mathematics

MDPI

Article

Interior Design Evaluation Based on Deep Learning: A Multi-Modal Fusion Evaluation Mechanism

Yiyan Fan [1], Yang Zhou [2,*] and Zheng Yuan [1]

[1] Shanghai Academy of Fine Arts, Shanghai University, Shanghai 200444, China; yiyanfn@gmail.com (Y.F.); 3267154304@shu.edu.cn (Z.Y.)

[2] School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200444, China

* Correspondence: zhouyang0410@shu.edu.cn

Abstract: The design of 3D scenes is of great significance, and one of the crucial areas is interior scene design. This study not only pertains to the living environment of individuals but also has applications in the design and development of virtual environments. Previous work on indoor scenes has focused on understanding and editing existing indoor scenes, such as scene reconstruction, segmentation tasks, texture, object localization, and rendering. In this study, we propose a novel task in the realm of indoor scene comprehension, amalgamating interior design principles with professional evaluation criteria: 3D indoor scene design assessment. Furthermore, we propose an approach using a transformer encoder–decoder architecture and a dual-graph convolutional network. Our approach facilitates users in posing text-based inquiries; accepts input in two modalities, point cloud representations of indoor scenes and textual queries; and ultimately generates a probability distribution indicating positive, neutral, and negative assessments of interior design. The proposed method uses separately pre-trained modules, including a 3D visual question-answering module and a dual-graph convolutional network for identifying emotional tendencies of text.

Keywords: interior design; 3D question answering; dual-graph convolutional networks; sentiment analysis

MSC: 68T07

check for updates

Citation: Fan, Y.; Zhou, Y.; Yuan, Z. Interior Design Evaluation Based on Deep Learning: A Multi-Modal Fusion Evaluation Mechanism. *Mathematics* **2024**, *12*, 1560. https://doi.org/10.3390/math12101560

Academic Editors: Fangfei Li and Jiapeng Xu

Received: 19 April 2024
Revised: 6 May 2024
Accepted: 8 May 2024
Published: 16 May 2024

1. Introduction

Interior design is a significant professional field. A comfortable indoor space holds crucial significance for both living and working environments. Assessment of interior design is not only a key goal in evaluating the comfort of living environments in the real world but is also crucial for virtual scenes, such as the layout of movie scenes and the construction of game scenes. In film production, the design and arrangement of scenes can convey different visual messages, thereby affecting the narrative and the audience's emotional experience. In game development, game artists can design virtual scenes to improve players' gaming experience and visual richness, thereby enhancing the attractiveness and entertainment of the game. Therefore, the need for interior design evaluation is not limited to the actual living environment but also extends to the creation of virtual scenes.

Considering the inherent characteristics of the task, comparable methods for visual evaluation tasks [1] face limitations imposed by the 2D format of image data within predefined scoring parameters, resulting in assessments confined to single-score analyses. Recent investigations centered on the visual assessment of indoor environments have witnessed a notable trend towards embracing multi-modal architectures aimed at seamlessly integrating visual data, predominantly images, with pertinent textual content. In such scenarios, textual cues often serve as supplementary information to enrich the interpretation of visual data [2,3], enabled by modal fusion techniques employing transformers. Recent scholarly inquiries have also seen the emergence of neuro-aesthetic estimators applied to three-dimensional settings [2]. Nevertheless, in these studies, the foundational approach continues to rely on virtual camera systems capturing 2D image frames from 3D scenes to conduct aesthetic evaluations, rather than directly evaluating within the 3D model itself. Furthermore, these investigations draw upon pre-existing textual knowledge to varying extents.

Previous studies on 3D visual question answering (3DVQA) have predominantly relied on the utilization of multi-modal features to tackle prediction tasks across predefined answer categories. Several seminal investigations have introduced foundational models, such as ScanQA [4], which utilizes a dual-branch architecture incorporating VoteNe and BiLSTM to address point cloud scenes and textual queries separately. In addition, CLIP-guide [5], integrates techniques for knowledge transfer. Nonetheless, these foundational approaches regard 3DVQA as a predictive task, thus being constrained by the predefined array of possible answers. Consequently, interactions between these models and human users often produce relatively simplistic feedback. In a recent groundbreaking development, Dwedari et al. [6] introduced an innovative transformer-based architectural framework specifically designed to generate open-ended responses to user inquiries regarding 3D scenes. This pioneering approach leverages language rewards throughout the model training process and incorporates reinforcement learning techniques to ensure that the generated sentences encompass comprehensive semantic content.

In such a scenario, sentiment analysis has attracted considerable attention [7,8]. It delves into the examination of natural language texts, extracting subjective information such as emotions and attitudes. Present sentiment analysis tasks commonly rely on deep neural networks like RNN and LSTM, paired with numerous text datasets to train versatile sentiment analysis models with robust generalization capabilities. A specialized area within this domain is aspect-based sentiment analysis (ABSA), which is particularly oriented towards finely detecting sentiment at the level of individual entities [7]. Our research utilizes a dual-graph convolutional network (DualGCN) architecture, which consists of an integrated semantic correlation-based graph convolutional network (SemGCN) and a syntax-based graph convolutional network (SynGCN), both facilitated by the shared BiAffine module. The motivation for the architectural choice stems from the desire to comprehensively consider both the syntactic structure and semantic relevance present in a given sentence. Compared with earlier methods oriented towards sentence-level and document-level sentiments, this method shows greater generality and applicability for analyzing short text paragraphs.

The collaborative interaction between the Gen3DQA and DualGCN modules forms the bulk of our overall framework. After embedding the user's textual questions and 3D point cloud scene information, we obtain vector spaces for both visual and textual data. These spaces are then fused through the transformer encoder to generate the fusion vector space. The fused vector space is passed through the transformer decoder to generate relevant descriptions based on the 3D scene. In a subsequent stage, these textual answers are fed into the ABSA module to identify the underlying aspect-based sentiments within them. This collaborative architecture facilitates the evaluation of 3D scenes based on user interests and evaluates the quality, atmosphere, style, etc., of interior design by detecting the fine-grained differences in emotion contained in the generated text descriptions. The combination of these modules provides a more complex and detailed assessment process. To conclude, our contributions are as follows: (1) We introduce a novel task: evaluating interior designs in 3D environments. (2) We propose a framework for this task, leveraging the transformer encoder–decoder architecture and DualGCN, to facilitate the evaluation of 3D indoor scene designs. (3) Our approach promotes user participation by allowing users to interact with the system by querying specific aspects of the 3D scene through text input.

2. Related Work

Significant traction in cross-modal tasks within both domains has been facilitated by the rapid advancements in natural language processing and computer vision, which have intersected to a great extent. Widespread attention has been garnered by tasks such as image captioning, visual question answering, and image retrieval. In the realm of 3D scene analysis, recent research has focused on modal fusion techniques. These techniques aim to integrate scene images with accompanying descriptive text, aligning them in a vector space to improve subsequent task performance. However, existing aesthetic estimators tailored for 3D scenes primarily utilize 2D image frame inputs, neglecting textual information. To address this research gap, we recognized the need for and developed a novel framework. This framework combines a context-aware, free-form answer generation system for questions within 3D scene models with a fine-grained sentiment analysis model focused on entity-level aspects. The goal is to conduct comprehensive evaluations of 3D interior scene designs across various semantic dimensions.

2.1. Visual Question Answering (VQA)

VQA was first introduced by Malinowski et al. [9] and usually combines images and natural language. The visual question-answering task involves presenting an image alongside a corresponding question, to which the visual question-answering model provides a prediction for the answer. Visual question answering (VQA) presents greater challenges compared to other cross-modal tasks due to its demand for nuanced semantic comprehension of both image and text data, as well as the necessity for visual reasoning to accurately anticipate the correct response. The most arduous aspect of the visual question-answering task lies in achieving holistic semantic alignment across multimodal inputs. The approach employing global feature fusion initially employs a conventional CNN to extract global image features, utilizes an RNN to extract textual features, and subsequently merges these disparate global features to derive fused features for answer prediction. This approach may fail to effectively capture the local correlations or contextual information between the image and text, potentially impacting the accuracy and diversity of the final answers. Moreover, the use of traditional CNN and RNN models to extract global features may be constrained by their ability to handle long sequences or complex image structures, thereby limiting the model's performance and applicability. Many important previous works on VQA for 2D images treat this task as a classification problem, where answers are generated from a predefined answer space [10–12]. However, treating VQA for 2D images as a classification problem from a predefined answer space may limit the flexibility and expressiveness of the model's responses. For the VQG task, there has been a paradigm shift from neural-network-based approaches to reinforcement learning settings [13], where the optimization reward is based on the informativeness of the generated questions [14]. While most significant VQA research has focused on 2D images, the task has also been extended to cover VQA of 360° panoramic images and videos [15,16]. Chou et al. [17] introduced VQA 360°, where the panoramic image, unlike the usual visual image, contains the complete visual content around the camera's optical center. This input representation offers a comprehensive understanding of entities in the image's full scene, demanding models with spatially sound reasoning to leverage the extra available information. Various VQA models have performed excellently in 2D environments [18–21].

2.2. Question Answering in 3D Scene

In 2D VQA tasks, CNNs are commonly used. However, the three-dimensional CNN method is costly. The spatial complexity and time complexity increase as the cube of N's resolution increases, and the computational cost is very high. Therefore, only a shallow resolution can be used, leading to quantized noise errors and limiting recognition accuracy. The PointNet++ backbone [22] is used as a scene encoder to obtain object proposals. A transformer is then employed to integrate the information from both modalities, resulting in a final vector for answer prediction. The core of PointNet++ proposes a multi-level feature extraction structure to extract local feature extraction and global features effectively.

Azuma et al. employed the PointNet++ architecture to introduce the VQA task into 3D scenes [4]. In this work, the authors utilized a dual-branch decoder to encode question tokens and point clouds separately. In the ScanQA method, the language-encoding layer encodes the question words $\mathrm{w}_{i}{}_{i=1}^{n_q}$ using GloVe (Global Vectors for Word Representation) embeddings. GloVe represents a notable breakthrough in natural language processing (NLP), offering effective word embeddings through the utilization of global co-occurrence statistics. At its core, GloVe introduces a crucial formula for computing word embeddings based on the co-occurrence probability ratio of words. This formula, encapsulated within its loss function, plays a pivotal role in training GloVe embeddings. Specifically, the definition of the loss function is as Equation (1):

$$\mathrm{J} = \sum_{i=1}^{V} \sum_{j=1}^{V} f\left(X_{ij}\right) \left(w_i^{\top} \widetilde{w}_j + b_i + \widetilde{b}_j - \log \mathrm{X}_{ij}\right)^2, \tag{1}$$

where X_{ij} represents the co-occurrence count of words i and j, f is a weighting function, w_i and $\widetilde{w}_j$ are the word vectors, and b_j and $\widetilde{b}_j$ are bias terms. Understanding this formula is essential for appreciating the underlying mechanisms of GloVe embeddings and their applications in various NLP tasks. The research on ScanQA introduces the VQA task into three-dimensional scenes. Due to its reliance on predefined semantic spaces for responses, the ScanQA model can only provide answers in the form of simple phrases rather than complete sentences. Furthermore, lacking significant prior knowledge, the model is limited to answering questions on basic physical information about the scene such as color, location, and quantity.

Furthermore, the effective integration of diverse modal information, particularly the alignment and consistent representation of heterogeneous data like point clouds and images, has garnered increasing attention from researchers. Subsequent research has endeavored to transfer prior knowledge from the 2D to the 3D domain [23]. Unlike previous methods that predict answers from the answer space, Dwedari et al. suggest employing an end-to-end transformer architecture for generating free-form answers in recent studies [6]. This improvement enables the model to generate free-form and complete-sentence answers. However, limitations in answering questions restricted to simple physical features persist.

2.3. Aspect-Based Sentiment Analysis Model (ABSA)

ABSA encompasses two fundamental tasks: aspect extraction (AE) and aspect sentiment classification (ASC). Traditional, fine-grained sentiment analysis (ABSA) typically involves three subtasks: attribute extraction, opinion extraction, and a series of three sequential tasks aimed at determining the emotional tendency of attribute–opinion pairs. Certain words appear more frequently in sentences within reviews, and these commonly occurring terms are identified as aspects. ABSA seeks to identify various aspects, such as food, service, and environment, discussed within reviews and evaluate the sentiment polarity associated with each one [24].

Traditional sentiment analysis methods relied heavily on handcrafted features [25–28]. These approaches were limited to sentence-level or document-level analysis, neglecting the connection between predefined aspects and content. Subsequently, attention-based neural networks gradually replaced feature engineering methods, addressing the need for manual feature extraction. Tang et al. proposed a hierarchical attention network, which captures sentiment information relevant to predefined aspects [26,29].

3. Method

The model framework proposed in this study, as depicted in Figure 1, is presented in detail in this section. The main objective of our research is to comprehensively evaluate interior design by integrating image generation and sentiment analysis methods. Our model takes two types of input data: a 3D point cloud consisting of RGB and standard features and a series of questions. These two types of input data are processed by the 3DQA module to generate corresponding textual answers, which are then fed into the DualGCN module for sentiment analysis, resulting in sentiment analysis results. The Gen3DQA module is responsible for generating natural language textual answers from input interior design images and questions, while the DualGCN module conducts sentiment analysis by analyzing these textual answers.

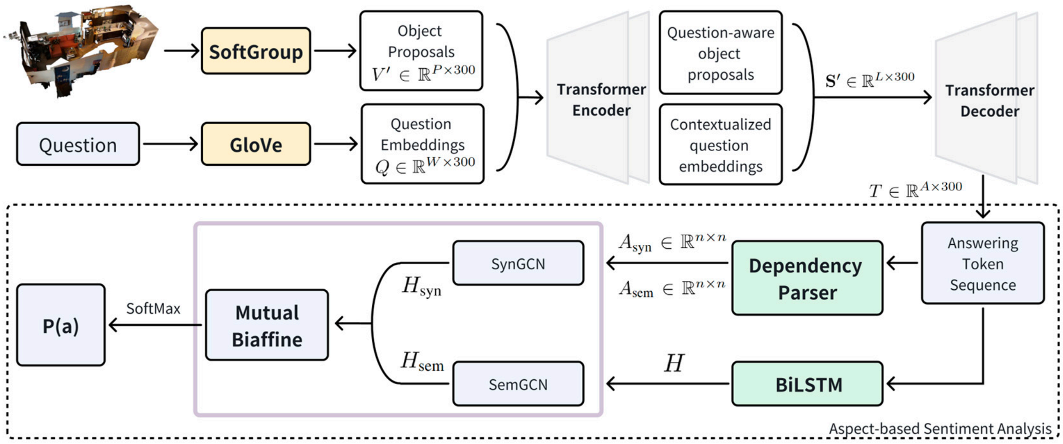

SoftGroup
Object Proposals $V' \in \mathbb{R}^{P\times 300}$
Question
GloVe
Question Embeddings $Q \in \mathbb{R}^{W\times 300}$
Transformer Encoder
Question-aware object proposals
Contextualized question embeddings
$\mathbf{S}' \in \mathbb{R}^{L\times 300}$
Transformer Decoder
$T \in \mathbb{R}^{A\times 300}$
Answering Token Sequence
Dependency Parser
$A_{\text{syn}} \in \mathbb{R}^{n\times n}$
$A_{\text{sem}} \in \mathbb{R}^{n\times n}$
SynGCN
H_{syn}
H_{sem}
SemGCN
H
BiLSTM
Mutual Biaffine
SoftMax
P(a)
Aspect-based Sentiment Analysis

Figure 1. Our framework leverages the generation of 3D question-answering and sentiment analysis to assess interior designs in 3D scenes. By combining a transformer architecture and dual-graph convolutional networks, we generate text descriptions of RGB-D interior scenes and analyze them comprehensively. The part inside the dotted box is the aspect-based sentiment analysis module. This approach reduces resource consumption while facilitating detailed evaluations.

3.1. 3D Question Answering

3D Point Cloud Embedding. In the 3D question-and-answer stage, we employ the sophisticated SoftGroup method [30] to extract semantic information from object proposals using 3D sparse convolutions. The opposite of SoftGroup methods is early hard prediction, which associates each point with a single category when performing semantic segmentation. Two problems arise when using previous methods for hard prediction classification: the IoU between prediction and Gt is low, and incorrect semantic prediction produces redundant false positives. The SoftGroup method allows each point to retain the prediction results of multiple semantic categories, thereby predicting more valid instance points and reducing incorrectly predicted instance results.

Token Sequence Embedding. The second input comprises the token sequence of the question. To encode these tokens and vectorize the words, we utilize the GloVe model [31], aiming to imbue the vectors with maximal semantic and grammatical information. At this stage, the user-input question vector Q is obtained, where W represents the token count, and 300 denotes the dimension of the GloVe word embedding, as in $Q \in R^{W\times 300}$. This vector holds the semantic details of every token within the question sequence, aiding in subsequent processing steps.

Transformer Encoder–Decoder. We use the original transformer positional encoding [32] for question embedding. The transformer model utilizes positional encoding generated through sine and cosine functions with different frequencies, which is subsequently integrated with the word vectors at corresponding positions. The dimension of the position vector should match the word vector dimension. We incorporate the generated object proposals into the last three dimensions of each (extended) object proposal. Initially, the attention scores are computed by applying the softmax function to obtain the attention weight for each time step after calculating the dot product between the query Q and the key V' prime. The formula for calculating our attention weight is as Equation (2):

$$\text{Attention}(\mathrm{Q}, V') = softmax\left(\frac{QV'^{\top}}{\sqrt{d_k}}\right), \tag{2}$$

where d_k denotes the dimensionality of the keys V', and the dot product operation $QV'^{\top}$ assesses the correlation between the question and the object proposals. We then use the obtained attention weights to perform a weighted sum on the values V', resulting in the generation of the contextual encoding sequence S. This process can be represented as Equation (3):

$$S = Attention\left(Q, V'\right) \cdot V', \tag{3}$$

where the symbol S denotes the contextual encoding sequence, possessing dimensions that mirror those of the values V' prime. This computational step effectively captures the semantic connections between the posed question and the proposed objects, thereby furnishing crucial contextual information essential for the subsequent decoding process to generate accurate responses.

Here, T' represents the output sequence generated by the transformer decoder, containing the token embeddings of the answer. This sequence is further processed by the DualGCN module to refine the generated answer. The answer sequence $T \in R^{\hat{A} \times 300}$, embedded with GloVe, serves as input to the transformer decoder for the training set, facilitating the generation of the final token sequence.

3.2. Text Sentiment Analysis Model

The generated textual responses from the 3D question-answering phase are fed into the ABSA model [33]. The DualGCN model maximizes the utilization of two crucial modules, namely, the syntax-based graph convolutional network (SynGCN) and the semantic-based graph convolutional network (SemGCN).

For the generated answer text t from the decoder, a text–aspect pair (t, a) is given, where t represents a text containing the context for sentiment analysis, and a represents one or more aspects (aspects) in the sentence, which are the target objects or topics for sentiment analysis. The following represents an example: t: "I really like the spacious layout and elegant furniture in this living room." a: ["layout", "furniture"]. In this example, t represents a sentence describing a living room, and a represents two aspects in the living room (i.e., layout and furniture). $a = a_1, a_2, \ldots, a_m$ represents an aspect, which is a subsequence of $t = \omega_1, \omega_2, \ldots, \omega_m$.

DualGCN. To extract hidden contextual representations, we employ both BiLSTM and BERT as sentence encoders. BiLSTM functions as the sentence encoder, extracting implicit contextual representations. The final aspect representation is obtained by pooling and concatenating the representations of all aspects, ultimately derived from the SynGCN and SemGCN modules.

Syntax-Based Graph Convolutional Network (SynGCN). The syntax-based graph convolutional network (SynGCN) module takes as its input a grammar encoding. Utilizing the dependency parsing model known as LAL-Parser [34], we incorporate syntactic details by employing the probability matrix containing all dependency arcs generated by the dependency parser in this study.

Using the adjacency matrix $A_{syn} \in R^{n \times n}$ of syntactic encoding, in the SynGCN module, the initial node representations within the syntactic graph are derived from the vector H obtained from the Bidirectional LSTM. The SynGCN module obtains the syntactic graph representation $H_{syn} = h_{syn_1}, h_{syn_2}, \ldots, h_{syn_n}$. In this context, the hidden representation $h_{syn_i} \in R^d$ represents the $i - th$ node. It is worth mentioning that for aspect nodes, we adopt the notation $\left\{h_{syn_{a_1}}, h_{syn_{a_2}}, \ldots, h_{syn_{a_m}}\right\}$ to denote these hidden representations.

Semantic-Based GCN (SemGCN). The SemGCN is crafted to capture semantic relationships within a sentence through self-attention mechanisms, without the necessity of explicit syntactic knowledge utilized by the SynGCN. This flexibility allows the SemGCN to adapt to various text inputs, including online reviews that may lack structured syntactic information.

Within our DualGCN framework, we utilize a self-attention layer to compute the attention score matrix denoted as $A_{sem} \in R^{n \times n}$, where attention scores are computed simultaneously between every pair of elements. The attention score matrix A_{sem} serves as the foundation for the SemGCN module and is structured as Equation (4):

$$A_{sem} = softmax\left(QW_Q Q^{\top} - KW_K^{\top}\right). \tag{4}$$

In this formulation, Q and K represent graph representations, while W_Q and W_K denote trainable weight matrices. The variable d represents the dimensionality of the input node features. We compute the attention score matrix by utilizing a single self-attention head. The SemGCN module, following a similar approach to the SynGCN module, derives the graph representation for every pair of elements using the BiAffine Module. To enable the sharing of pertinent features between the SynGCN and SemGCN modules, we employ a mutual BiAffine transformation, formulated as Equations (5) and (6):

$$H_{syn0} = softmax\left(H_{syn}W_1(H_{sem})^{\top}\right), \tag{5}$$

$$H_{sem0} = softmax\left(H_{sem}W_2(H_{syn})^{\top}\right), \tag{6}$$

where W_1 and W_2 denote trainable parameters.

Through pooling operations and concatenation performed on the aspect nodes within the SynGCN and SemGCN modules, the ultimate feature representation for the ABSA task is Equations (7)–(9):

$$h_{syn_a} = avg - pooling\left(h_{syn_{a_1}}, h_{syn_{a2}}, \ldots, h_{syn_{am}}\right), \tag{7}$$

$$h_{sem_a} = avg - pooling\left(\left\{h_{sem_{a_1}}, h_{sem_{a2}}, \ldots, h_{sem_{am}}\right\}\right), \tag{8}$$

$$r = \left[h_{syn_a}; h_{sem_a}\right]. \tag{9}$$

After processing through a linear layer, the representation denoted as r undergoes transformation, followed by the application of a softmax function, resulting in the generation of a probability distribution p for sentiment Equation (10):

$$p(a) = softmax\left(W_{p^r} + b_p\right). \tag{10}$$

Within this framework, D constitutes a comprehensive set comprising pairs of sentences and their associated aspects, encapsulating the entirety of the sentiment analysis task. Conversely, C serves as the repository for an array of diverse sentiment polarities, providing the necessary context for evaluating the sentiment expressed within the textual data. This structured representation facilitates a systematic analysis of sentiment across various aspects, enhancing the model's ability to discern emotional nuances within the input sentences.

Regularization and Loss Function

Regularizer. These regularization techniques serve to refine the learned features within the SemGCN module, promoting a more nuanced understanding of semantic relationships. By encouraging orthogonal features, the model can effectively disentangle distinct aspects of the input data, facilitating clearer interpretation and reducing the risk of overfitting. Additionally, the incorporation of the differential regularizer further enhances the model's adaptability to diverse semantic structures, fostering robustness in capturing semantic nuances.

The orthogonal regularizer plays a crucial role in fostering the acquisition of interpretable feature representations and fostering the acquisition of features that are mutually orthogonal. By enforcing this regularization, it is assumed that items with semantic relationships within a sentence should be positioned in distinct regions, thereby minimizing redundancy and bolstering the model's capacity to extrapolate to unseen data. This regularization is formally defined as Equation (11):

$$R_O = \left|\left|A_{sem}A_{sem}^{\top} - I\right|\right|_F. \tag{11}$$

Within this context, the identity matrix is denoted by the symbol I, and the Frobenius norm is represented by $\| \cdot \|_F$. This regularization technique operates by minimizing the off-diagonal elements of $A_{\text{sem}}A_{\text{sem}}^{\top}$ to preserve orthogonality, thereby ensuring that the learned features remain mutually independent.

The differential regularizer serves to promote distinct representations obtained from both the SynGCN and SemGCN modules by constraining the difference between their adjacency matrices. This technique aims to ensure that the semantic information captured by each module remains complementary rather than redundant. Its mathematical expression is given as Equation (12):

$$R_D = \frac{1}{\left|\left|A_{\text{sem}} - A_{\text{syn}}\right|\right|_F}. \tag{12}$$

Loss Function. The primary objective during training is to minimize the comprehensive loss function, which encompasses both the cross-entropy loss designed for sentiment analysis and additional regularization terms. This loss function formulation ensures that the model not only learns to predict sentiment accurately but also maintains desirable properties such as feature interpretability and parameter stability through regularization techniques as Equation (13):

$$\mathcal{L}_T = \mathcal{L}_C + \lambda_1 R_O + \lambda_2 R_D + \lambda_3 |\theta|_2^2, \tag{13}$$

where the regularization coefficients λ_1, λ_2 and λ_3 are introduced to balance the impact of regularization terms. All trainable parameters within the model are denoted by θ. Tailored for the ABSA task, the standard cross-entropy loss $\mathcal{L}_C$ measures the disparity between predicted sentiment probabilities and ground truth labels.

$$\mathcal{L}_C = -\sum_{(s,a)\in D}\sum_{c\in C} logp(a), \tag{14}$$

where the set D comprehensively captures the associations between sentences and their respective aspects, facilitating a thorough analysis of sentiment within the textual data. Similarly, the collection C encompasses a diverse range of sentiment polarities, providing a nuanced framework for evaluating the sentiment expressed across different aspects. This structured representation enables a systematic exploration of sentiment nuances, enhancing the model's capacity to discern subtle emotional variations within the input sentences.

3.3. Proposed 3D Visual Question-Answering Model

We employ a pretrained 3DQA model for our training process. The pretrained model has been previously trained by other researchers and is used as a backbone for our approach. Here, we detail the training procedure for our visual-language model, focusing on fine-tuning the pretrained 3DQA model.

Initially, our training process involves the pre-training of SoftGroup using ScanRefer [35]. This decision is informed by the findings of Dwedari et al., whose experiments revealed that the combination of RGB and regular features resulted in the highest overall score. However, it is worth noting that the forward pass for SoftGroup entails a significant computational cost. Therefore, during predictions on the test set, the SoftGroup undergoes reactivation, ensuring that a complete forward pass is executed to obtain accurate results.

Throughout the end-to-end training process of the SoftGroup model, we utilize a multitask loss function called SoftGroup. This loss function encapsulates the collective loss incurred during the initial training phase, providing a comprehensive measure of model performance. By optimizing this multitask loss function, the SoftGroup model adapts and refines its parameters to better capture the semantic relationships between visual and textual inputs, enhancing its overall effectiveness in the subsequent tasks.

We then refine our question-answering model through a two-stage training regimen. During the initial stage, the question-answering model undergoes training with a focus on minimizing the word-level cross-entropy (XE) loss, denoted as Lans. This loss function is formulated as Equation (15):

$$\mathcal{L}_{\text{ans}} = -\sum_{t\in T}\sum_{z\in Z} y_{t,z}\log\left(\hat{y}_{\text{t,z}}\right), \tag{15}$$

where the symbol T represents the ground truth answer, providing a reference for training the model. The vocabulary used during training is denoted by Z, encompassing the set of all possible tokens within the dataset. Each predicted probability $\hat{y}_{t,z}$ corresponds to the likelihood of token z in the softmax output at step t, serving as a crucial component for assessing model performance.

As for the localization branch, its cross-entropy loss $\mathcal{L}_{loc}$ follows a similar formulation to established methods, ensuring consistency with prior approaches in the field. Consequently, the total loss for the second training phase is determined by the summation of $\mathcal{L}_{\text{ans}}$ and $\mathcal{L}_{loc}$, expressed as $\mathcal{L} = \mathcal{L}_{\text{ans}} + \mathcal{L}_{\text{loc}}$.

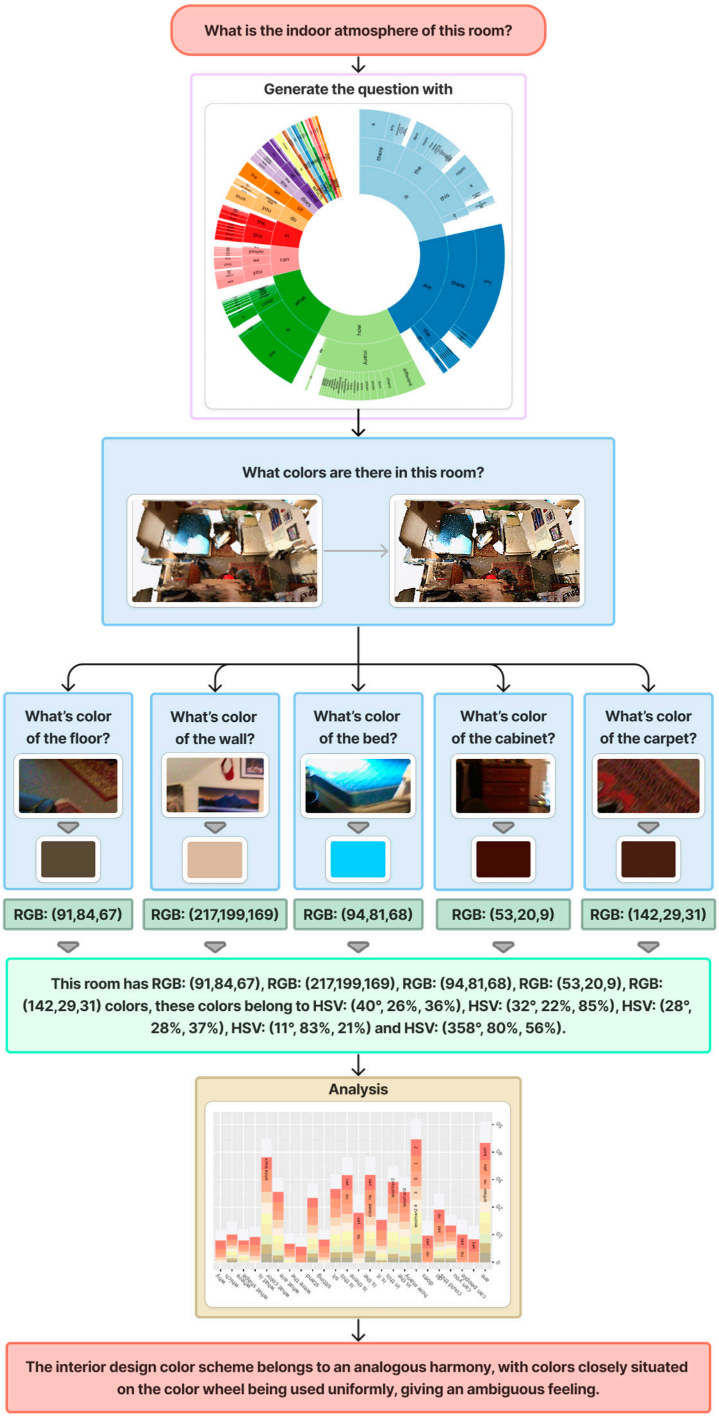
Following this training phase, we proceed to fine-tune our pre-trained 3DQA model, aligning it with the specific requirements of our task and ensuring its optimal performance in subsequent evaluations.

3.4. Procedure of Interior Design Evaluation Model

The detail procedure of the proposed interior design evaluation model is shown in Figure 2. It can be seen from Figure 2 that for a vague conceptual question, the proposed framework supports splitting the question into specific sub-questions. The final judgment is obtained as total question–sub-question–total question. During testing, we reactivate the SoftGroup module to generate object proposals. For the transformer decoder, we adopt greedy decoding to generate answer sequences, as carried out in prior work. The answer sequences are then fed into the aesthetic rule module for comprehensive analysis. Initially, text questions are provided by the user and are then fed into the DualGCN for analysis to obtain sub-questions. The sub-questions are fed into Glove in text form, fused with the 3D point cloud scene embedded through Softgroup. Ultimately, text answers are generated by the decoder. This example primarily focuses on implementing combinations of RGB colors. We identify the five most voluminous objects in the indoor scene and analyze their color combination relationships. To begin, we need to compute the hue values based on the RGB values using the following formula:

$$\text{H} = \begin{cases} \text{undefined}, \quad \text{if } max = min \\ 60^\circ \times \frac{g-b}{max-min} + 0^\circ, \quad \text{if } max = r \text{ } and \text{ } g \geq b \\ 60^\circ \times \frac{g-b}{max-min} + 360^\circ, \quad \text{if } max = r \text{ } and \text{ } g < b \\ 60^\circ \times \frac{b-r}{max-min} + 120^\circ, \quad \text{if } max = g \\ 60^\circ \times \frac{r-g}{max-min} + 240^\circ, \quad \text{if } max = b \end{cases} \tag{16}$$

The difference in H values determines the angular separation of each color on the hue circle, which helps assess the color harmony. For example (Figure 2), the main colors in the room have the following H values: RGB (91, 84, 67) corresponds to H value 42, RGB (217, 199, 169) corresponds to H value 37, RGB (94, 81, 68) corresponds to H value 30, RGB (53, 20, 9) corresponds to H value 15, and RGB (142, 29, 31) corresponds to H value 359. After calculation, the maximum difference in their H values is 43 degrees (since hue values cycle between 0 and 360 degrees). This color information is fed into the large language model ChatGPT. We leverage relevant interior design knowledge to process the large language model. In this example, if the difference in hue values among the five colors is less than or equal to this threshold, they can be considered to exhibit analogous harmony. According to Zhang et al. [36], color combinations in analogous harmony do not create strong contrasts but instead present a soft and unified effect, evoking feelings of comfort and relaxation. This color scheme is often used to create warm and peaceful indoor environments, suitable for spaces such as bedrooms and living rooms that require a cozy atmosphere. In addition to the examples provided, disciplines such as art and design typically classify color combinations uniformly and associate them with corresponding psychological effects.


What is the indoor atmosphere of this room?
Generate the question with
What colors are there in this room?
What's color of the floor?
What's color of the wall?
What's color of the bed?
What's color of the cabinet?
What's color of the carpet?
RGB: (91,84,67)
RGB: (217,199,169)
RGB: (94,81,68)
RGB: (53,20,9)
RGB: (142,29,31)
Analysis

This room has RGB: (91,84,67), RGB: (217,199,169), RGB: (94,81,68), RGB: (53,20,9), RGB: (142,29,31) colors, these colors belong to HSV: (40°, 26%, 36%), HSV: (32°, 22%, 85%), HSV: (28°, 28%, 37%), HSV: (11°, 83%, 21%) and HSV: (358°, 80%, 56%).

The interior design color scheme belongs to an analogous harmony, with colors closely situated on the color wheel being used uniformly, giving an ambiguous feeling.

Figure 2. For a vague conceptual question, our framework supports splitting the question into specific sub-questions. The final judgment is obtained as total question–sub-question–total question.

4. Experiments

According to professional knowledge related to interior design and art, different combinations of materials and colors produce different design atmospheres [36]. Therefore, we can design a set of rules by exploring patterns in the combinations of these colors and materials, thereby determining the feelings evoked by interior spaces. The atmosphere of a space is also related to the style of interior furnishings and the distance between objects. For an abstract question raised by users, such as atmosphere, quality, and feelings, our framework can reconstruct the problem and split a main question into multiple sub-problems. As computers excel in describing precise information rather than abstract feelings, we reconstruct and break down the questions raised by humans into concrete sub-problems related to color, position, orientation, size, and so on. Eventually, by aggregating the answers to these sub-problems, a comprehensive description can be obtained.

We tested our model using scene data from ScanNet [37]. The proposed model scores using the XE loss trained on a single object, multiple objects, and no object localization at all are given in Table 1. It can be seen that the proposed model exhibits different effects under different conditions. In our observations, the model demonstrated a slight performance enhancement when handling multiple objects compared to single-object scenarios. In analyzing the experimental results, we found that the model exhibited different effects under different conditions. Specifically, when we applied the model to handle multiple objects, we observed a slight performance improvement. This indicates that our model has a certain advantage in handling complex scenarios, enabling it to better deal with the relationships and interactions among multiple objects. However, in contrast, we found a moderate decline in the model's performance when object localization information was removed. This suggests that the position information of objects is crucial for the accuracy and effectiveness of the model. In interior design, the position of objects not only affects the layout and atmosphere of a space but also directly influences people's perception and experience within that space. Therefore, the model may lose its grasp of the overall spatial characteristics when lacking object localization information, thereby affecting its performance.

Table 1. The proposed model scores using the XE loss trained on a single object, multiple objects, and no object localization at all.

Model	BLEU-1	BLEU-4	BLEU-4	BLEU-4
Gen3DQA+DGCN (single object)	36.22	10.31	34.15	63.82
Gen3DQA+DGCN (multiple object)	36.71	10.17	33.29	63.17
Gen3DQA+DGCN (w/o object localization)	35.83	9.89	31.85	60.85

To further validate the model performance of the proposed method, Table 2 shows image captioning metric scores for different types of question validation sets based on different methods. Our experiments tested the answer text effects of several different framework combinations. These metrics are employed to evaluate the effectiveness of various models in natural language tasks, reflecting the proximity of the generated text to human performance. Notably, among all indicators, the Gen3DQA+DGCN model demonstrates superior performance. This indicates the model's capability in handling scenarios with multiple objects and its robustness in object localization. This result provides important insights into understanding the model's performance in real-world applications and guides future efforts to further improve model performance.

Table 2. Image captioning metric scores for different types of question validation sets based on different methods.

Model	BLEU-1	BLEU-4	Rouge	Meteor
Gen3DQA+DGCN	39.72	12.41	36.32	72.41
Baseline1+DGCN	34.14	9.37	31.36	66.37
Baseline2+DGCN	29.78	8.47	28.46	61.48
Baseline3+DGCN	32.58	8.92	29.33	66.92
VoteNet+MCAN	29.46	6.08	30.97	12.07
ScanRefer+MCAN	27.85	7.46	30.68	11.97
Gen3DQA+ATAE-LSTM	38.72	10.46	33.74	71.39
Gen3DQA+DGEDT	39.43	11.18	33.98	70.38
Gen3DQA+InterGCN	39.15	11.84	35.17	70.93

Baseline 1. Lxmert: Learning cross-modality encoder representations from transformers. Baseline 2. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Baseline 3. 12-in-1: Multi-task vision and language representation learning.

Experiments were conducted using scene data from ScanNet, and the interior design evaluation accuracy is given in Table 3. The varying effects of our model under different conditions can be observed. Specifically, our model exhibited a slight performance enhancement when handling multiple objects compared to single-object scenarios. This suggests that our model has an advantage in dealing with complex scenarios, enabling it to better capture the relationships and interactions among multiple objects. However, we also noticed a moderate decline in the model's performance when object localization information was removed. This highlights the crucial role of object position information for the accuracy and effectiveness of the model, as the spatial arrangement of objects directly influences people's perception and experience within a space. These findings underscore the importance of considering object localization in interior design applications and suggest avenues for further improving our model's performance in real-world scenarios.

Table 3. Comparison of the interior design evaluation accuracy based on different methods.

Model	Accuracy
Gen3DQA+DGCN	91.53
Baseline1+DGCN	74.82
Baseline2+DGCN	69.06
Baseline3+DGCN	71.92
VoteNet+MCAN	70.38
ScanRefer+MCAN	72.49
Gen3DQA+ATAE-LSTM	76.39
Gen3DQA+DGEDT	84.29
Gen3DQA+InterGCN	81.53

In the user interaction process, individuals start by submitting an indoor point cloud scene and subsequently inputting text to ask questions about the scene. However, our model has restrictions on the types of questions it can handle. Although it can respond to queries concerning physical attributes like color, position, and distance, it can also address inquiries regarding the style or atmosphere of the indoor scene.

5. Conclusions

In this study, we leverage advanced techniques of transformer encoder–decoder and graph convolutional networks to address the challenge of evaluating interior designs from different aspects in a 3D environment. Our framework integrates a 3D question-answering task (3DQA) with a dual-graph convolutional network (DualGCN) to comprehensively evaluate 3D indoor scene design. By generating textual descriptions of RGB-D internal scenes through text-form questions and sentiment analysis, we can leverage textual data to obtain 3D scene information effectively.

Leveraging their ability to extract hidden contextual representations, we employ pre-trained models such as BiLSTM and BERT for sentence encoding. Incorporating both the SynGCN and SemGCN modules within the DualGCN framework allows us to account for both semantic correlation and syntactic structure in sentiment analysis. This integration enhances the overall accuracy and robustness of the evaluation process.

We introduce a new task: evaluating interior designs in a 3D environment. By providing a framework designed for this task, we enable users to evaluate 3D interior scene designs efficiently. Our system allows users to interact with scenes through text input, promoting user engagement and enhancing the overall user experience.

Discussion: Although our framework shows promising results, some issues are still worthy of discussion and follow-up. First, the effectiveness of our approach can be further evaluated and validated through large-scale user studies involving designers, architects, and end users. In addition, incorporating user feedback mechanisms can enhance the model's understanding of personal preferences and design styles. Expanding to include additional modalities such as audio and video data can enhance the evaluation process, providing a more holistic comprehension of the scene. In terms of practical applications, this technology could offer valuable guidance for people's everyday interior decoration choices and serve as a guiding tool for the design of movie sets and game environments. This is because humans, especially those without a background in design, often have basic perceptions of space but may lack the ability to deduce specific details. In the future, our model may integrate and infer textual information further, potentially incorporating large language models like ChatGPT for enhanced analysis and synthesis.

Limitations and Future Work: Despite its potential, our framework has certain limitations. Relying on pre-trained models could potentially introduce biases or limitations inherent in the training data. Moreover, the scalability of the system to large and complex scenes warrants further investigation. Traditional interior design assessment is primarily conducted by humans. In comparison to human evaluation, our model still exhibits differences in terms of diversity, comprehensiveness, and detail. Regarding diversity, our model's capability is limited to addressing specific types of questions, such as those concerning style/atmosphere that can be inferred based on factors like color and placement. However, it cannot answer subjective questions related to factors like the reasonableness of design arrangements or direct judgments of good/bad. In terms of comprehensiveness and detail, factors such as the texture details of furniture also influence the style and atmosphere of a design, but our current model lacks the precision to address such aspects.

Moving forward, our focus will be on overcoming these limitations by investigating methods to mitigate bias in pre-trained models and improving the scalability and efficiency of the framework. Additionally, we plan to extend our approach to other domains such as virtual reality and augmented reality, enabling immersive and interactive evaluations of spatial designs. In conclusion, our research establishes a groundwork for future explorations in evaluating interior design within 3D environments, opening avenues for innovative advancements in the field.

Author Contributions: Conceptualization, Y.F., Z.Y. and Y.Z.; methodology, Y.F. and Z.Y.; software, Y.F. and Y.Z.; validation, Y.F. and Y.Z.; formal analysis, Y.F.; investigation, Y.F. and Z.Y.; resources, Y.Z.; writing—original draft preparation, Y.F.; writing—review and editing, Y.Z.; supervision, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 62303296.

Data Availability Statement: This data can be found here, https://tianchi.aliyun.com/competition/entrance/231573/information (accessed on 1 April 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wei, Z.; Zhang, J.; Shen, X.; Lin, Z.; Mech, R.; Hoai, M.; Samaras, D. Good view hunting: Learning photo composition from dense view pairs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5437–5446.
2. Xie, D.; Hu, P.; Sun, X.; Pirk, S.; Zhang, J.; Mech, R.; Kaufman, A.E. Gait: Generating aesthetic indoor tours with deep reinforcement learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7409–7419.
3. Shao, Z.; Yu, Z.; Wang, M.; Yu, J. Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 14974–14983.
4. Azuma, D.; Miyanishi, T.; Kurita, S.; Kawanabe, M. Scanqa: 3D question answering for spatial scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19129–19139.
5. Parelli, M.; Delitzas, A.; Hars, N.; Vlassis, G.; Anagnostidis, S.; Bachmann, G.; Hofmann, T. Clip-guided vision-language pre-training for question answering in 3D scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5606–5611.
6. Dwedari, M.M.; Niessner, M.; Chen, Z. Generating Context-Aware Natural Answers for Questions in 3D Scenes Supplementary Material. *arXiv* **2013**, arXiv:2310.19516.
7. Zhu, L.; Xu, M.; Bao, Y.; Xu, Y.; Kong, X. Deep learning for aspect-based sentiment analysis: A review. *PeerJ Comput. Sci.* **2022**, *8*, e1044. [CrossRef] [PubMed]
8. Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A. *A Practical Guide to Sentiment Analysis*; Springer: Cham, Switzerland, 2017; Volume 5.
9. Malinowski, M.; Fritz, M. A multi-world approach to question answering about real-world scenes based on uncertain input. *Adv. Neural Inf. Process. Syst.* **2014**, 27.
10. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
11. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 2425–2433.
12. Zhou, Y.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6281–6290.
13. Jain, U.; Zhang, Z.; Schwing, A.G. Creativity: Generating diverse questions using variational autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5415–5424.
14. Li, Y.; Duan, N.; Zhou, B.; Chu, X.R.; Ouyang, W.; Wang, X. Visual question generation as dual task of visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6116–6124.
15. Jang, Y.; Song, Y.; Kim, C.D.; Yu, Y.; Kim, Y.; Kim, G. Video question answering with spatio-temporal reasoning. *Int. J. Comput. Vis.* **2019**, *127*, 1385–1412. [CrossRef]
16. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; Liu, J. Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7331–7341.
17. Chou, S.-H.; Chao, W.-L.; Lai, W.-S.; Sun, M.; Yang, M.-H. Visual question answering on 360° images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Online, 2–5 March 2020; pp. 1607–1616.
18. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32897–32912.
19. Herrera, D.A.B. Towards an Image-Term Co-Occurrence Model for Multilingual Terminology Alignment and Cross-Language Image Indexing. Ph.D. Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2014.
20. Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of the Conference on Robot Learning PMLR, Zhongshan, China, 9–11 December 2023; pp. 2165–2183.
21. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 13–23.
22. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5105–5114.
23. Cao, A.-Q.; De Charette, R. Monoscene: Monocular 3D semantic scene completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3991–4001.

24. Chauhan, G.S.; Nahta, R.; Meena, Y.K.; Gopalani, D. Aspect-based sentiment analysis using deep learning approaches: A survey. *Comput. Sci. Rev.* **2023**, *49*, 100576. [CrossRef]
25. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
26. Tang, D.; Qin, B.; Liu, T. Aspect-level sentiment classification with deep memory network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 214–224.
27. Zeng, J.; Liu, T.; Jia, W.; Zhou, J. Relation construction for aspect-level sentiment classification. *Inf. Sci.* **2022**, *586*, 209–223. [CrossRef]
28. Huang, B.; Ou, Y.; Carley, K.M. Aspect-level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, 10–13 July 2018*; Springer: Cham, Switzerland, 2018; pp. 197–206.
29. Tan, X.; Cai, Y.; Zhu, C. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3426–3431.
30. Vu, T.; Kim, K.; Luu, T.M.; Nguyen, T.; Yoo, C.D. Softgroup for 3D instance segmentation on point clouds. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2698–2707.
31. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
33. Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; Hovy, E. Dual graph convolutional networks for aspect-based sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 6319–6329.
34. Ma, L.; Zhang, P.; Luo, D.; Zhu, X.; Zhou, M.; Liang, Q.; Wang, B. Syntax-based graph matching for knowledge base question answering. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 8227–8231.
35. Chen, D.Z.; Chang, A.X.; Nießner, M. Scanrefer: 3d object localization in rgb-d scans using natural language. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 202–221.
36. Zhang, Q.; Zheng, S. *Interior Design Materials Collection*; China Architecture & Building Press: Beijing, China, 1991.
37. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5828–5839.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.