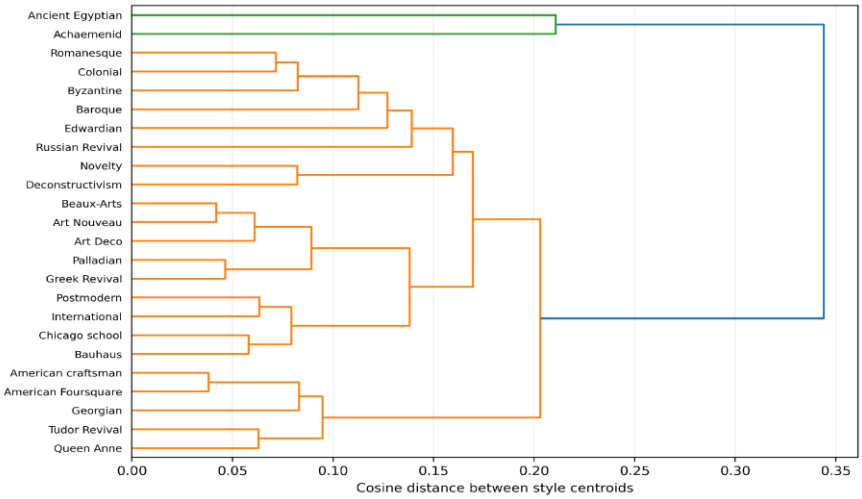

Figure 7. Average-linkage clustering of normalized style centroids using cosine distance. Short branches identify styles with similar held-out prompt projections.

Table 11. Embedding and score-space clustering quality

Representation	Cosine silhouette	Davies-Bouldin	Calinski-Harabasz
Shrinkage LDA scores	-0.0827	5.1213	4.4123

Representation	Cosine silhouette	Davies-Bouldin	Calinski-Harabasz
Label-prompt projection	-0.1128	5.7745	3.5270
LLM ensemble projection	-0.0987	5.1562	4.6807
PGAR calibrated hybrid	-0.0815	5.2244	4.2273

Table 12. Most frequent top-rank confusions for the PGAR calibrated hybrid

Ground truth	Top-ranked style	Count	Class share %
Georgian	American Foursquare	7	41.2
American craftsman	American Foursquare	6	28.6
American Foursquare	American craftsman	6	19.4
International	American Foursquare	5	27.8
Chicago school	Bauhaus	5	26.3
Art Deco	American Foursquare	4	20.0
Greek Revival	Palladian	4	20.0
Palladian	Greek Revival	4	20.0
Novelty	American Foursquare	4	19.0
Queen Anne	American Foursquare	3	21.4

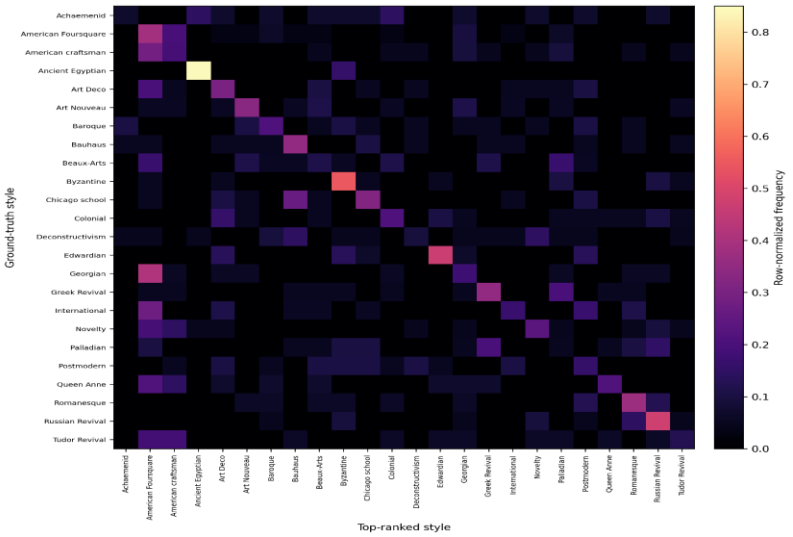

Figure 8. Row-normalized top-rank confusion matrix for the PGAR calibrated hybrid. Bright diagonal cells indicate high per-style Recall@1.

Robustness, significance, and execution time

Table 13 and Figure 9 report perturbation testing. Grayscale retained 28.95% Recall@1 and 62.28% Recall@5; blur yielded 28.07% and 63.38%. JPEG quality 25 reduced Recall@1 by 1.32 points and MRR by 0.0089. Center crop and low brightness reduced MRR to 0.4230 and 0.4244. Horizontal reflection was most damaging, lowering

Recall@1 to 26.10% and MRR to 0.4180. Grayscale and blur preserved gradient structure, whereas crop and reflection changed facade layout.

Table 14 summarizes paired tests. For hybrid versus label-only retrieval, the bootstrap p value was below 0.001 for Recall@5 and 0.003 for MRR. Recall@1 had p = 0.327, and the exact McNemar comparison of top-ranked correctness had p = 0.366. Thirty-four images were correct only under the hybrid, while 26 were correct only under label prompts. The test therefore did not establish a top-rank difference despite a positive numerical change.

Model fitting was lightweight. Across the five outer folds, nearest-centroid fitting required 0.067 s, ridge classification 0.101 s, shrinkage LDA 0.621 s, text-calibrated prompt fusion 2.179 s, and PGAR hybrid 3.486 s. The hybrid's added cost came from repeated inner-fold selection of the fusion weight. The complete main run, using cached image descriptors, finished in 25.24 s.

Table 13. Robustness with models fitted on original training images

Condition	Prompt R@1 %	Prompt R@5 %	Hybrid R@1 %	Hybrid R@5 %	Hybrid MRR
Original	26.97	61.18	28.73	62.50	0.4484
Grayscale	26.75	62.28	28.95	62.28	0.4445
Center crop (80%)	26.10	57.46	26.75	60.75	0.4230
Gaussian blur	27.19	62.06	28.07	63.38	0.4452
Low brightness	25.22	57.89	26.54	59.65	0.4244
JPEG quality 25	25.22	61.18	27.41	61.84	0.4395
Horizontal flip	24.78	57.68	26.10	60.53	0.4180

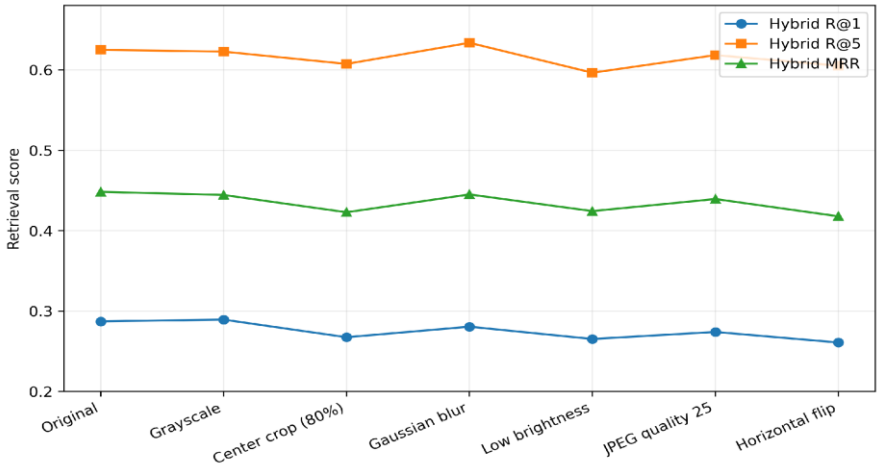

Figure 9. Hybrid retrieval under six held-out image perturbations. Models and transformations were fitted on original training images only.

Table 14. Paired hybrid-versus-label tests and selected fit times

Comparison or cost	Estimate	95% interval / scope	p value
R@1 difference	1.75 points	[-1.54, 5.04]	0.327
R@5 difference	7.68 points	[3.95, 11.40]	<0.001

Comparison or cost	Estimate	95% interval / scope	p value
MRR difference	0.0340	[0.0111, 0.0566]	0.003
McNemar top-1	34 hybrid-only / 26 label-only	60 discordant	0.366
Label projection fit time	0.214 s	five outer folds	—
Shrinkage LDA fit time	0.621 s	five outer folds	—
PGAR hybrid fit time	3.486 s	outer + inner folds	—
Complete main run	25.239 s	cached descriptors	—

DISCUSSION

Design descriptions improved the position of the correct style more reliably than they changed the first prediction. The hybrid added 7.68 points at Recall@5 and 0.0340 MRR over label-only prompts, while its 1.75-point Recall@1 gain was not significant. This distinction matters because researchers and practitioners often inspect a short candidate list; moving a style into the top five has practical value even when top-1 classification remains difficult.

Text-to-image retrieval provides the clearest evidence for prompt grounding. The hybrid ranked the collection with mAP 0.2894, above label-only prompts and every visual baseline. Descriptions expand names such as Palladian or Edwardian into observable attributes, while class-score calibration makes image scores comparable. The result also supports a design principle found in grounded rationale and retrieval-explanation systems: ranked evidence is more useful when users can inspect why candidates were surfaced [27]–[29].

Prompt wording changed the geometry of the text space. Shared words in the generic template produced high inter-style similarity and weak retrieval, whereas the LLM descriptions introduced discriminative attributes. The visual-only ensemble showed that those attributes, rather than repeated class names, carried the gain. Three descriptions did not beat two at the first rank, indicating that equal averaging can dilute useful cues. Prompt selection or learned weighting is therefore preferable to unrestricted accumulation [21], [22].

HOG captured facade geometry, rooflines, bay rhythm, and openings and accounted for most of the accuracy. Color and texture improved the combined ranking for characteristic brick, stone, timber, and patterned surfaces. Grayscale conversion and blur consequently had little effect, while cropping and reflection disrupted spatial organization. The reflection result also exposes a limitation: an asymmetrical facade can change its descriptor even though its architectural style is unchanged.

Errors followed architectural relationships. Greek Revival and Palladian share classical orders; Foursquare, craftsman, Georgian, Tudor Revival, and Queen Anne overlap in domestic scale, pitched roofs, and textured facades; Chicago school, Bauhaus, and International share rectilinear grids and large windows. Negative silhouette values confirm that the categories are not visually isolated. Calibrated candidate lists therefore represent the corpus more faithfully than a geometry that assumes strict class separation.

The study is limited to 456 images, 14 to 31 examples per style, and one label per photograph even when a building contains hybrid or altered elements. The text bank is defined at class level, and the image representation is handcrafted rather than based on a large pretrained transformer. These choices keep the experiment compact and make the role of the descriptions easy to isolate, but they limit invariance to viewpoint, occlusion, and localized architectural parts. The findings therefore describe this corpus and protocol rather than every possible vision-language architecture.

Future work should learn class-specific prompt weights, target reciprocal confusions with localized facade components, and permit multi-label annotations for hybrid buildings. Pretrained CLIP-, ALIGN-, SigLIP-, or BLIP-derived image encoders [5]–[9] could replace the handcrafted descriptor while retaining the fold discipline and bidirectional evaluation. Human-facing studies could test whether short attribute rationales improve expert review,

as suggested by recent work on design consistency, critique, and structured briefs [25]–[28]. More adaptive fusion could incorporate explicit uncertainty estimates rather than a single linear weight, extending the calibration logic used here and in uncertainty-aware late fusion [30].

CONCLUSIONS

Prompt-Grounded Architectural Retrieval connected 456 building photographs with 72 LLM-generated design descriptions across 24 styles. The evaluation showed that descriptive prompts substantially improved ranking depth and text-to-image retrieval compared with style names alone. PGAR calibrated hybrid reached 62.50% Recall@5, 0.4484 MRR, and 0.2894 text-to-image mAP. Its gains over label-only prompts were statistically supported for Recall@5 and MRR, while the top-rank difference was not significant. Structural gradient features carried most of the visual signal; color and texture improved combined rankings. Ancient Egyptian, Byzantine, and Russian Revival were among the clearest styles, whereas classical and domestic revival families produced the largest reciprocal confusions. Embedding and clustering analyses confirmed substantial cross-style overlap, and perturbation tests showed stability to grayscale conversion, blur, and JPEG compression. The study demonstrates that concise language grounded in visible massing, roofs, openings, materials, and ornament is an effective retrieval interface for architectural imagery, especially when description scores are calibrated with training-fold visual evidence.

REFERENCES

[1] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, “What makes Paris look like Paris?,” ACM Trans. Graph., vol. 31, no. 4, Art. no. 101, pp. 1–9, Jul. 2012.

[2] Justin Sun, Xiaoming Xiao, and Huichao Dong, “From Tell-to-Design to Healthcare Test-Fit Constraint Checking: An LLM-Compatible Requirements-to-Constraints Framework”, JACS, vol. 3, no. 11, pp. 52–67, Nov. 2023, doi: 10.69987/JACS.2023.31105.

[3] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[4] X. Huang et al., “Urban Building Classification (UBC)—A dataset for individual building detection and classification from satellite imagery,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2022, pp. 1413–1421.

[5] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. 38th Int. Conf. Mach. Learn., PMLR, vol. 139, 2021, pp. 8748–8763.

[6] C. Jia et al., “Scaling up visual and vision-language representation learning with noisy text supervision,” in Proc. 38th Int. Conf. Mach. Learn., PMLR, vol. 139, 2021, pp. 4904–4916.

[7] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2023, pp. 11975–11986.

[8] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in Proc. 39th Int. Conf. Mach. Learn., PMLR, vol. 162, 2022, pp. 12888–12900.

[9] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in Proc. 40th Int. Conf. Mach. Learn., PMLR, vol. 202, 2023, pp. 19730–19742.

[10] A. Dosovitskiy et al., “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021.

[11] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005, vol. 1, pp. 886–893.

[12] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[13] I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” Philos. Trans. R. Soc. A, vol. 374, no. 2065, Art. no. 20150202, 2016.

[14] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, Feb. 1970.

[15] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” Ann. Eugenics, vol. 7, no. 2, pp. 179–188, 1936.

[16] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, pp. 273–297, 1995.

[17] S. Menon and C. Vondrick, “Visual classification via description from large language models,” in Proc. Int. Conf. Learn. Representations, 2023.

[18] S. Pratt, I. Covert, R. Liu, and A. Farhadi, “What does a platypus look like? Generating customized prompts for zero-shot image classification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2023, pp. 15691–15701.

[19] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” Int. J. Comput. Vis., vol. 130, no. 9, pp. 2337–2348, Sep. 2022.

[20] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 16816–16825.

[21] J. U. Allingham, J. Ren, M. W. Dusenberry, X. Gu, Y. Cui, D. Tran, J. Z. Liu, and B. Lakshminarayanan, “A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models,” in Proc. 40th Int. Conf. Mach. Learn., PMLR, vol. 202, 2023, pp. 547–568.

[22] X. Qu, G. Gou, J. Zhuang, J. Yu, K. Song, Q. Wang, Y. Li, and G. Xiong, “ProAPO: Progressively automatic prompt optimization for visual classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2025, pp. 25145–25155.

[23] T. B. Brown et al., “Language models are few-shot learners,” in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 1877–1901.

[24] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 27730–27744.

[25] Y. Chen and M. Li, “From Hand-Drawn Sketches to Interactive Web Prototypes: A Reproducible Vision-Language Approach with Structural and Visual Consistency Evaluation,” J. Technol. Informatics Eng., vol. 4, no. 2, pp. 364–384, Aug. 2025, doi: 10.51903/jtie.v4i2.490.

[26] Y. Li, S. Lu, and L. Zhao, “LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment,” Int. J. Graph. Des., vol. 3, no. 1, pp. 196–215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[27] Q. Wu, S. Meng, and J. Zhao, “Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards,” Int. J. Graph. Des., vol. 3, no. 1, pp. 216–240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[28] J. Mu, Y. Lu, and E. Hwang, “Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards,” Int. J. Graph. Des., vol. 4, no. 1, pp. 192–208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[29] Y. Li, “Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset,” J. Adv. Comput. Syst., vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[30] Q. Xin, “Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning),” J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[31] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 86, pp. 2579–2605, 2008.

[32] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.

[33] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[34] Justin Sun, Xiaoming Xiao, and Huichao Dong, “LEED-Aligned Embodied Carbon Decision Support for Hospital Expansion: An LLM-Assisted Comparison of Mass Timber, Structural Steel, and Reinforced Concrete Structural Strategies”, JACS, vol. 4, no. 12, pp. 95–109, Dec. 2024, doi: 10.69987/JACS.2024.41208.

[35] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.
[36] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.