Multimodal Late Fusion for Depth-Aware Spatial Scene Understanding

Kirthana Natarajan
Columbia University
kmn2161@columbia.edu

Trisha Maturi
Columbia University
tm3530@columbia.edu

Abstract

Depth information provides explicit geometric cues that are often missing from RGB images, making it a potentially valuable modality for indoor scene understanding. While recent vision–language models such as CLIP learn powerful semantic representations from RGB–text pairs, their effectiveness for encoding depth images and integrating geometric information remains underexplored. In this work, we investigate the role of depth embeddings in multimodal indoor scene classification using the SUNRGBD dataset. We generate spatially focused captions for RGB images using a pretrained vision–language model and extract image, text, and depth embeddings using frozen encoders. These embeddings are combined using multiple fusion architectures, including shallow and deep late fusion, gated fusion, and transformer-based fusion, and evaluated under controlled ablation settings.

In addition, we conduct a systematic analysis of CLIP as a depth encoder, examining embedding–geometry correlations, neighborhood consistency, cross-modal alignment, and retrieval performance. Our results show that CLIP depth embeddings capture coarse geometric structure and local depth similarity, despite not being trained for this modality. For scene classification, multimodal fusion improves performance over unimodal baselines, with image–text combinations providing the largest gains and depth contributing modest improvements in specific settings. However, increased architectural complexity does not consistently yield better performance, and transformer-based fusion underperforms under limited data. Overall, our findings highlight both the potential and limitations of using CLIP-based depth embeddings for multimodal scene understanding. The code can be accessed here: https://github.com/kmn01/DepthAware.

1 Introduction

Recent advances in deep learning have led to powerful visual representations that capture rich semantic information from RGB images. Models trained on large-scale datasets excel at encoding appearance-based cues such as color, texture, lighting, and object identity. However, RGB images provide only a two-dimensional projection of the visual world; this can obscure important information about a scene's three-dimensional structure. As a result, representations based solely on RGB often struggle to distinguish scenes that are visually similar but differ in their underlying spatial layout.

Depth information provides a complementary modality by explicitly encoding geometric structure, including object distance, spatial arrangement, and overall scene layout. Unlike RGB, depth maps are largely invariant to lighting conditions and color variations, instead emphasizing 3D geometry. This distinction is especially important for indoor scene understanding, where spatial relationships frequently define scene categories more strongly than appearance alone. For example, bedrooms

Submission for 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

and living rooms may share similar objects and textures yet differ in their geometric configurations, while hallways and offices can exhibit comparable color palettes but possess distinct depth profiles due to differences in room shape and scale. In such cases, depth supplies information that is difficult or unreliable to infer from RGB alone.

The importance of geometric reasoning is well established in domains such as robotics and autonomous systems, where understanding spatial layout is essential for navigation, manipulation, and obstacle avoidance. Indoor robots rely on depth sensing to identify free space, estimate distances, and reason about traversability, while autonomous driving systems use depth cues for tasks such as object localization and motion planning, particularly under challenging visual conditions. Depth information has also become increasingly relevant in visual reasoning and multimodal tasks, including visual question answering, where questions involving relative position, occlusion, or distance cannot be answered reliably using appearance cues alone.

Motivated by these observations, this work investigates whether depth embeddings can improve indoor scene classification when combined with RGB embeddings. Using paired RGB and depth data from standard RGB-D datasets, we systematically compare RGB-only, depth-only, and fused RGB–depth representations within a unified classification framework. By evaluating different fusion strategies and analyzing class-level performance, we aim to better understand when and how depth contributes complementary information beyond strong appearance-based representations.

2 Related Works

Recent work has explored RGB-D contrastive learning to jointly model appearance and geometry. Methods such as P4Contrast (5) align point–pixel pairs to learn unified geometric and visual representations, while other approaches (9) employ pixel-level contrastive objectives to integrate cross-modal cues for RGB-D scene and object understanding. Additional work (2) introduces spatial priors into contrastive frameworks to better capture depth-aware relationships. While these methods demonstrate the effectiveness of coupling depth and appearance features, they are primarily limited to visual modalities and do not incorporate language-level semantics.

In parallel, vision–language models (VLMs) such as CLIP (6) and BLIP-2 (4) have shown strong generalization by aligning image and text representations through large-scale contrastive learning. However, these models operate on two-dimensional imagery and do not explicitly encode geometric structure or spatial relationships.

Despite these advances, the extent to which pretrained VLMs, particularly CLIP, can meaningfully encode depth images remains underexplored. Existing work has largely focused on enhancing CLIP with additional spatial supervision, rather than systematically evaluating its ability to represent geometric information when applied directly to depth data.

Building on prior work in RGB-D learning and vision–language representation, this work investigates whether CLIP can serve as an effective depth encoder for indoor scene understanding. By analyzing depth embeddings extracted using CLIP and comparing multiple fusion architectures for RGB, depth, and text, we aim to clarify the role of depth information in multimodal classification and assess when geometric cues provide complementary benefits over appearance and language alone.

3 Methodology

Figure 1: **Multimodal Pipeline.** End-to-end workflow for multimodal feature extraction, fusion, and classification.

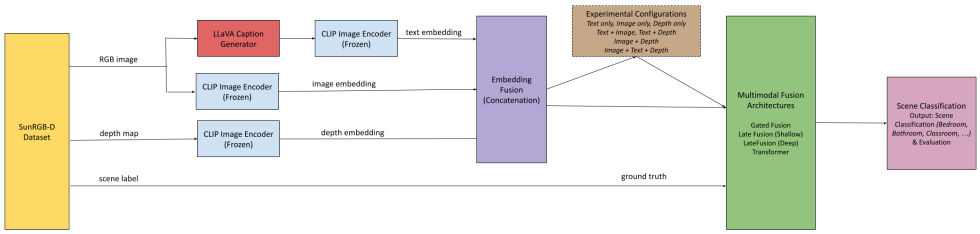


Table 1: **Overview of the proposed pipeline.** Stages of the proposed method, associated tools, and their purposes.

Stage	Tool	Purpose
1. Caption Generation	LLaVA	Generate spatially focused textual descriptions of RGB images
2. Embedding Extraction	CLIP Encoder	Encode RGB images, depth maps, and captions into fixed-dimensional embeddings.
3. Fusion & Classification	MLP / Transformer	Combine selected modalities and predict indoor scene class.
4. Evaluation	-	Measure accuracy and per-class performance across fusion strategies.

3.1 Dataset

We use the SUNRGBD dataset (8), which provides aligned RGB images, depth maps, and ground-truth indoor scene labels. From the dataset, we select images with valid depth information and corresponding scene classification labels (e.g., bedroom, bathroom, classroom). All samples are preprocessed to ensure consistent input formatting across modalities.

3.2 Caption Generation

To incorporate textual information describing scene layout, we generate captions for each RGB image using a pretrained LLaVA vision–language model. Captions are produced using the following prompt:

> "Describe the scene in one sentence, focusing on the relative positions of objects and background elements. Avoid repetition and unnecessary detail."

This prompt is designed to emphasize spatial relationships and scene structure rather than object enumeration. Each image is associated with a single generated caption, which is later embedded and used as an additional modality for classification.

3.3 Embedding Extraction

For each sample, we extract embeddings from three modalities using CLIP (analysis in 4.1):

- Text embedding: computed from the generated caption using a pretrained text encoder.
- Image embedding: computed from the RGB image using a pretrained image encoder.
- Depth embedding: computed from the depth map, which is processed and encoded using the same embedding pipeline as the RGB images.

All encoders are kept frozen during training to isolate the effect of multimodal representation fusion. The resulting embeddings are concatenated as needed depending on the experimental configuration.

3.4 Fusion Architectures

To evaluate how different fusion strategies impact multimodal scene classification, we experiment with four classification models that combine image, depth, and text embeddings in different ways. In all cases, the input embeddings are extracted using frozen encoders, and only the fusion and classification components are trained.

3.4.1 Gated Fusion

The Gated Fusion model applies learnable modality-specific gates to the text, image, and depth embeddings prior to fusion. Each modality is scaled by a learned linear transformation followed by a soft weighting mechanism, allowing the model to emphasize or suppress individual modalities during training. The gated embeddings are concatenated and passed to a shallow multilayer perceptron for classification.

$$g_i = \sigma(W_i x_i + b_i), \qquad g_d = \sigma(W_d x_d + b_d), \quad g_t = \sigma(W_t x_t + b_t), \tag{1}$$

$$x_{\text{fused}} = [g_i \cdot x_i,\ g_d \cdot x_d,\ g_t \cdot x_t], \tag{2}$$

$$x_{\text{norm}} = \frac{x_{\text{fused}}}{\|x_{\text{fused}}\|_2}, \tag{3}$$

$$y = W_2\,\text{ReLU}(W_1 x_{\text{norm}} + b_1) + b_2, \tag{4}$$

$$\text{gates} = \text{softmax}([W_i x_i, W_d x_d, W_t x_t]) \tag{5}$$

This design enables the model to adaptively control the relative contribution of each modality while maintaining a simple fusion structure.

3.4.2 Late Fusion (Shallow)

The Late Fusion (Shallow) model concatenates the image, depth, and text embeddings into a single feature vector, which is then passed through a lightweight MLP consisting of a single hidden layer. This architecture serves as a minimal late-fusion baseline, allowing us to assess whether simple concatenation is sufficient to leverage multimodal information.

$$x_{\text{fused}} = \text{concat}(x_i, x_d, x_t), \tag{6}$$

$$x_{\text{norm}} = \frac{x_{\text{fused}}}{\|x_{\text{fused}}\|_2}, \tag{7}$$

$$y = W_2\,\text{ReLU}(W_1 x_{\text{norm}} + b_1) + b_2 \tag{8}$$

3.4.3 Late Fusion (Deep)

The Late Fusion (Deep) model extends the shallow late-fusion approach by using a deeper MLP with multiple fully connected layers, batch normalization, and nonlinear activations. By increasing model capacity, this architecture evaluates whether deeper interactions among concatenated embeddings improve classification performance compared to simpler fusion strategies.

$$x_{\text{fused}} = \text{concat}(x_i, x_d, x_t), \tag{9}$$

$$x_{\text{norm}} = \frac{x_{\text{fused}}}{\|x_{\text{fused}}\|_2}, \tag{10}$$

$$h_1 = \text{ReLU}\big(\text{BN}_1(W_1 x_{\text{norm}} + b_1)\big), \tag{11}$$

$$h_2 = \text{ReLU}\big(\text{BN}_2(W_2 h_1 + b_2)\big), \tag{12}$$

$$y = W_3\,\text{Dropout}(h_2) + b_3 \tag{13}$$

3.4.4 Transformer-Based Fusion

The Transformer Fusion model treats each modality embedding as a token and applies a transformer encoder to model interactions between modalities. Image, depth, and text embeddings are first projected into a shared latent space and augmented with modality-specific embeddings. A learnable classification token is prepended to the token sequence, and the transformer encoder captures cross-modal relationships via self-attention.

$$g = \text{softmax}([g_i, g_d, g_t]), \tag{14}$$

$$z_i = g_0 \cdot f_i(x_i), \qquad z_d = g_1 \cdot f_d(x_d), \qquad z_t = g_2 \cdot f_t(x_t), \tag{15}$$

$$z_{\text{cls}} \text{ is a learnable CLS token}, \tag{16}$$

$$\text{tokens} = [z_{\text{cls}},\ z_i + e_i,\ z_d + e_d,\ z_t + e_t], \tag{17}$$

$$Z_{\text{encoded}} = \text{TransformerEncoder}(\text{tokens}), \tag{18}$$

$$z_{\text{pooled}} = \frac{1}{4} \sum_{k=0}^{3} Z_{\text{encoded},k}, \tag{19}$$

$$y = W_{\text{cls}} z_{\text{pooled}} + b_{\text{cls}} \tag{20}$$

In addition, this model incorporates learnable modality gates, allowing the network to dynamically reweight each modality before attention-based fusion. The final representation is obtained by pooling over the transformer outputs and passed to a linear classification head.

This architecture enables explicit modeling of higher-order interactions between modalities and provides interpretability through learned modality weights.

3.5 Experimental Comparison

All four fusion models are evaluated under identical training and evaluation settings. By comparing performance across these architectures, we analyze the trade-offs between model complexity, fusion strategy, and classification accuracy, with particular focus on the contribution of depth embeddings in multimodal settings.

We evaluate the contribution of each modality by training and testing the classifier under the following input configurations:

- Text only
- Image only
- Depth only
- Text + Image
- Text + Depth
- Image + Depth
- Text + Image + Depth

4 Evaluating CLIP Depth Embeddings

4.1 CLIP Depth Embedding Analysis

While CLIP is primarily trained on RGB images and natural language, it is increasingly used as a general-purpose visual embedding model. In this work, CLIP is used to encode depth maps by treating them as image-like inputs. To assess whether this choice is reasonable, we evaluate whether CLIP depth embeddings preserve geometric similarity, organize depth structure meaningfully in latent space, and align with RGB embeddings in a geometry-aware manner.

4.1.1 Pairwise Depth–Embedding Correlation

Do depth-similar images have similar CLIP depth embeddings?

We observed a weak but consistent positive correlation with embedding similarity (Pearson $r = 0.22$, Spearman $\rho = 0.24$, $p \ll 10^{-40}$). Depth histogram similarity yielded similar but slightly lower correlations (Pearson $r = 0.20$, Spearman $\rho = 0.20$, $p \ll 10^{-40}$). The higher Spearman coefficients suggest a monotonic but non-linear relationship, suggesting that depth accounts for only a limited portion of the variance in CLIP depth embeddings.

4.1.2 CLIP Neighborhood Depth Consistency

Are nearest neighbors in CLIP space geometrically similar?

We measured the depth histogram similarity between each image and its nearest neighbors in CLIP space. Despite weak global correlations, local neighborhoods in CLIP space are depth-consistent. Nearest neighbors exhibit a very high depth histogram similarity (mean of 0.98), indicating that CLIP embeddings are locally consistent with respect to coarse geometric structure.

4.1.3 Depth Bucket Retrieval Accuracy

Can CLIP embeddings separate scenes by coarse depth regime?

To test coarse depth separability, we perform retrieval over four depth buckets (chance accuracy ≈ 0.25). CLIP embeddings achieved an accuracy of 0.49, nearly doubling chance performance. This result indicates that while depth alone does not strongly govern the global structure of the embedding space, CLIP representations retain sufficient depth information to reliably distinguish broad depth categories.

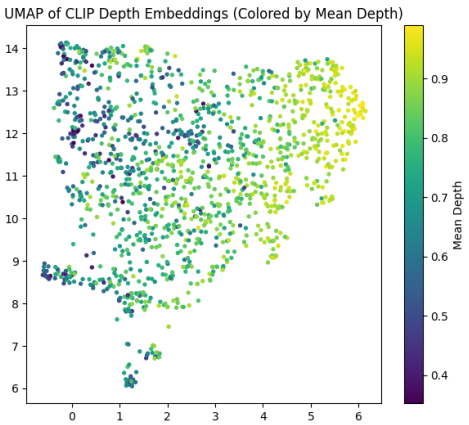
4.1.4 Label-Conditioned Depth Similarity Analysis

Is depth similarity driven by geometry or semantics?

To assess whether depth similarity in CLIP embeddings is driven by scene semantics or geometric structure, we computed depth similarity metrics conditioned on semantic labels. Across 11 labels with sufficient samples, mean Pearson correlations for global depth statistics and depth histogram similarity were 0.26 and 0.22, with corresponding Spearman correlations of 0.30 and 0.29. These values are only marginally higher than random-label baselines (Pearson 0.25 and 0.23), indicating that semantic labels explain little of the overall depth similarity. Although some scene categories exhibit notably higher correlations (maximum Pearson ~ 0.45), suggesting label-specific amplification effects, depth similarity in CLIP is largely driven by geometric rather than semantic factors.

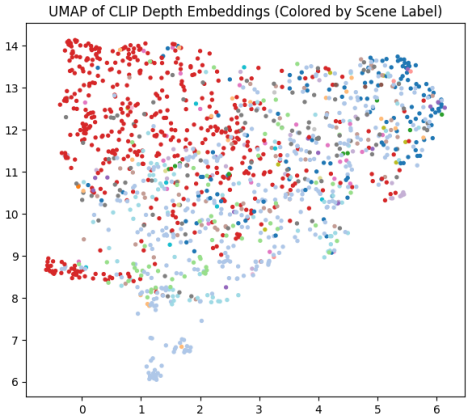
4.2 CLIP Embedding Space Visualization (UMAP and t-SNE)

1. Global Depth Sensitivity



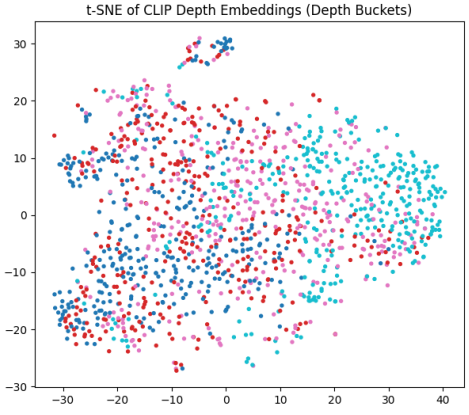
We first analyze whether CLIP depth embeddings reflect global depth statistics of the input scenes. Using UMAP projections of CLIP embeddings, colored by mean scene depth, we observe a smooth and continuous gradient across the embedding space. Scenes with similar average depth tend to be embedded nearby, while progressively deeper scenes are distributed along a dominant axis of variation. This behavior is consistent across multiple random seeds and dimensionality reduction settings.
Quantitatively, we find weak-to-moderate but statistically significant correlations between pairwise embedding similarity and depth similarity (Spearman $\rho = 0.22$). These results indicate that CLIP embeddings encode coarse depth-related information, particularly global scene scale and near–far layout, despite the absence of explicit depth supervision during training.

2. Semantic Dominance over Geometric Structure



To differentiate geometric depth, we further visualize the same embedding space colored by scene category labels. In this setting, embeddings exhibit clearer organization by scene semantics than by depth alone. While certain regions of the space correspond to specific scene types (e.g., corridors versus open rooms), substantial overlap remains across categories.
This observation suggests that depth information in CLIP embeddings is largely mediated by semantic cues. Scene types that are naturally associated with characteristic depth distributions (e.g., hallways, outdoor scenes) form loose clusters, and depth gradients emerge implicitly through this semantic organization rather than through explicit geometric encoding.

3. Lack of Discrete Depth Separability



Finally, we assess whether CLIP embeddings support separability across discretized depth regimes. Using t-SNE visualizations colored by coarse depth buckets, we observe extensive overlap between depth bins, with no clear cluster boundaries. While a gradual transition from shallow to deep scenes is apparent, depth categories are not cleanly separable in the embedding space.
This behavior contrasts with embeddings produced by depth-specialized models (e.g., DPT, Depth-MAE, DepthAnything), which exhibit stronger alignment with geometric depth measures. The lack of discrete depth clustering in CLIP embeddings indicates that depth is not represented as an independent latent factor, but rather as a continuous, secondary attribute.

4.3 Statistical Significance Testing

Robustness of Depth–Embedding Correlations

We evaluate the robustness of the relationship between CLIP depth embeddings and geometric depth statistics using bootstrap and permutation tests. A Spearman correlation with 1,000 bootstrap resamples yields a mean correlation of $\rho = 0.221$ with a 95% confidence interval of [0.186, 0.257]. The observed correlation lies well within this interval, indicating a stable association rather than an artifact of sampling variability.

To assess statistical significance, we conduct permutation testing by randomly permuting depth histograms across samples. This results in a p-value of 0.0, strongly rejecting the null hypothesis of no relationship. A within-label permutation test that controls for scene category remains significant ($\rho \approx 0.166$, $p = 0.0$), suggesting that CLIP embeddings encode depth-related information beyond semantic scene labels, albeit with a modest effect size.

Cross-Modal Alignment Between RGB and Depth Embeddings

We examine whether depth embeddings align with their corresponding RGB embeddings in CLIP's joint embedding space. Cosine similarity between paired RGB–depth embeddings (mean = 0.612, $\sigma = 0.063$) is higher than for randomly matched pairs (0.579). This difference is highly significant under a two-sided t-test ($p \approx 8.4 \times 10^{-64}$).

These results indicate that CLIP learns a non-trivial cross-modal alignment between RGB and depth representations despite depth not being an explicit training modality. However, the modest gap between paired and random similarities suggests that this alignment is coarse and likely driven by shared semantic content rather than precise geometric correspondence.

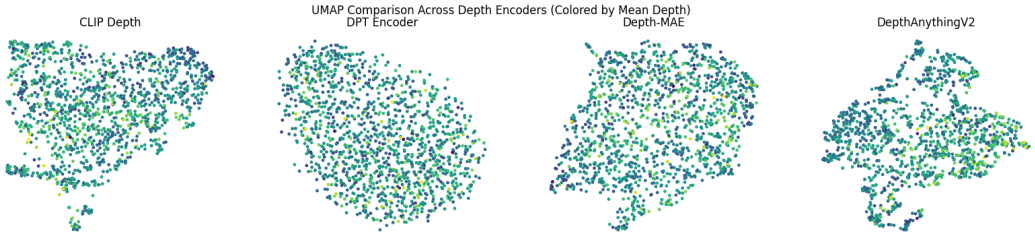
Cross-Modal Retrieval and Geometry-Aware Control

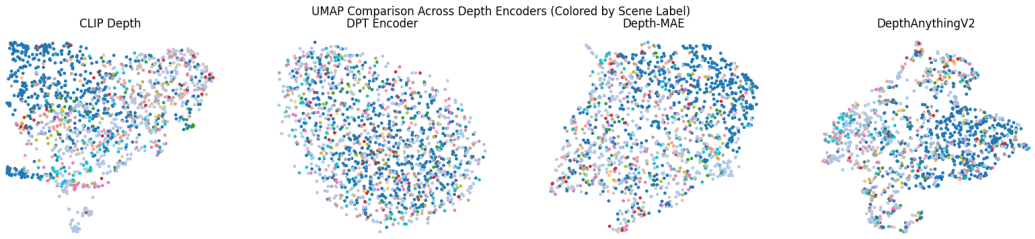
To further assess RGB–depth alignment, we evaluate cross-modal retrieval by querying depth embeddings with RGB embeddings. CLIP achieves a Recall@5 of 5.3%, only slightly above chance, indicating weak instance-level alignment.

When retrieval is restricted to samples with similar depth statistics, Recall@5 increases to 10.5%. Although still low in absolute terms, this improvement suggests that coarse geometric similarity aids cross-modal alignment. Overall, retrieval performance remains limited, supporting the conclusion that CLIP does not encode depth as a distinct or discriminative latent dimension.

4.4 Comparing UMAPs Across Depth Encoders

We evaluate four depth encoding approaches: CLIP, DPT (MiDaS)(1), Depth-MAE, and DepthAnythingV2(11). For each encoder, depth maps are embedded into a fixed-dimensional representation space. We analyze the resulting embeddings using cosine similarity to assess representation consistency, separability, and stability across samples.





We compare CLIP's embedding structure to depth-native models using UMAP visualizations colored by mean depth and scene label. This helps assess whether CLIP organizes depth information in a way similar to models trained explicitly on geometry.

When colored by mean depth 4.4, CLIP embeddings show a smooth but weak depth gradient, suggesting sensitivity to coarse depth cues without strong geometric ordering. Depth-MAE displays a clearer and more structured depth progression, indicating stronger alignment between embedding similarity and geometric depth. DepthAnythingV2 forms distinct clusters but lacks a consistent depth gradient, reflecting its focus on pixel-level depth prediction rather than global embedding structure.

These patterns are supported by Spearman correlations between embedding similarity and depth histogram similarity. CLIP shows a moderate correlation ($\rho = 0.229$), consistent with implicit depth sensitivity. Depth-MAE achieves a higher correlation ($\rho = 0.344$), indicating more explicit encoding of depth relationships. In contrast, DepthAnythingV2 shows near-zero correlation ($\rho = -0.034$), suggesting that its global embeddings do not preserve depth similarity despite strong depth estimation performance.

Coloring UMAPs by scene label 4.4 highlights differences in semantic influence. CLIP embeddings are strongly organized by scene category, with depth variation appearing mainly within semantic groups. Depth-MAE shows weaker semantic dominance and more evenly distributed depth structure. DepthAnythingV2 exhibits strong semantic clustering but little alignment with depth structure.

Overall, CLIP captures depth indirectly through semantic organization, leading to moderate but meaningful depth correlations. Depth-MAE learns embeddings that better reflect geometric depth, while DepthAnythingV2 prioritizes accurate depth prediction over geometry-aware embedding structure. This shows that strong depth estimation alone does not guarantee depth-aware embeddings and motivates our approach to explicitly align geometric structure with semantic representations.

5 Results

Table 2: **Classification Accuracy.** Accuracies for different modality combinations (image, text, depth) across four fusion architectures.

Input	**Gated Fusion**	**Late Fusion Shallow**	**Late Fusion Deep**	**Transformer Based Fusion**
Text Only	0.758	0.747	0.675	0.328
Image Only	0.877	0.870	0.826	0.328
Depth Only	0.386	0.375	0.346	0.328
Text and Image	0.909	0.902	0.891	0.328
Text and Depth	0.797	0.779	0.725	0.328
Image and Depth	0.873	0.870	0.851	0.328
Image, Text, and Depth	0.906	0.906	0.916	0.328

Table 3: **Performance Comparison of Fusion Architectures.** Overall classification performance using text, image, and depth embeddings.

Input	Gated Fusion	Late Fusion Shallow	Late Fusion Deep	Transformer Based Fusion
Accuracy	0.91	0.91	0.92	0.33
Macro Precision	0.59	0.63	0.72	0.02
Macro Recall	0.60	0.60	0.66	0.06
Macro F1	0.59	0.60	0.68	0.03
Weighted F1	0.88	0.88	0.90	0.16

5.1 Ablation Across Modalities

Table 2 reports classification accuracy across different input modality combinations and fusion architectures. Several consistent trends emerge.

5.1.1 Unimodality

First, RGB image embeddings provide the strongest unimodal signal, achieving accuracies of minimum 0.8 across all non-transformer architectures. In contrast, depth-only embeddings perform poorly, indicating that depth information alone is insufficient for reliable indoor scene classification. Text-only inputs achieve moderate performance, suggesting that spatially focused captions capture useful semantic cues but lack the full discriminative power of visual features.

5.1.2 Multimodality

Second, multimodal fusion improves performance over unimodal baselines, particularly when combining image and text embeddings. Across all three non-transformer architectures, the Text + Image configuration yields the highest or near-highest accuracy, with Gated Fusion achieving 0.909 accuracy. This suggests that textual descriptions generated by LLaVA provide complementary high-level semantic information that augments appearance-based image embeddings.

Notably, adding depth embeddings does not consistently improve performance beyond image–text fusion. While depth contributes modest gains in some settings (e.g., Text + Depth vs. Text only), the Image + Depth configuration performs similarly to Image only, indicating that depth information may be partially redundant when strong RGB embeddings are available. However, the Text + Image + Depth configuration achieves the highest overall accuracy (0.916) under most model, suggesting that depth can provide incremental benefits when combined with both visual and textual context.

5.2 Comparison of Fusion Architectures

Table 3 summarizes overall performance metrics for the four fusion architectures using all three modalities.

Among the non-transformer models, Late Fusion Deep achieves the best overall performance, with the highest accuracy (0.92), macro F1 (0.68), and weighted F1 (0.90). This indicates that increasing model capacity enables better learning of cross-modal interactions after embedding concatenation. The shallow late-fusion model and gated fusion model perform similarly in terms of accuracy, though the gated model shows slightly lower macro precision and recall, suggesting limited benefits from simple gating mechanisms.

The Transformer-Based Fusion model performs substantially worse, with accuracy close to random chance (0.33) and near-zero macro precision and recall. This suggests that the transformer architecture was unable to effectively learn cross-modal relationships under the current experimental setup. Possible contributing factors include the limited dataset size, insufficient training signal for attention-based fusion, or suboptimal hyperparameter or model architecture choices. These results highlight that more complex fusion architectures do not necessarily yield improved performance, particularly when training data is limited and embeddings are frozen.

5.3 Implications for Depth Embeddings

Overall, the results indicate that depth embeddings provide very limited standalone value, but can offer modest complementary benefits in multimodal settings. Depth contributes most effectively when paired with both RGB and text embeddings and processed by higher-capacity classifiers. These findings suggest that while geometric information is relevant for indoor scene understanding, its utility depends strongly on the fusion strategy and the presence of complementary semantic signals.

5.4 Comparison with ResNet50 Baseline

To contextualize the performance of our multimodal fusion models, we compare them with a standard ResNet-50 model trained on RGB images only. The ResNet-50 baseline achieved a training loss of 0.2864 with 93.85% training accuracy, and a validation loss of 0.5758 with 84.84% validation accuracy. Incorporating the additional modalities of text and depth through our fusion architectures improves overall classification accuracy. Our approach achieves 92% accuracy for the Late Fusion Deep model, exceeding the ResNet-50 validation accuracy by over 6%, highlighting the value of semantic and geometric information for indoor scene classification.

6 Discussion

6.1 Limitations

All modality encoders are kept frozen, which isolates the effect of fusion strategies but may limit the ability of depth embeddings to adapt to the classification task, but was implemented in this project due to resource and compute constraints. End-to-end fine-tuning could improve cross-modal alignment, particularly for depth representations. Additionally, the transformer-based fusion model may require larger datasets or different training regimes to fully leverage attention-based cross-modal interactions.

6.2 Future Work

While this work focused on indoor scene classification, our findings suggest that depth information may be more valuable in tasks that require explicit spatial reasoning. A natural next step is to extend this analysis to visual question answering (VQA), where questions often involve object relationships, relative position, distance, and occlusion. In such cases, depth can provide geometric cues that are difficult to infer from RGB images alone. When combined with image and caption embeddings, depth has the potential to ground language queries in the three-dimensional structure of a scene, improving performance on geometry-aware and relational questions.

Future work will explore multimodal VQA settings where image, caption, and depth embeddings are jointly used to answer questions that depend on spatial layout. This includes evaluating depth-aware fusion strategies for questions involving "in front of," "behind," "near," or "far," and analyzing whether depth contributes more strongly under specific question types. Incorporating depth-native encoders or fine-tuning vision–language models with depth supervision may further improve geometric reasoning in these tasks.

In addition, we conducted preliminary experiments on 3D object detection using image, caption, and depth embeddings. A promising direction for future work is to apply multimodal fusion architectures to jointly combine these embeddings for enhanced object detection. By integrating depth and textual cues alongside RGB images, such models could better capture spatial relationships, object geometry, and context, improving detection and localization in cluttered indoor scenes. This approach would allow the network to leverage complementary information from each modality, enabling more accurate and robust 3D object recognition.

Overall, these directions aim to better leverage depth as a complementary signal for language-guided visual reasoning, moving beyond scene-level classification toward more geometry-aware multimodal understanding.

6.3 Contribution

Both team members contributed equally to all aspects of this project, including dataset preparation, caption generation, embedding extraction, model implementation, experimental evaluation, result analysis, and writing the paper.

7 Conclusions

This work studies whether depth information can improve indoor scene classification when combined with RGB images and text, using CLIP as a shared embedding model. We first analyzed how well CLIP represents depth and then evaluated the effect of depth in different multimodal fusion architectures.

Our analysis shows that CLIP captures depth only implicitly. Depth embeddings contain weak but statistically significant geometric information and preserve coarse depth similarity, but depth is not a dominant or separable factor in CLIP's embedding space. Instead, depth cues are largely tied to scene semantics. Compared to depth-native models, CLIP is less effective at organizing representations by geometric structure.

These properties are reflected in downstream results. RGB embeddings provide the strongest unimodal performance, while depth alone performs poorly. Combining image and text yields the largest gains, showing that textual descriptions add useful semantic context. Adding depth provides small but consistent improvements only when image and text are both present, suggesting that depth contributes complementary geometric cues rather than acting as a primary signal. Among fusion methods, deeper late-fusion models perform best, while transformer-based fusion struggles under limited data and frozen encoders.

Overall, depth information can help indoor scene classification, but its impact is modest when encoded through CLIP.

References

[1] BIRKL, R., WOFK, D., AND MÜLLER, M. Midas v3.1 – a model zoo for robust monocular relative depth estimation, 2023.

[2] CHEN, H., CHEN, Z., WU, Y., AND CHEN, H. Spatial-aware multi-modal contrastive learning for rgb-d salient object detection and beyond. *Information Fusion 124* (2025), 103362.

[3] JANOCH, A., KARAYEV, S., JIA, Y., BARRON, J. T., FRITZ, M., SAENKO, K., AND DARRELL, T. A category-level 3d object dataset: Putting the kinect to work. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (2011).

[4] LI, J., LI, D., SAVARESE, S., AND HOI, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[5] LIU, Y., YI, L., ZHANG, S., FAN, Q., FUNKHOUSER, T., AND DONG, H. P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding, 2020.

[6] RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., KRUEGER, G., AND SUTSKEVER, I. Learning transferable visual models from natural language supervision, 2021.

[7] SILBERMAN, N., HOIEM, D., KOHLI, P., AND FERGUS, R. Indoor segmentation and support inference from rgb-d images. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2012).

[8] SONG, S., LICHTENBERG, S., AND XIAO, J. Sun rgb-d: A rgb-d scene understanding benchmark suite, 2015.

[9] WU, J., HAO, F., LIANG, W., AND XU, J. Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection. *IEEE Transactions on Multimedia 26* (2024), 1011–1026.

[10] XIAO, J., OWENS, A., AND TORRALBA, A. Sun3d: A database of big spaces reconstructed using SfM and object labels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2013).

[11] YANG, L., KANG, B., HUANG, Z., ZHAO, Z., XU, X., FENG, J., AND ZHAO, H. Depth anything v2, 2024.