forests

MDPI

Article

Parameterization before Meta-Analysis: Cross-Modal Embedding Clustering for Forest Ecology Question-Answering

Rui Tao [1,2], Meng Zhu [3], Haiyan Cao [2] and Hong-E Ren [1,4,*]

[1] College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China; trlx20@nefu.edu.cn

[2] College of Artificial Intelligence and Big Data, Hulunbuir University, Hulunbuir 021008, China; ske159@163.com

[3] College of Information Engineering, Harbin University, Harbin 150076, China; zhum913@163.com

[4] Heilongjiang Forestry Intelligent Equipment Engineering Research Center, Harbin 150040, China

[*] Correspondence: nefu_rhe@163.com

Abstract: In the field of forestry ecology, image data capture factual information, while literature is rich with expert knowledge. The corpus within the literature can provide expert-level annotations for images, and the visual information within images naturally serves as a clustering center for the textual corpus. However, both image data and literature represent large and rapidly growing, unstructured datasets of heterogeneous modalities. To address this challenge, we propose cross-modal embedding clustering, a method that parameterizes these datasets using a deep learning model with relatively few annotated samples. This approach offers a means to retrieve relevant factual information and expert knowledge from the database of images and literature through a question-answering mechanism. Specifically, we align images and literature across modalities using a pair of encoders, followed by cross-modal information fusion, and feed these data into an autoregressive generative language model for question-answering with user feedback. Experiments demonstrate that this cross-modal clustering method enhances the performance of image recognition, cross-modal retrieval, and cross-modal question-answering models. Our method achieves superior performance on standardized tasks in public datasets for image recognition, cross-modal retrieval, and cross-modal question-answering, notably achieving a 21.94% improvement in performance on the cross-modal question-answering task of the ScienceQA dataset, thereby validating the efficacy of our approach. Essentially, our method targets cross-modal information fusion, combining perspectives from multiple tasks and utilizing cross-modal representation clustering of images and text. This approach effectively addresses the interdisciplinary complexity of forestry ecology literature and the parameterization of unstructured heterogeneous data encapsulating species diversity in conservation images. Building on this foundation, intelligent methods are employed to leverage large-scale data, providing an intelligent research assistant tool for conducting forestry ecological studies on larger temporal and spatial scales.

Keywords: forestry ecology; meta-analysis; cross-modal; question-answering; embedding clustering

check for updates

Citation: Tao, R.; Zhu, M.; Cao, H.; Ren, H.-E. Parameterization before Meta-Analysis: Cross-Modal Embedding Clustering for Forest Ecology Question-Answering. *Forests* **2024**, *15*, 1670. https://doi.org/10.3390/f15091670

Academic Editor: Marcello Vitale

Received: 19 August 2024
Revised: 11 September 2024
Accepted: 19 September 2024
Published: 22 September 2024

1. Introduction

Deep learning-based image captioning models represent a relatively mature approach to cross-modal generation from images to language. Upon inputting an image, these models can generate a brief description, as illustrated in Figure 1. Such models are trained on internet data, and the descriptions they generate often fail to meet the specialized requirements of forestry ecology and do not support follow-up queries. When additional knowledge or information is sought, these models cannot provide further assistance. To enable deep learning models to generate specialized descriptions for images in the field of forestry ecology, it is necessary to construct a dataset for training, which requires annotating a large number of images with expert knowledge. It is evident that literature on forestry ecology provides an excellent interpretation of the image of natural conservation. However, the task of extracting expert knowledge relevant to specific images from the literature is immense and surpasses manual efforts, necessitating the use of artificial intelligence methods. As shown in Figure 2, we propose a deep learning model designed to extract factual information from natural conservation imagery and learn expert knowledge from forestry ecology literature. Users can query the model in a question-and-answer manner to extract the acquired information and knowledge, providing an intelligent tool for researchers in forestry ecology to leverage large-scale data and conduct scientific research on broader temporal and spatial scales in less time. When integrating image and language data, the core scientific problem we face is how to cluster the data effectively to facilitate knowledge and information extraction through question-answer reasoning. We will now proceed to discuss this in further detail, starting with the scale of natural conservation image data.

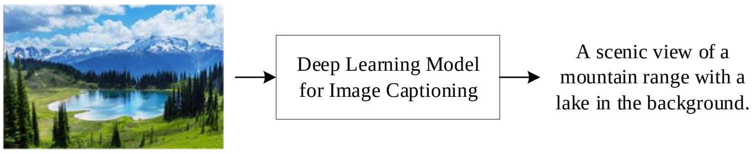


Figure 1. Examples of generations from image captioning model.

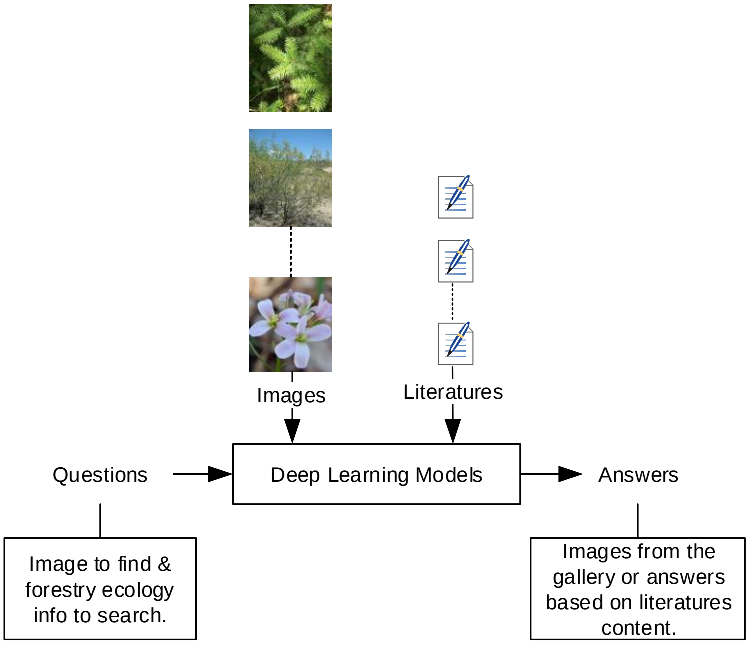


Figure 2. AI-assisted Meta-analysis.

Images are a crucial component of monitoring data in nature reserves, which can be spatially categorized into aerial, terrestrial, and ground levels [1]. With the widespread deployment of cameras in nature conservation and continuous shooting, the volume of ground-level image data is increasing daily. A single field-deployed camera can capture up to 40,000 photos per day triggered by events [2], and the number of images from cameras monitoring areas as vast as tens of thousands of square kilometers for forest fire monitoring is astronomical [3,4]. Each image from a nature reserve captures visual information of a specific time and a local area of the earth's surface, limited by time and space. Although monitoring, management, and research should be closely integrated, monitoring without scientific research is insignificant [5]. However, these image data are fragmented, with relatively limited information, making the research value for auxiliary monitoring and management limited when studying fixed-point or small-area ground images. Conducting cross-regional research by aggregating multiple nature reserve image datasets is more conducive to uncovering the scientific value within these data, thereby forming a closed loop with the monitoring and management of nature reserves, mutually promoting and developing positively. However, image data from ground-based cameras in nature reserves are vast in quantity, diverse in sources, rapidly growing, and involve numerous species, making the pre-processing workload far beyond the scope of human capability. Deep learning offers a method to leverage large-scale data with relatively small labeled datasets, yet this method heavily relies on the quality and quantity of pre-labeled datasets. Given the massive scale of nature reserve image data and the need for expert knowledge, labeling these data is costly. For example, a project mentioned in [6] enlisted thousands of technical volunteers, working for three months, to label 48 species in the image data from 125 camera traps. By this estimate, labeling 5000 species and 10,000 cameras would require over 30 million person-months of work. With such labor demands, expanding this to multiple nature reserve monitoring areas or even globally makes manual labeling nearly impossible, leading to a dilemma where data remain untapped. In forestry ecology meta-analyses, image and literature data sources are often combined to conduct studies such as landmark discovery, species diversity prediction, and habitat analysis. Through cross-verification of factual information within images and expert knowledge in literature, further conclusions or predictions are derived [7–11]. During meta-analyses, image and literature retrieval are frequent and fundamental operations, yet there is a lack of efficient technical means to search across these two heterogeneous databases. Furthermore, forestry ecology involves interdisciplinary knowledge integration across biology, environmental science, and geography, making the classification of the corpus within forestry ecology literature necessary for improving retrieval efficiency. The knowledge contained within forestry ecology literature provides ready-made annotations for nature reserve image data, while the objects of interest within images naturally classify the corpus of knowledge within the literature. Establishing effective connections between these two data types could resolve the bottleneck of manual image annotation and enhance the cross-modal retrieval (for instance, using images to search for text or employing language descriptions to retrieve images) efficiency for forestry ecology researchers across the image and literature heterogeneous databases.

To address this challenge, we propose a novel method: cross-modal embedding clustering (CMEC). This technique employs a dual-encoder (a pair of encoders, one for image encoding and the other for language encoding) architecture, representing species images and the knowledge corpus within the literature in a shared vector space aligned cross-modally, solving the computational problem of cross-modal heterogeneous data. Building on this, we integrate information from a multi-task perspective, combining image recognition, cross-modal retrieval, and cross-modal question-answering, designing a multi-task, cross-modal reasoning model supported mainly by cross-modal clustering with image factual information as the centroid. Through this cross-modal clustering, images can be annotated with expert knowledge from the literature. For example, a photograph of a specific tree species can be annotated with ecological significance, growth patterns, and other information from forestry ecology literature. Simultaneously, the factual information within images can serve as natural clustering centers for textual data, aiding in organizing and retrieving relevant literature.

Our research aims to integrate unstructured heterogeneous data in the field of forestry ecology, providing intelligent data services for scientific research in this field, and helping researchers quickly locate relevant content. As illustrated in Figure 2, we first design a model to parameterize the forestry ecology literature collection and nature reserve image collection. We then use this model to search for the required knowledge or images through a series of questions. Unlike keyword retrieval, this AI-assisted meta-analysis method allows researchers to maintain cognitive continuity, improving the efficiency of knowledge retrieval and organization by eliminating the need to pause, read, organize the retrieved pages and literature, think about the next search keyword, and frequently interrupt their train of thought. In this process, the deep learning model acts like a research assistant, having pre-learned and organized forestry ecology literature, allowing researchers to ask questions as needed, and the model infers and answers based on the learned literature knowledge.

In summary, the main contributions of our research include the following: providing an effective approach for cross-modal information fusion of conservation image data and forestry ecology literature; supporting the parameterization of these two types of heterogeneous and unstructured data, which facilitates the integration of multi-source data in forestry ecological research; and offering an efficient method for intelligent, large-scale forestry ecology research, designed to handle cross-regional, long-term time series data.

2. Related Works

Meta-analysis is a statistical method used to synthesize the results of multiple independent studies to enhance the reliability and generalizability of conclusions. Its underlying logic is based on two core principles: first, by aggregating samples from multiple studies, it increases statistical power and reduces the impact of random errors inherent in individual studies; second, it quantifies the heterogeneity among study results to explore potential underlying causes, thereby providing a more comprehensive understanding of the subject matter [12–14].

The basic steps of a meta-analysis typically include: (1) identifying the research question and retrieval strategy; (2) conducting a systematic literature search and screening for eligible studies; (3) extracting key data and assessing study quality; (4) integrating and analyzing the data using appropriate statistical methods; (5) evaluating heterogeneity and interpreting the results in the context of their significance. The advantage of meta-analysis lies in its ability to synthesize a large number of study outcomes, offering more robust conclusions. However, its limitations include susceptibility to selection bias and heterogeneity issues, and the interpretation of results depends heavily on the quality of the included studies [15–17].

In the field of forestry ecology, meta-analysis has gradually become an important quantitative review tool to reveal the dynamics and changes of complex ecosystems. By integrating research data across multiple locations and temporal scales, meta-analysis can uncover patterns related to forest management, climate change, biodiversity conservation, and more. For example, meta-analyses have enabled scholars to more accurately assess the correlation between deadwood volume and biodiversity in forest ecosystems [18], the impact of forest management on bird community conservation in Eastern North America [19], and the effects of fire on soil microbial metabolic quotient [20].

However, conducting meta-analysis in forestry ecology also faces several challenges. First, research heterogeneity is high; the complexity and diversity of ecosystems often result in poor comparability across studies [21,22]. Second, data sources are scattered; many ecological studies are case-based, with inconsistent data collection methods and statistical analyses, making data integration difficult [23,24]. Additionally, limitations in sample size and spatial scale may lead to bias, affecting the generalizability of the conclusions [25]. Existing studies often overlook the impact of temporal scales on ecological processes, and meta-analysis still faces challenges in handling long-term series data [26].

With advancements in artificial intelligence (AI) technology, AI-assisted meta-analysis aimed at improving efficiency is emerging. The main research methods include AI-assisted literature search and screening, data extraction, data analysis, and heterogeneity detection, as well as result interpretation and visualization [27–30]. In terms of cross-modal representation of image and language data, AI has empowered researchers to handle larger-scale datasets. For instance, integrating global conservation image data to overcome spatial limitations and aggregating decades of forestry ecology literature to transcend temporal constraints. To achieve both, cross-modal representation techniques for images and language are essential. CLIP [31] laid the foundation in this field, followed by developments such as BLIP [32], BEITv3 [33], and COCA [34], establishing the groundwork for cross-modal representation learning of images and literature. After the model has parameterized the information and knowledge from images and literature, it often requires the application of chain-of-thought techniques (a method that guides AI models to perform reasoning based on a coherent sequence of logical cues) [35] to facilitate on-demand retrieval through a question-answering approach.

Literature reviews and meta-analyses are indispensable steps in scientific research. The more literature reviewed, the less biased the conclusions, and the more images analyzed, the more comprehensive the facts. The integration of large-scale conservation image data with forestry ecology literature represents a cross-validation of facts and knowledge, thereby enhancing the reference value of literature reviews and meta-analyses. Table 1 summarizes the representative literature at key technological milestones relevant to our proposed method. From this, we can outline the primary statistical approaches for data and knowledge in the field of forestry ecology, which include: human+algorithm integration, software management, statistical machine learning, expert systems, unimodal (primarily language modality) AI-assisted methods, and multimodal AI-assisted methods. AI-assisted models are further divided into support for statistical analysis and support for reasoning.

Compared with related work, our proposed question-answering cross-modal retrieval method for forestry ecology and conservation image data achieves the parameterization of large-scale, heterogeneous(as discussed in this paper, refers to data from different sources, with different structures, and of different types, which are inherently incompatible), and unstructured data from conservation images and forestry ecology literature, while simultaneously improving visual recognition accuracy and literature classification efficiency. This method is designed to function as an intelligent assistant, enabling researchers in the field of forestry ecology to harness large-scale data. As a result, they can conduct research on a larger spatiotemporal scale with less time investment.

Table 1. Key Methodologies and Insights.

Reference	Proposed	Finding	Limitation
Meng et al., 2020 [26]	Long-term forest ecosystem resilience assessment based on the DTW algorithm.	Elimination of phenological and spectral noise by combining the effects of sudden changes and gradual transitions.	Focusing solely on remote sensing indices may lead to the omission of important dynamic features due to incomplete data.
Urbano et al., 2024 [23]	data management of nature reserves based on software tools.	structured datasets are organized for ease of research use.	High costs and low efficiency.
Zhu et al., 2023 [24]	Literature meta-analysis conducted using a machine learning model guided by personal experience.	Avoiding statistical biases known from prior experience.	A limited sample of 148 papers was analyzed.
Graham et al., 2021 [25]	Crowdsourcing to integrate expertise from interdisciplinary groups.	A guiding framework built from expert knowledge.	Lack of openness and interactivity.
Roy et al., 2024 [30]	AI-assisted insect biodiversity monitoring based on visual analysis.	Automated and intelligent biodiversity monitoring.	AI tools for single-modality assistance.
Radford et al., 2021 [31]	Cross-modal representations in vision-language learning based on contrastive learning.	A base encoder for cross-modal alignment.	Lack of cross-modal reasoning capability.
Wei et al., 2022 [35]	Proposal for guiding inference through introductory prompts.	Guiding models to generate longer descriptions following logical cues.	Limited to language-based single-modality reasoning.

3. Preliminary and Methods

This section primarily explains how we leverage cross-modal clustering as a key method to design deep learning models from a multi-task perspective. In brief, we project both textual and visual data into a shared vector space, fuse their features, and then input the combined representation into a language generation model to generate answers to user queries. The overall workflow is illustrated in Figure 3. For model training, we constructed the NACID dataset based on the iNaturalist 2017 dataset, as shown in Tables 2 and 3, and further detailed in Table 4. We then compared our proposed cross-modal clustering method (Figure 4) with image classification methods represented by ImageNet (Figure 5). To ensure accurate alignment of the image and literature embeddings in the shared space, we trained a momentum encoder using contrastive learning, as depicted in Figure 6. To guide the model in question-answering reasoning, we built a corresponding dataset, as shown in Figure 7. To fully integrate the information from both the literature and image modalities, we designed a dedicated deep network, as illustrated in Figure 8. The modular structure of our model during deployment is shown in Figure 9.

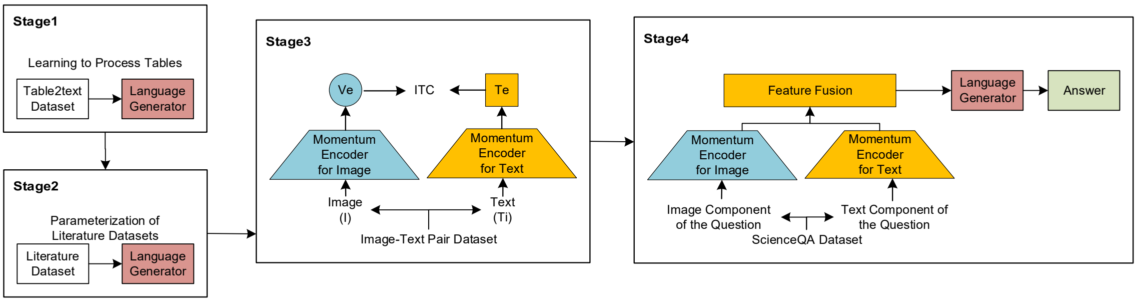

Stage1
Learning to Process Tables
Table2text Dataset
Language Generator
Stage2
Parameterization of Literature Datasets
Literature Dataset
Language Generator
Stage3
Ve
ITC
Te
Momentum Encoder for Image
Momentum Encoder for Text
Image (I)
Text (Ti)
Image-Text Pair Dataset
Stage4
Feature Fusion
Language Generator
Answer
Momentum Encoder for Image
Momentum Encoder for Text
Image Component of the Question
Text Component of the Question
ScienceQA Dataset

Figure 3. The process of model training and fine-tuning.

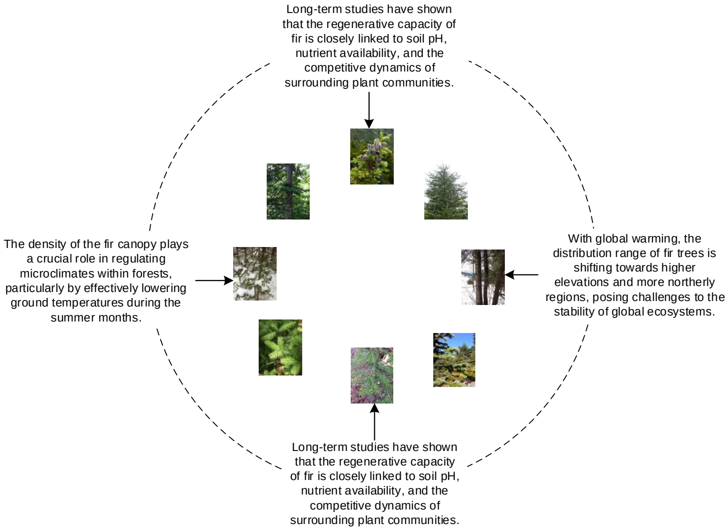

Long-term studies have shown that the regenerative capacity of fir is closely linked to soil pH, nutrient availability, and the competitive dynamics of surrounding plant communities.
The density of the fir canopy plays a crucial role in regulating microclimates within forests, particularly by effectively lowering ground temperatures during the summer months.
With global warming, the distribution range of fir trees is shifting towards higher elevations and more northerly regions, posing challenges to the stability of global ecosystems.
Long-term studies have shown that the regenerative capacity of fir is closely linked to soil pH, nutrient availability, and the competitive dynamics of surrounding plant communities.

Figure 4. Visual facts as the centroid for clustering.

Table 2. The category details of the iNaturalist 2017 dataset.

Super Category	Category Count	Train Images	Val Images
Plantae	2101	158,407	38,206
Insecta	1021	100,479	18,076
Aves	964	214,295	21,226
Reptilia	289	35,201	5680
Mammalia	186	29,333	3490
Fungi	121	5826	1780
Amphibia	115	15,318	2385
Mollusca	93	7536	1841
Animalia	77	5228	1362
Arachnida	56	4873	1086
Actinopterygii	53	1982	637
Chromista	9	398	144
Protozoa	4	308	73
Total	5089	579,184	95,986

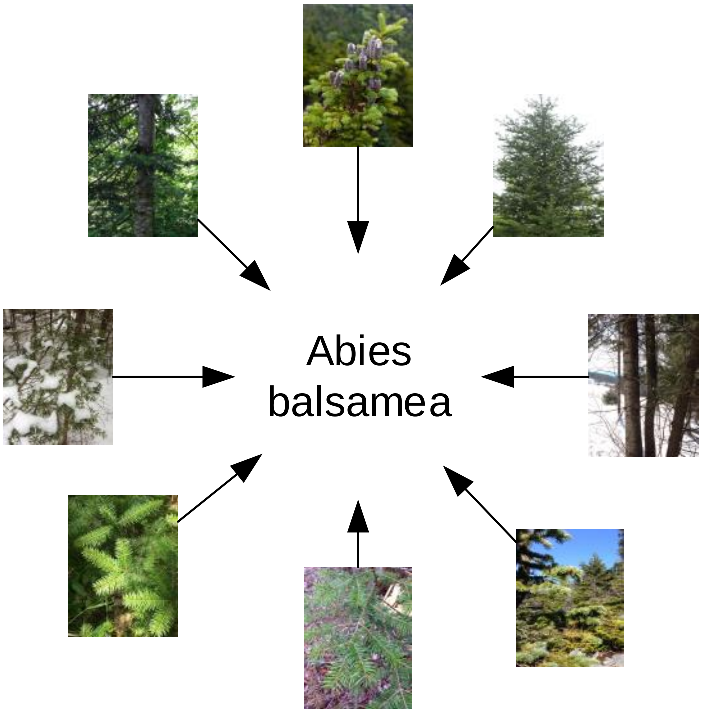
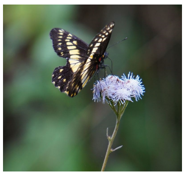
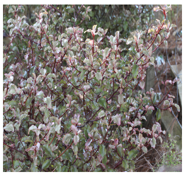


Figure 5. Species name as the centroid for clustering.

Table 3. The instances of the iNaturalist 2017 dataset.

Category Name	Image
Asplenium flaccidum	
Marmota flaviventris	
Rhus aromatica	
Strymon istapa	

Table 4. The instances of the NACID dataset.

Annotation	Images
A butterfly (*Catasticta nimbice*) perches on a floss flower (*Ageratum houstonianum*). *Catasticta nimbice* is commonly found in tropical and subtropical regions of South America, typically active during dawn and dusk; *Ageratum houstonianum* is a dwarf shrub with oval or elliptical leaves, boasting a lengthy flowering period and thriving in warm, moist climates.	
Heterotheca subaxillaris, native to North America, is a perennial herbaceous plant belonging to the Asteraceae family. Its leaves are elongated or lanceolate, and its flowers display a yellow or golden hue. This species is characterized by its drought resistance, heat tolerance, and strong adaptability, making it a valuable ornamental plant.	
Mesembryanthemum crystallinum, known as the ice plant, is native to South Africa and is a drought-resistant and salt-tolerant perennial plant. Its leaves are thick and fleshy, with a gray–green color and sometimes reddish or purplish edges. The flowers of the ice plant are white or pink, and they produce hard capsules as fruits. It is suitable for horticultural beautification.	
Myrsine australis, a small evergreen shrub or tree native to New Zealand, is extensively utilized in horticulture and ecological conservation. Its leaves typically exhibit an elliptical or lanceolate shape, and the plant demonstrates rapid growth and high adaptability, yielding red or black berries upon maturity.	

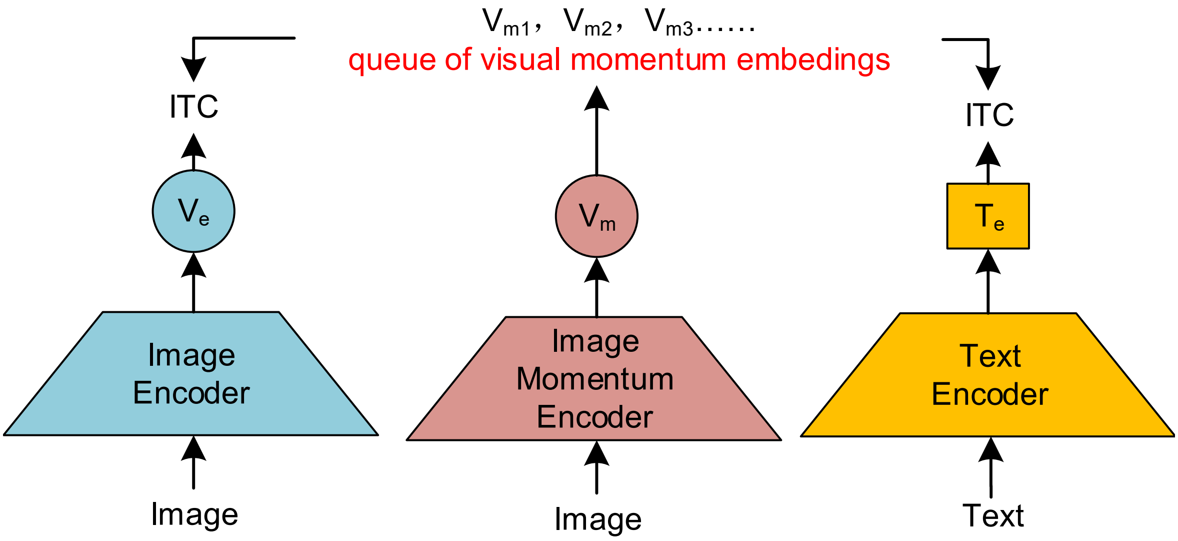


Figure 6. The training process of the momentum encoder.

To achieve cross-modal question-answering, we integrated perspectives from multiple tasks, including image recognition, cross-modal retrieval, and visual question-answering. By considering the characteristics of large-scale, fragmented conservation image data and the complexity of cross-disciplinary knowledge integration in forestry ecology literature, we proposed a novel method that centers cross-modal clustering around factual information within images. This approach not only enhances the model's understanding of images (as reflected in improved accuracy in image recognition and cross-modal retrieval) but also provides cross-modal validation for meta-analyses in forestry ecology, thereby improving the interpretability of the model's outputs. The specific details will be elaborated in Section 3.2 in conjunction with the construction of the dataset.

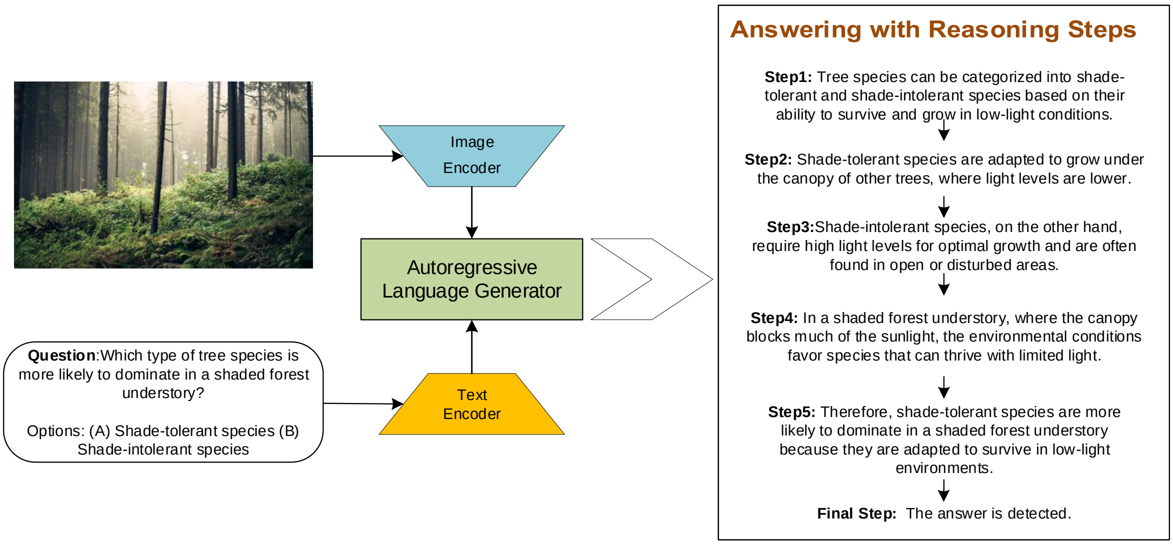


Image Encoder
Autoregressive Language Generator
Text Encoder
Answering with Reasoning Steps

Question:Which type of tree species is more likely to dominate in a shaded forest understory?

Options: (A) Shade-tolerant species (B) Shade-intolerant species

Step1: Tree species can be categorized into shade-tolerant and shade-intolerant species based on their ability to survive and grow in low-light conditions.

Step2: Shade-tolerant species are adapted to grow under the canopy of other trees, where light levels are lower.

Step3: Shade-intolerant species, on the other hand, require high light levels for optimal growth and are often found in open or disturbed areas.

Step4: In a shaded forest understory, where the canopy blocks much of the sunlight, the environmental conditions favor species that can thrive with limited light.

Step5: Therefore, shade-tolerant species are more likely to dominate in a shaded forest understory because they are adapted to survive in low-light environments.

Final Step: The answer is detected.

Figure 7. The instances of question-answering.

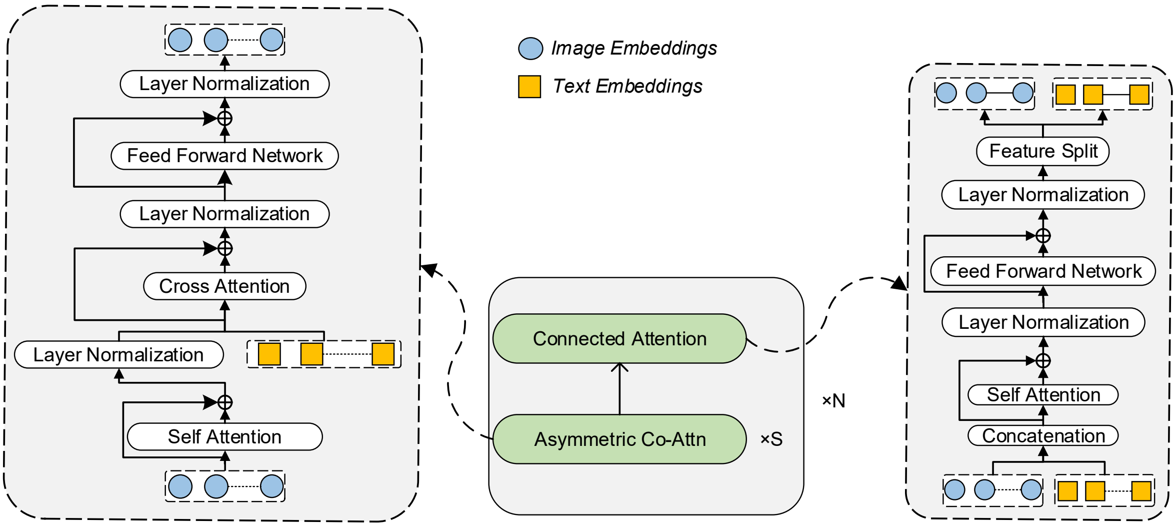


Figure 8. Deep neuro network for fusion of image and text features.

In summary, we employ a pair of CLIP-based cross-modal encoders to project image and text data into a shared vector space for unified processing. Once the deep learning model has been trained to capture the statistical patterns within the data, it can handle larger-scale datasets, enabling the exploration of forestry ecology literature and images across longer time spans and broader spatial scales. Furthermore, by leveraging chain-of-thought techniques, the model can learn to perform reasoning, allowing users to retrieve the information and knowledge acquired by the model through a question-and-answer manner. To illustrate how we leverage deep learning techniques for on-demand extraction of forestry ecology knowledge through question-answering using images and text as inputs, we have designed a pipeline, as shown in Figure 3, divided into four stages. The content of forestry ecology literature primarily comprises three data formats: text, illustrations, and tables. We retained only the text and tables, converting the tables into text for uniform processing, as depicted in stage 1 of Figure 3. The "Language Generator" in the figure represents the natural language generator responsible for parameterizing forestry ecology literature and providing readable natural language output for users. After processing in stage 1, the Language Generator learns how to handle tables within the literature. Subsequently, in stage 2, the Language Generator learns all content from the forestry ecology literature(the literature dataset we used is detailed in Appendix A.3) except for illustrations, completing the initial parameterization process. In stage 3, we introduce image data. To enable the model to process both images and text simultaneously, we project the data from these two modalities into a unified vector space using a pair of momentum encoders. These two encoders are trained jointly using a contrastive learning approach, wherein they are compared to each other during training. The goal of contrastive learning is to ensure alignment between image-text pairs after vectorization, such as aligning images of various species with their corresponding textual descriptions after projection (see Section 3.4 for details). To ensure that the encoders capture the features of the data rather than the differences between the encoders themselves, we employed a momentum encoding method. This method allows the encoders to learn from each other's evolving process, rather than simply learning the final outcomes. Essentially, this involves taking a moving average of the encoders' long-term changes (see Section 3.5 for more details). stage 4 involves training the forestry ecology question-answering model. In this stage, the input images and text for the question-answering task are first processed by the encoders from stage 3 to achieve cross-modal information fusion. The fused information is then passed to the language generator for decoding, producing a natural language output, i.e., the "answer", which users can directly read. In summary, the entire process involves the model autonomously reading forestry ecology literature, projecting images and text in the forestry ecology domain into a shared vector space for computation, and generating a professional forestry ecology answer through the model by fusing the cross-modal information of the input images and text during the question-answering phase.

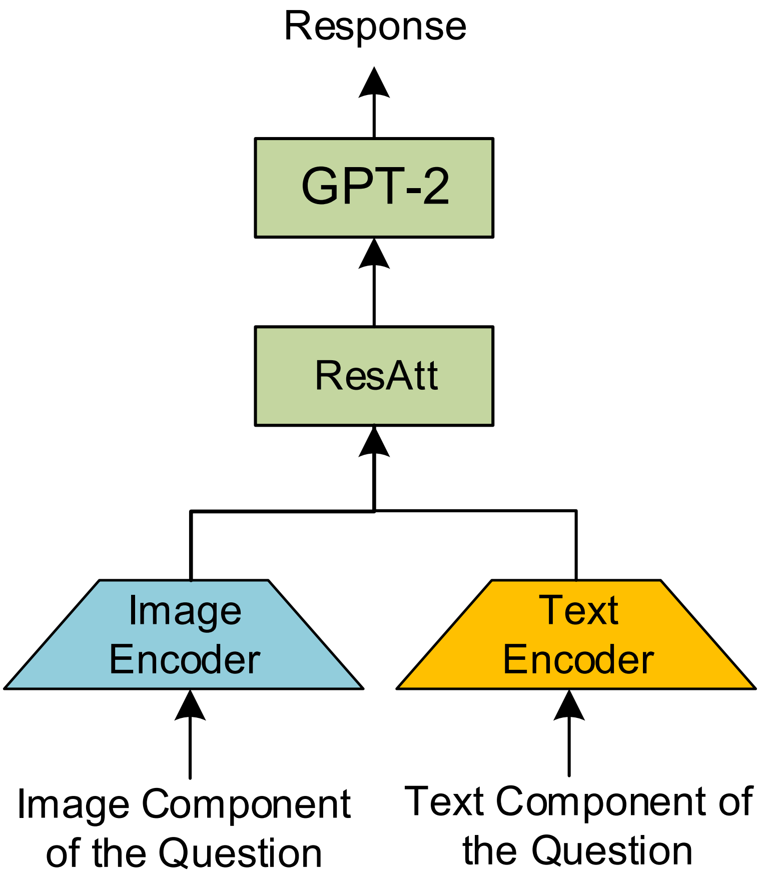

Response
GPT-2
ResAtt
Image Encoder
Text Encoder
Image Component of the Question
Text Component of the Question

Figure 9. Model Structure During Inference.

3.1. Problem Definition and Preparation

In this section, we predefine the variables shown in Figure 3 from the perspective of deep learning and lay the groundwork for the corresponding model design details. The input question consists of both an image and text. The image is fed into a cross-modal image encoder to extract features, which are then combined with the corresponding token sequence of the text and fed into a language model to generate the answer. Let the language model be denoted as LM, the input image as I, the token sequence of the input text as T_i, the cross-modal image encoder as $Encoder_v$, and the generated answer as A. Therefore, the QA (question-answering) process can be expressed as $A = LM(Encoder_v(I), T_i)$, as illustrated in Figure 3 stage 3.

The three main forms of data within the literature include text, images, and tables. To address this, we performed a two-stage preprocessing on the language generation model, as shown in stages 1 and 2 in Figure 3. The language model we utilized is GPT-2 [36]. In stage 1, we fine-tuned GPT-2 following the method proposed in Reference [37]. In stage 2, we trained the language model GPT-2 in an autoregressive manner to parameterize the literature dataset. The method we propose is implemented by the momentum encoder shown in stage 3, and its effectiveness is validated through experiments on the question-answering model shown in stage 4. Specifically, after stage 1, the model gains the ability to convert tables into text. In stage 2, we fine-tuned the language model with the text and tables from the literature. At this point, the text and tabular data from the literature are parameterized into the knowledge of the language model. The next step is how to extract information on demand in a question-answering manner. To enable the question-answering model to understand the context formed by images and language, we mapped them into a shared semantic space. This step is achieved by the momentum encoder shown in stage 3. To train this pair of encoders, we constructed an image-text pair dataset containing images of animals and plants along with their corresponding species names and descriptions, refer to Section 3.2 for details.

3.2. Underlying Logic of NACID Dataset Construction

To ensure that our designed deep learning model can accurately identify species within images and provide corresponding forestry ecology descriptions, we made improvements to publicly available species datasets. Specifically, we utilized the iNaturalist 2017 dataset [38], which consists of 5089 species, with 579,184 training set images and 95,986 validation set images. The sample sizes for each species classification and the distribution of samples in the training and validation sets are presented in Table 2. Examples of the original sample instances from the iNaturalist 2017 dataset are shown in Table 3.

However, the dataset does not include text descriptions corresponding to the images. In order to generate text descriptions paired with images we followed the pipeline of Laion COCO 600 M [39] to curate our Nature Conservation Image-text Pair Dataset (NACID). For specific details, please refer to Appendix A.4. Sample examples are shown in Table 4.

Training the encoder pair shown in Figure 3 directly on the native iNaturalist 2017 dataset, as illustrated in Table 3, yields cross-modal embedding results as depicted in Figure 5. This type of embedding is suitable only for image classification. However, to fully integrate images and literature, a more richly annotated image-text pair dataset, such as the one shown in Table 4, is required. The challenge lies in determining how the textual content of forestry ecology literature should be correlated with image content, which is currently unknown. This uncertainty is the reason for introducing a multi-task perspective, as language models (e.g., GPT-2) are inherently aware of correlations between textual data. This serves as the basis for our approach to classifying literature based on factual information present in images. As the model is applied and receives user feedback, further refinement of the cross-modal representation clustering, as shown in Figure 4, can be achieved through expert calibration, leading to clusters centered on factual content within images. The primary motivation behind our effort to enrich image annotation with more diverse corpora is to leverage the knowledge embedded in models trained on various tasks for cross-modal information fusion. Our objective is to achieve cross-modal clustering centered on visual facts.

3.3. Cross-Modal Embedding Clustering

Language expression is inherently ambiguous; the same image can be interpreted differently across various disciplines, and even within the same discipline, different perspectives can yield different interpretations. Given that the factual information contained within an image remains constant, we use this as the centroid (clustering center) to optimize the cross-modal shared space. Specifically, various interpretations centered around the same visual information are clustered together. This approach mitigates the impact of ambiguity on the performance of the forestry ecology question-answering model.

To map the image-text pairs depicted in Table 4 into a shared semantic space, we utilized a contrastive learning approach to train a pair of encoders. The visual encoder is responsible for extracting image features, while the language encoder extracts text features. The purpose of contrastive learning is to bring similar feature points closer together in the representation space, and conversely, to push dissimilar points farther apart. Specifically, the text description corresponding to an image serves as a positive sample, while all other text descriptions act as negative samples. Through training, the model can measure the degree of match between images and their corresponding textual descriptions. We begin by elucidating the contrastive learning loss function, followed by a discussion on the design process of the cross-modal momentum encoder.

3.4. Contrastive Learning Loss Function

For a deep learning model to effectively integrate image data with forestry ecology literature, the foundation is cross-modal alignment. Specifically, the species information contained in an image needs to be aligned with the professional terminology of the forestry ecology field. For instance, the image information of an ice plant needs to be aligned with the term Mesembryanthemum crystallinum. In this section, our goal is to align the information of interest within the images to their interpretations in forestry ecology.

Following ConVIRT [40], the image-text contrastive learning loss function (ITC) is formulated based on InfoNCE [41]. As illustrated in Table 4, assuming there are N pairs of image-text pairs in the dataset, corresponding to N pairs of vectors, where the vectors for the image are denoted as $V_1, V_2, \ldots, V_N$ and for the text as $T_1, T_2, \ldots, T_N$ (an arbitrary query text). We denote the vector for the $i-th$ image-text pair as (V_i, T_i), where T_i is the positive sample for V_i, and $T_j (j \neq i)$ is the negative sample for V_i. We calculate the visual-to-language loss by multiplying V by T, as shown in Equation (1); and the language-to-visual loss by multiplying T by V, as shown in Equation (2). Because cross-modal contrastive learning is asymmetric, we perform vector similarity calculation in two directions. The contrastive loss for the $i-th$ pair in the image → text direction:

$$\ell_i^{(V \rightarrow T)} = -\log \frac{\exp\left(S_C(\mathbf{V}_i, \mathbf{T}_i)/\tau\right)}{\sum_{k=1}^{N} \exp\left(S_C(\mathbf{V}_i, \mathbf{T}_k)/\tau\right)}, \tag{1}$$

where $S_C(\cdot)$ is the cosine similarity, i.e., $S_C(a, b) = a^\top b/(\|a\|\|b\|)$, and τ is a temperature hyperparameter. Similarly, an image-to-text loss for a single pair is calculated as follows:

$$\ell_i^{(T \rightarrow V)} = -\log \frac{\exp\left(S_C(\mathbf{T}_i, \mathbf{V}_i)/\tau\right)}{\sum_{k=1}^{N} \exp\left(S_C(\mathbf{T}_i, \mathbf{V}_k)/\tau\right)}. \tag{2}$$

where $k \in \{1, 2, \ldots, N\}$, while i is a specific value within $\{1, 2, \ldots, N\}$.

Finally, the training objective is a weighted sum:

$$\mathcal{L}_{\textbf{ITC}} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda \ell_i^{(V\to T)} + (1-\lambda)\ell_i^{(T\to V)}\right), \tag{3}$$

where $\lambda \in [0,1]$ is a hyperparameter to balance vision and language in the total loss function, and in practice N refers to the batch size.

3.5. Cross-Modal Momentum Encoder

As shown in stage 3 of Figure 3, the role of the pair of encoders is to project the data features of each modality into a shared space based on their relevance. For example, the image features of an ice plant should be similar to the descriptions related to Mesembryanthemum crystallinum in forestry ecology literature, while being marked as unrelated to descriptions of other species. This is achieved through contrastive learning of the outputs from the two encoders. However, in practical implementation, the updates of these two encoders are asynchronous, leading to the model learning the differences between the encoders rather than the differences in the data itself. To improve the accuracy of cross-modal feature comparison, we employ momentum encoding to suppress the update speed of the encoders, ensuring that the model learns the inherent relevance of the data itself rather than the discrepancies between the encoders.

CLIP [31] stands as a seminal work in training image-text cross-modal encoders using the contrastive learning method described in Section 3.4, serving as the foundational model in this domain. We utilized the rich open-source resources of openCLIP [42] and followed the momentum encoding concept of MoCo [43] to design our cross-modal momentum encoder. The visual encoder utilized ViT-B/32 [44], while the language encoder employed BERT [45]. Both encoders consisted of 12 layers of transformers.

The pre-trained models we employed, including the language generator GPT-2 and the cross-modal encoder CLIP, were trained on large-scale internet datasets. For instance, CLIP's training set comprises 400 million image-text pairs, encompassing a vast array of prior linguistic knowledge. One result of this is that for the same image, the model can generate descriptions from multiple perspectives, such as physics, biology, sociology, and so forth. These descriptions are intertwined in the language representation space, compounded by the compressed nature of language, which exhibits polysemy across different contexts. Furthermore, question-answering poses a challenging reasoning task, presenting significant challenges to generative models.

In our proposed method, the language generation model serves as both an extractor from the model's learned knowledge and a crucial human–computer interaction module. Therefore, enhancing the accuracy of the language generation model is pivotal for on-demand extraction of parameterized literature.

The contrastive learning described in Section 3.4 is a widely adopted method. However, when two encoders engage in mutual contrastive learning and one of them exhibits unstable representation outputs, such as rapid changes in the representation of certain samples, it leads to the other encoder learning only the differences between the encoders rather than the differences between the samples themselves. This ultimately causes the model to converge via shortcuts. The consequence is that the encoders fail to capture the crucial information in the data, thereby affecting the quality of feature vectors in the embedding space and the relationships between them.

Therefore, we adopted the momentum encoder concept, as shown in Equation (4):

$$\theta \leftarrow m\theta_k + (1-m)\theta_q, \tag{4}$$

where $m \in [0,1)$ is the momentum hyperparameter. The query encoding θ_q is updated based on gradient backpropagation. Typically, m takes a value greater than 0.9, which is equivalent to taking a moving average of the encoding updates. In other words, over 90% of the features extracted by the encoder in the next iteration are inherited from the previous iteration. This concept enables one of the two encoders engaged in contrastive learning to generate stable, continuously changing features as pseudo-labels for the other encoder. And it minimizes the interference of encoder noise on sample features.

Like most deep learning methods, the essence of contrastive learning lies in classification. The difference between classification tasks lies in whether the centroids are set manually or based on comparisons between samples. For contrastive learning, the number of samples determines the number of categories. However, image-text pair datasets comprise two modalities: images and text. Therefore, the first consideration is whether to use the image modality or the text modality as centroids. Given that photographs in the natural world contain factual information, whereas scientific and technological knowledge in literature is predominantly encoded in text. Considering the high degree of information compression, polysemy, and diverse perspectives in language, we opt to use images as centroids, enabling knowledge to cluster around facts. Building upon these ideas, we propose a cross-modal embedding clustering method. The specific steps are outlined as follows:

1. Utilize a pair of ViT-B/32 image encoders to initially classify images in the NACID dataset. One of the encoders adopts momentum updates and stores its outputs in a first-in-first-out (FIFO) queue q in chronological order as pseudo-labels;
2. Train the other image encoder using contrastive learning methods based on the pseudo-labels in q;
3. Train a language encoder using contrastive learning methods based on the pseudo-labels in q.

Ideally, when extracting features from each sample using contrastive learning methods, it is preferable to compare it with all samples, i.e., the length of the queue q should equal the number of samples in the training set. However, as the training set size increases, the batch size differs significantly from the length of q, making it impractical to train an online momentum encoder based on contrastive learning. To address this issue, we store the queue q offline and load it during training, thus decoupling a large queue from a small training batch. The training method for the cross-modal momentum encoder is illustrated in Figure 6.

In Figure 6, V_e represents visual embedding, V_m represents visual momentum encoding, T_e represents text embedding, and ITC represents the contrastive learning loss function. Taking q with a length of 4096 and a batch size of 32 during training as an example, in the training process, the momentum encodings of the 32 samples learned are removed from the queue according to the FIFO principle, and the newly generated momentum encodings are added to the queue. The encoder learns samples with semantic consistency in chronological order, thereby reducing interference caused by encoder fluctuations. This method not only enhances the quality of feature extraction but also ensures a reasonable distribution of feature points in embedding space. Continuing, we feed the prepared encoder with visual and language features extracted from the question part into an autoregressive language generator. The language generator predicts subsequent expressions based on the input question encoding.

3.6. Autoregressive Language Generator

We selected an autoregressive language model to learn from forestry ecology literature and employed a human–machine question-answering format for knowledge extraction. Specifically, we supplemented the ScienceQA dataset [46] with 2000 forestry ecology visual question-answering (VQA) samples to train our model to extract knowledge and information in a question-and-answer format. An example of our augmented dataset is shown in Figure 7.

Letting the textual part of the question be denoted as Q_t, the image part as Q_v, and the subsequent expressions generated word by word by the autoregressive language model as $w_1, w_2, \ldots, w_k$, the generation of any word $w_i (i > 2)$ is represented as Equation (5).

$$\mathcal{L}_{\text{LM}} = \log p(w_i | Q_t, Q_v, w_1, w_2, \ldots, w_{i-1}). \tag{5}$$

In the context of Figure 7, let Step 1 be denoted as S1, Step 2 as S2, and so forth. Then, the generation of any macroscopic Step Si can be represented as shown in Expression 6.

$$\mathcal{L}_{\mathrm{LM}} = \log p(S_i|Q_t, Q_v, S_1, S_2, \ldots, S_{i-1}). \tag{6}$$

In other words, from the macroscopic perspective of human–computer interaction, autoregressive generation proceeds step by step, while from the microscopic perspective, it is generated word by word.

Due to the large number of pixels required to carry a certain amount of information in image data, which exhibits sparsity, compared to images, language has a significantly higher degree of information compression. In the question-answering process, as the language sequences within the context become longer, sparse visual information is overwhelmed. To leverage the centroid role of visual representation and improve the accuracy of language generation model predictions, we performed fusion before feeding the encoded image-text pairs to the language generation model. For this purpose, we designed a cross-modal feature fusion deep neural network.

3.7. Cross-Modal Feature Fusion Network

Building on the aforementioned cross-modal representation, we project the images and text input by the user into a shared space. Before feeding this input into the language generation model, we perform cross-modal feature fusion to enable the model to generate a more accurate response based on the input data. Following mPLUG [47], we designed the cross-modal feature fusion deep neural network as shown in Figure 8.

The primary modules as shown in the middle of Figure 8 include the asymmetric co-attention block (referred to as AC) and the connected attention block (referred to as CA), where every S layer of AC is followed by a CA to form a skip-connection module. In practice, a cross-modal feature fusion network is constructed by stacking N skip-connection modules. In the AC block, visual features are predominantly considered, while language features are supplementary. Through multiple residual connections and gradual fusion, the output comprises visual features integrated with language information. The repeated residuals enhance the weight of sparse visual information during the feedforward process, ensuring that visual information remains prominent despite module stacking and increasing language context. In the CA block, visual and language features exchange information in the feedforward attention network, ultimately yielding fully integrated visual and language feature outputs. In practice, these outputs take the form of vector sequences. We refer to the feature fusion network shown in Figure 4 as ResAtt (residual and attention).

3.8. Cross-Modal Question-Answering Model

The model architecture during usage is illustrated in Figure 9. When the user inputs an image and text, these inputs are processed by an image encoder and a text encoder, respectively, to extract their features. These features are then projected into a shared space. After cross-modal feature fusion, the combined features are fed into a language generation model (GPT-2) to generate an answer for the user.

3.9. Settings

The experimental setup includes 4 Nvidia RTX 3090 24 G GPUs, with a learning rate of 1×10^{-4}, temperature coefficient of 0.07, loss function coefficient λ set to 0.75, momentum queue of 4096, the momentum set to 0.999, batch size set to 32 and the AdamW optimizer [48] chosen for optimization. We conducted further manual optimization on the NACID dataset proposed in Section 3.2. For forestry ecological question-answering, we created templates for questions and curated question-answer samples to fine-tune the model for knowledge extraction in forestry ecology. Additionally, we created a vector database for the abstract content and BibTeX format data of each document, as detailed in Appendix A. In summary, to validate the effectiveness of the proposed method and facilitate comparison with previous approaches, we selected the image classification task on the iNaturalist 2017 dataset to test the efficacy of our image encoder. We assessed its performance by calculating recognition accuracy and compared it with other methods on this dataset. To evaluate the cross-modal alignment performance of our image encoder and text encoder, and to facilitate comparison with similar methods, we selected the Cross-modal Retrieval task on the MS COCO dataset. We employed Recall at K (R@K) as the standard evaluation metric for retrieval tasks. This metric measures the proportion of queries where the correct matches are found within the top K retrieved results. A higher R@K value indicates better performance.

4. Experiments and Results Analysis

In this section, we conducted experiments on standardized tasks using public datasets, verifying the effectiveness of the basic modules through performance comparisons with similar models. Specifically, to assess the impact of cross-modal alignment, we performed species image recognition experiments on the iNaturalist 2017 dataset, as shown in Table 5, and cross-modal retrieval experiments on the MSCOCO dataset, as shown in Table 6. This set of experiments aims to verify whether the alignment between images and text is achieved after projecting them into a shared space using a pair of momentum encoders. We also conducted multi-species text-to-image retrieval on the NACID dataset, as shown in Table 7. The purpose of this set of experiments is to validate the clustering of language descriptions around images in the cross-modal space through cross-modal correlation search. After validating the fundamental functionality of the model, we performed cross-modal question-answering experiments on the public ScienceQA dataset and compared our results with leading models on this dataset, as shown in Table 8. Finally, we demonstrated the performance of our proposed method in completing cross-modal question-answering tasks in forestry ecology, as illustrated in Figure 10. The purpose of this set of experiments is to validate the ability of a pair of momentum encoders and the language generation module to collaboratively perform question-answering inference.

Table 5. Comparison of image classification on iNaturalist 2017 (%).

Method	Top1 Accuracy
MetaFormer [49]	80.4
FixSENet-154 [50]	75.4
SEB+EfficientNet-B5 [51]	72.3
TransFG [52]	71.7
IncResNetV2 SE [38]	67.3
SpineNet-143 [53]	63.6
MetaSAug [54]	63.3
Graph-RISE [55]	31.1
PaMA	82.1

Table 6. Quantitative analysis of cross-modal retrieval on MS COCO (%).

Method	Retrieval I2T			Retrieval T2I		
	R@1	R@5	R@10	R@1	R@5	R@10
Oscar [56]	57.5	82.8	89.8	73.5	92.2	96.0
METER [57]	57.1	82.7	90.1	76.2	93.2	96.8
ViSTA [58]	52.6	79.6	87.6	68.9	90.1	95.4
ALADIN [59]	51.3	79.2	87.5	64.9	88.6	94.5
PaMA	60.8	83.8	91.3	75.7	92.6	96.3

Table 7. Top 3 Cross-Modal Retrieval Instances from Text to Image.

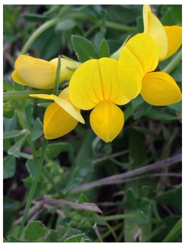
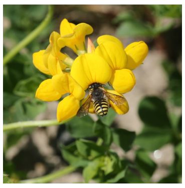
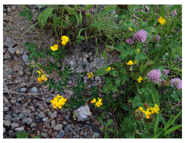
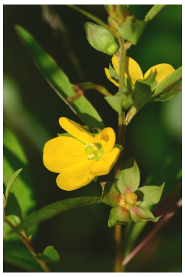
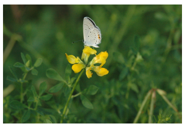
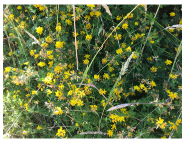
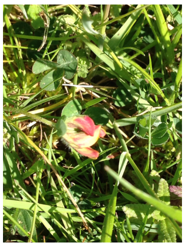
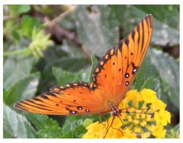
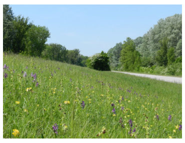
Query	**Close-up Images of Lotus Corniculatus Flowers Showing Morphological Variations.**	**Interactions between Lotus Corniculatus and Pollinators in Its Natural Habitat.**	**Lotus Corniculatus Coexisting with Other Species.**
Top1 Result			
Top2 Result			
Top3 Result			

Table 8. Model Evaluation Results on ScienceQA Test Split (%).

Model	**NAT**	**SOC**	**LAN**	**TXT**	**IMG**	**NO**	**G1-6**	**G7-12**	**AVG**
Human [46]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [46]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
GPT-4 [60]	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaMA-Adapter [61]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
$MM - CoT_{Base}$ [62]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
PaMA	95.63	85.84	91.32	93.19	89.37	94.15	93.92	89.62	91.63

4.1. Main Results

We abbreviate the method proposed as PaMA (Parameterization before Meta-Analysis). To validate the effectiveness of the model, we conducted experiments on standardized tasks using public datasets and compared them with state-of-the-art (SOTA) models on the Leaderboard. Firstly, to validate the effectiveness of the image encoder, we performed image classification tasks on the iNaturalist 2017 dataset [38] and compared them with SOTA models on the Leaderboard. Secondly, to validate the effectiveness of the cross-modal encoder, we conducted image-text cross-modal retrieval experiments on the MSCOCO dataset [63] and compared them with SOTA models on the Leaderboard. Lastly, to validate the effectiveness of the question-answering model, we conducted experiments on the ScienceQA dataset [46] and compared them with SOTA models on the Leaderboard.

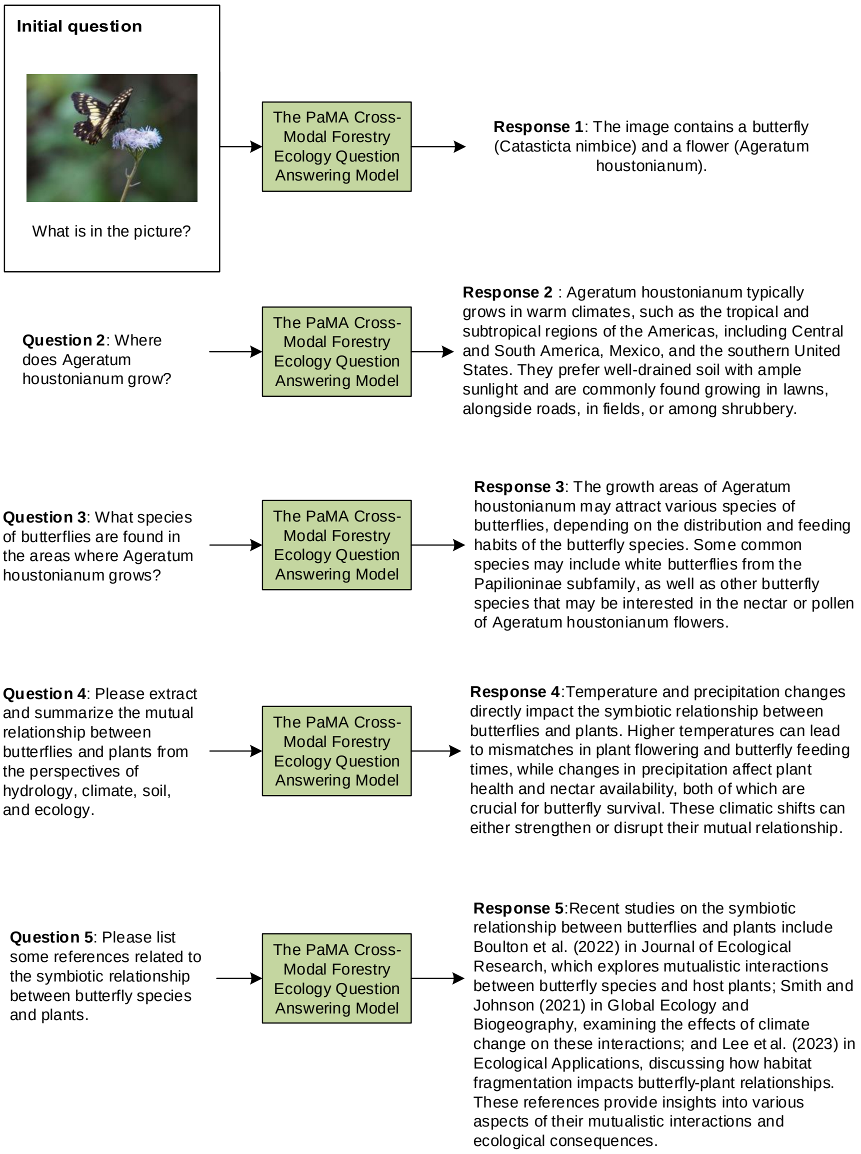

Initial question
What is in the picture?
The PaMA Cross-Modal Forestry Ecology Question Answering Model
Response 1: The image contains a butterfly (Catasticta nimbice) and a flower (Ageratum houstonianum).
Question 2: Where does Ageratum houstonianum grow?
Response 2 : Ageratum houstonianum typically grows in warm climates, such as the tropical and subtropical regions of the Americas, including Central and South America, Mexico, and the southern United States. They prefer well-drained soil with ample sunlight and are commonly found growing in lawns, alongside roads, in fields, or among shrubbery.
Question 3: What species of butterflies are found in the areas where Ageratum houstonianum grows?
Response 3: The growth areas of Ageratum houstonianum may attract various species of butterflies, depending on the distribution and feeding habits of the butterfly species. Some common species may include white butterflies from the Papilioninae subfamily, as well as other butterfly species that may be interested in the nectar or pollen of Ageratum houstonianum flowers.
Question 4: Please extract and summarize the mutual relationship between butterflies and plants from the perspectives of hydrology, climate, soil, and ecology.
Response 4:Temperature and precipitation changes directly impact the symbiotic relationship between butterflies and plants. Higher temperatures can lead to mismatches in plant flowering and butterfly feeding times, while changes in precipitation affect plant health and nectar availability, both of which are crucial for butterfly survival. These climatic shifts can either strengthen or disrupt their mutual relationship.
Question 5: Please list some references related to the symbiotic relationship between butterfly species and plants.
Response 5:Recent studies on the symbiotic relationship between butterflies and plants include Boulton et al. (2022) in Journal of Ecological Research, which explores mutualistic interactions between butterfly species and host plants; Smith and Johnson (2021) in Global Ecology and Biogeography, examining the effects of climate change on these interactions; and Lee et al. (2023) in Ecological Applications, discussing how habitat fragmentation impacts butterfly-plant relationships. These references provide insights into various aspects of their mutualistic interactions and ecological consequences.

Figure 10. A human–machine question-answering instance.

As shown in Table 5, we conducted experiments on the image classification task of the iNaturalist 2017 dataset, comparing it with the SOTA methods on the leaderboard. The experimental results show that the image encoder trained using the momentum method proposed in Section 3.5 achieved a top-1 accuracy higher by 1.7% to 51% compared to similar methods, demonstrating the effectiveness of momentum encoding.

The performance of the image encoder determines whether image information can be accurately extracted and expressed in cross-modal question-answering tasks within forestry ecology. In other words, if the image encoder lacks sufficient accuracy in visual object recognition and does not possess high discriminative power for species identification within the domain of forestry ecology, it will not support the cross-modal question-answering tasks of this study. Comparing our method with similar approaches on standardized tasks in public datasets can more rigorously validate its effectiveness and performance level. The results indicate that the image encoder in our proposed method can accurately identify visual objects in images, ranking among the top methods. The selected iNaturalist 2017 dataset, which includes 5089 species, demonstrates that our proposed image encoder can distinguish fine-grained visual species targets. In summary, the image encoder in our method is capable of encoding visual information for question-answering tasks in forestry ecology.

As mentioned in Section 3, our proposed method includes a pair of encoders: an image encoder for extracting visual features and a text encoder for extracting linguistic features. These encoders project the extracted features from both modalities into a shared space, aiming to bring similar features closer and push dissimilar ones further apart. To evaluate the performance of this pair of encoders, a suitable experiment is cross-modal retrieval. The goal of this experiment is to retrieve similar text from a text dataset given an image or to retrieve similar images from an image dataset given a text. Higher retrieval accuracy indicates better performance of the cross-modal encoders. Conducting standardized experiments on public datasets and comparing them with similar methods can both validate the effectiveness of the proposed method and measure its performance level. For these reasons, we chose the cross-modal retrieval experiment on the MSCOCO dataset, which has many comparable methods and a high degree of standardization, to verify the performance of our proposed pair of cross-modal encoders. The experimental results are shown in Table 6. PaMA achieved comparable performance to SOTA models on the leaderboard in both image-to-text and text-to-image retrieval tasks, demonstrating the effectiveness of our proposed cross-modal momentum encoder. The evaluation metric used is R@K, which reflects the accuracy of cross-modal retrieval. A higher R@K value indicates higher retrieval accuracy.

To more intuitively demonstrate the cross-modal retrieval capabilities of the model, the top 3 retrieval results for conservation image data using forestry ecology descriptions are presented in Table 7.

The experimental results indicate that the performance of the cross-modal encoders in our proposed method ranks among the top compared to similar approaches, ensuring the accuracy of data feature extraction in cross-modal question-answering tasks within forestry ecology. Specifically, this pair of cross-modal encoders can effectively represent data from their respective modalities and reflect the correlation between image and text features through cross-modal representation space embedding. Higher performance of the cross-modal encoders implies greater accuracy in extracting features from each modality and in computing cross-modal similarity. As shown in Table 7, the accuracy of cross-modal retrieval is generally high, although there are instances where errors occur, such as mistaking Ludwigia alternifolia for Lotus corniculatus. Nevertheless, the cross-modal alignment between the literature knowledge and the image data are semantically consistent. In summary, the pair of cross-modal encoders in our proposed method provides robust support for cross-modal question-answering in forestry ecology.

The experiments above indicate that using fact features within images as centroids for cross-modal representation clustering in a shared semantic space is feasible. The results in Table 5 demonstrate that the image encoder can effectively classify image data even after projecting visual features into the cross-modal space, indicating that the embedding of image features in space conforms to the distribution of similarities among individual samples in the image set. The results in Table 6 indicate that after projecting the representations of each modality into the shared space, the image encoder and text encoder can accurately calculate their mutual similarity, demonstrating the effectiveness of the proposed embedding clustering method.

After validating the effectiveness of the cross-modal encoder module, we proceeded to verify the overall effectiveness of the proposed method, specifically its performance in executing cross-modal question-answering tasks when all modules are integrated into a forestry ecology question-answering model. Consistent with the principles of the previous experiments, we conducted standardized experiments on public datasets. These experiments not only validated the model's effectiveness but also assessed the performance level of the proposed method by comparing it with similar approaches.

The scores in Table 8 represent the percentage of correct answers. The questions in the ScienceQA dataset are divided into several categories, with the publicly available leaderboard mainly listing the following question categories: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1–6 = grades 1–6, G7–12 = grades 7–12, the term "Avg" represents the average score across the aforementioned eight categories. Compared to GPT-3.5, PaMA achieved an average score that was 16.46% higher, with a particularly notable increase of 21.94% in the IMG category, demonstrating that the proposed method possesses the capability for cross-modal question-answering.

The performance of PaMA on QA tasks stems from two aspects: effective encoding and decoding of the model, as well as the effect of orthogonal information superposition. As seen in the data transformation process illustrated in Figure 9, the raw data are first extracted for features through an encoder, then pass through a feature fusion network, and are finally decoded by a language generator. In this encoding–decoding process, ensuring the effectiveness of both encoding and decoding is crucial for successfully completing the QA task. From the experimental results, it is evident that our proposed method achieves this. However, considering the performance of models like GPT-3.5 in the IMG category, as shown in Table 8, there is a significant difference compared to models incorporating visual information. What could be the reason for this? Firstly, in terms of model comparison, our method adopts GPT-2, which is not as proficient in natural language processing as GPT-3.5. Secondly, all comparisons are made on the same ScienceQA dataset and task; the difference lies in our model's incorporation of momentum visual encoding and cross-modal feature fusion.

How does this difference enable a relatively weaker language model to exhibit better performance? The primary reason is the effect of orthogonal information superposition. Because GPT fundamentally involves prediction, as indicated in Equation (5), more known information leads to lower information entropy and higher prediction accuracy. Furthermore, the information content in the same dataset remains constant. Moreover, since GPT-3.5 outperforms GPT-2, it can extract more information to a greater extent from the same dataset. In summary, our method's ability to significantly surpass GPT-3.5 in the IMG category is due to effectively introducing information gained from visual features.

From the perspective of cross-modal representation spaces, the effective superposition of visual and linguistic information depends on two factors. Firstly, whether the encoder can extract features from data samples to the maximum extent, and secondly, the cross-modal shared semantic space embedding, as it determines whether the embedding positions of feature points in the space better reflect the data distribution and their similarities. These two aspects are precisely addressed by the cross-modal momentum encoder, also known as cross-modal embedding clustering. In other words, the cross-modal momentum encoder is an effective method for performing cross-modal embedding clustering.

Then, to what extent does the fusion module ResAtt play a role in the process of cross-modal information superposition between visual and linguistic features? To address this, we conducted two ablation experiments.

4.2. Ablation Study

In cross-modal question-answering tasks within forestry ecology, the model needs to accurately distinguish between species with high intra- and inter-species similarity in visual information. Additionally, it must align these fine-grained visual details with specialized forestry ecology knowledge. To more precisely encode image and textual data in the forestry ecology domain, we optimized the model performance using a momentum encoder. To evaluate its optimization effect, we replaced the momentum encoder with two widely used visual encoders and compared their performance with and without the momentum encoder. This comparison validated the effectiveness of the momentum encoder and its contribution to the overall model performance. Before guiding the language model to generate answers for users, we performed cross-modal information fusion on the input-side visual and language features. To assess its contribution to the overall model performance, we conducted ablation experiments comparing models with and without the cross-modal fusion module. Our ablation experiments were conducted on 6532 samples that include both textual and image context, and the results of the ablation experiments are presented in Table 9.

Table 9. Accuracy (%) of different visual encoders with and without cross-modal feature fusion.

Model	With ResAtt	Without ResAtt
PaMA/ResNet	83.35	79.61
PaMA/CLIP	84.75	81.69

In Table 9, PaMA/ResNet represents replacing the momentum visual encoder with ResNet [64], PaMA/CLIP represents replacing the momentum visual encoder with CLIP [31], "with ResAtt" indicates the presence of a feature fusion network, and "without ResAtt" indicates the absence of a feature fusion network.

Compared to the 89.37% accuracy of PaMA in the IMG column of Table 8, the models with a feature fusion network in Table 9 experienced a decrease in accuracy of approximately 5% to 8%, while models without a feature fusion network saw a decrease of about 7% to 12%. In the "with ResAtt" column, the relatively lower performance decrease indicates that our proposed momentum encoder performs better than ResNet and CLIP when the feature fusion network is retained, highlighting the higher quality of visual features extracted by momentum-based contrastive learning methods. In the "without ResAtt" column, the relatively larger performance decrease shows that removing both the visual encoder and the feature fusion network leads to greater performance drops, demonstrating the indispensability of both the momentum encoder and the feature fusion network.

Moreover, the reasoning chain in the question-answering task is relatively long, resulting in a lengthy textual context. In such cases, our proposed feature fusion network with visual residual connections is more suitable for integrating visual and language features within extended textual context, thereby effectively reducing uncertainties during the language model's prediction process.

4.3. *Qualitative Analysis of Forestry Ecological Question-Answering*

A human–machine question-answering instance using PaMA, as shown in Figure 10.

The Response1 in Figure 10 illustrates the effective cross-modal alignment achieved after projecting visual and language features into a shared space. This alignment is demonstrated by the accurate interpretation of image-text pairs created for forestry ecology on the NACID dataset during the inference process. Furthermore, it indicates that the embeddings of feature points conform to the original similarity distribution of the dataset, thereby proving the effectiveness of the image encoder and text encoder in embedding clustering.

Responses 2–5 demonstrate the process by which the model extracts knowledge from the literature to respond to user queries. This experimental result indicates that PaMA's language generation module can accurately predict subsequent text based on preceding text and output fluent natural language. On the other hand, this also demonstrates the effectiveness of our proposed method, which utilizes factual information from images as centroids for cross-modal embedding. In terms of PaMA's model structure, two designs are crucial to achieving this effectiveness. First, the ResAtt module introduces residual connections for visual features, ensuring that visual information continually contributes to enhancing predictions in the language module during question-answering processes, rather than being overshadowed by increasingly lengthy language sequences. Second, the invariant factual information contained within images ensures invariance of clustering centroids in visual and language embeddings within the shared semantic space. This approach avoids issues such as centroid drift caused by linguistic ambiguity, which could lead to chaotic or illusory language generation in single-modality question-answering tasks due to high uncertainty [65]. The effectiveness of our proposed cross-modal momentum encoder ensures the quality of feature extraction and embedding in the shared semantic space, enabling orthogonal information superposition across modalities and reducing uncertainty during the language generation process, thereby improving model performance.

Reference [62] proposes a two-stage framework that first infers the rationale and then the answer, allowing the answer inference process to leverage cross-modal information provided by the rationale, resulting in more accurate answers. Reference [46] presents a method that first uses an image captioning model to map images to text, then employs this text to drive a language generation model for inference. This approach leads to information loss from the image data and lacks a shared cross-modal representation space, resulting in insufficient or inaccurate mutual information representation and computation. From the combined analysis of these two references, it is evident that cross-modal question-answering is a reasoning process that integrates vision and language, grounded in cross-modal representation space. These representations enable accurate and effective mutual information computation from data of different modalities. The rationale guides the reasoning process and plays a crucial role in enhancing the accuracy of cross-modal question-answering, as demonstrated in Reference [62] and our proposed method. Furthermore, the experimental comparison results shown in Table 8 indicate that our proposed cross-modal embedding clustering method performs better, demonstrating the effectiveness of our approach.

Admittedly, our proposed method also has its limitations. From the perspective of the model, its computations and outputs are black-box in nature, lacking interpretability. This makes it challenging to rigorously validate the reliability of the model's output during practical applications. From the perspective of training data, the public datasets used for model testing do not cover all scenarios in forestry ecology question-answering. Specifically, the datasets employed in the experiments, such as iNaturalist 2017, MSCOCO, and ScienceQA, do not share the same distribution as the data encountered in actual forestry ecology question-answering tasks. Even with the improvements made to the iNaturalist 2017 dataset by expert annotation as described in this study, it is still impossible to cover all real-world situations. Feeding such manually annotated data into AI models extends the subjective judgments of human experts, rather than creating universally applicable intelligent models. From the perspective of the tasks used for testing (such as image classification tasks and cross-modal retrieval tasks mentioned earlier), the predefined input-output relationships and evaluation standards constrain the ultimate behavior of the model. These constraints and evaluations are difficult to align with the conditions encountered in practical applications. The primary limitation of the multi-task perspective is that different tasks exhibit varying degrees of data fitting. Specifically, this manifests as the model overfitting on some tasks while underfitting on others. If a particular task leads the model to learn noise rather than meaningful features, it can further deteriorate model convergence. Moreover, balancing multiple tasks relies on empirical adjustments of a limited set of hyperparameters, which further increases the model's uncertainty. In summary, our research has limitations and biases in terms of interpretability, dataset distribution, and task design.

5. Discussion

We propose a method for extracting information and knowledge from forestry ecology literature and nature conservation images through a question-answering framework. This approach leverages cross-modal clustering, dataset construction, and chain-of-thought techniques to train a language model that integrates species-related visual data with forestry ecology literature, generating user-relevant answers based on logical reasoning. Experimental results on the standardized tasks of the ScienceQA dataset demonstrate that our method improves performance in visual question-answering tasks by 21.94 percentage points compared to GPT-3.5, validating the effectiveness of the proposed approach. Feature extraction and classification form the foundation of deep learning models, and to this end, we optimized our cross-modal encoder from a multi-task perspective. We conducted image recognition and cross-modal retrieval tasks on the iNaturalist 2017 and MSCOCO datasets, respectively. The experimental results show that our encoder outperforms similar methods, further validating the effectiveness of multi-task optimization for cross-modal encoders.

The interdisciplinary nature of the model provides an advanced decision-support tool for forestry practitioners and policymakers. Practitioners can extract relevant insights for sustainable forestry management, while policymakers benefit from evidence-based recommendations derived from complex ecological datasets. The integration of deep learning models with traditional ecological theories elevates the accuracy of ecological modeling. The model could be extended to estimate the value of ecosystem services (e.g., carbon sequestration, water filtration) by mapping forest health and ecosystem functions across diverse temporal and spatial scales. This is critical for establishing economic incentives for conservation. The existing model relies on static datasets, meaning it cannot dynamically update based on new data or user feedback after training. This makes it less adaptive to evolving ecological patterns, new research findings, or shifts in user information needs. Online learning allows the model to continuously refine its parameters in response to real-time data and interactions [66,67]. This enables the model to stay current with the latest ecological data and scientific literature, making its predictions and retrievals more relevant and timely. For example, if new satellite imagery shows sudden changes in forest cover, the model can quickly adjust its outputs to reflect these updates. The model may be prone to biases, especially if the training data are skewed toward specific regions, species, or ecological conditions. This limits its generalizability across diverse ecosystems and climates. By incorporating online learning, the model can gradually incorporate diverse datasets over time, learning from new data sources as they become available. This helps mitigate initial biases by ensuring the model is exposed to a broader range of ecological conditions, leading to improved generalization across different ecosystems. The static nature of the current model may not fully align with the dynamic needs of practitioners and policymakers, whose decisions are influenced by changing environmental, economic, and social conditions. With the ability to learn and adjust in real-time, the model can better support real-world decision-making. Practitioners can interact with the model, feeding it new information or datasets specific to their local context, which allows the model to generate more customized and actionable insights. This is particularly useful in a context like forest management, where conditions can change rapidly due to factors like climate variability or deforestation. The model's predictions are based on pre-trained patterns and may struggle to adapt to unseen or novel ecological phenomena, especially in regions with limited historical data. Online learning will enhance the model's predictive capabilities by allowing it to adapt to novel data in real-time. For instance, when faced with unfamiliar forest conditions due to unexpected environmental changes, the model can quickly adjust its predictions as new data streams in. This leads to more accurate forecasts of forest dynamics, species distributions, and biodiversity changes. The model lacks a mechanism to incorporate user feedback, meaning any misclassifications or irrelevant results cannot be corrected through interaction. The integration of reinforcement learning or other user-centered feedback mechanisms will allow the model to learn from user interactions. This creates a feedback loop where users can correct or guide the model's responses, improving the accuracy of future interactions. Over time, this leads to a more personalized and responsive model that better caters to the specific needs of different stakeholders, whether they are researchers, forest managers, or policymakers. In the domain of forestry ecology, where environmental conditions change (e.g., due to climate change or deforestation), models that adapt to new data through meta-learning [68] can provide more reliable and up-to-date predictions, supporting the need for further research in this direction.

6. Conclusions

Expert knowledge in forestry ecology literature offers the most accurate interpretation of conservation image data, while factual information contained in images naturally serves as the clustering center for such knowledge. We first employ a pair of encoders to project images and language into a shared vector space through cross-modal alignment. Considering the complexity of question-answering reasoning, we organize the logical relationships between visual information and literature knowledge from a multi-task perspective, optimizing the shared vector space embeddings. Experimental results demonstrate that our proposed cross-modal embedding clustering, which uses factual information from images as the clustering center, is an effective cross-modal classification method for the intricate species and complex forestry ecological knowledge. By adopting a multi-task perspective and focusing on the quality and distribution of cross-modal embeddings, systematic optimization can effectively enhance the performance of cross-modal question-answering models in forestry ecology.

Forestry ecological research is highly dependent on scale. As the statistical range of image data expands from a single conservation area to a global scope, as the number of literature sources grows from hundreds to millions, and as the temporal span extends from decades to centuries, new discoveries and research findings are bound to emerge. While researchers have limited time and energy, intelligent models empower us to leverage large-scale data, providing new tools for conducting extensive-scale forestry ecological studies. To this end, we attempt to integrate conservation image data with forestry ecology literature and provide a question-answering interface for on-demand knowledge and information retrieval, thereby exploring the infrastructure and solutions necessary for building intelligent forestry ecological big data systems.

Through this study, we propose that parameterization based on deep learning technologies is an effective method for integrating multimodal monitoring data from forestry ecological domains such as climate, hydrology, soil, and ecology. This method offers a solution for merging, analyzing, and statistically processing heterogeneous data, information, and knowledge, and is likely to become the foundation for constructing big data and intelligent infrastructure in forestry ecology.

Author Contributions: R.T. was responsible for model design, model training, dataset construction, and code writing and debugging. M.Z. was responsible for model design and code debugging. H.C. contributed to dataset annotation checking, article editing, and revision. H.-E.R. contributed to conceptualization, methodology, draft writing, reviewing, and providing experimental conditions, including the artificial intelligence laboratory and experimental equipment. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the project of Harbin Science and Technology Plan Research on Computer Vision Recognition Technology of Wood Species Based on transfer learning Fusion Model Project (ZC2022ZJ010027), the Young Doctoral Research Initiation Fund Project of Harbin University "Research on Wood Recognition Methods Based on Deep Learning Fusion Model" Project (HUDF2022110), the National Natural Science Foundation (62466012), and the Fundamental Research Funds for the Central Universities (2572017PZ10).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Acknowledgments: We would like to express our gratitude to Pan Lu et al. for sharing the ScienceQA dataset. We also extend our thanks to the ChatGPT platform for providing natural language processing services. I would also like to express my gratitude to Peng Hu from the College of Forestry at Northeast Forestry University for his guidance and assistance in the field of forestry ecology.

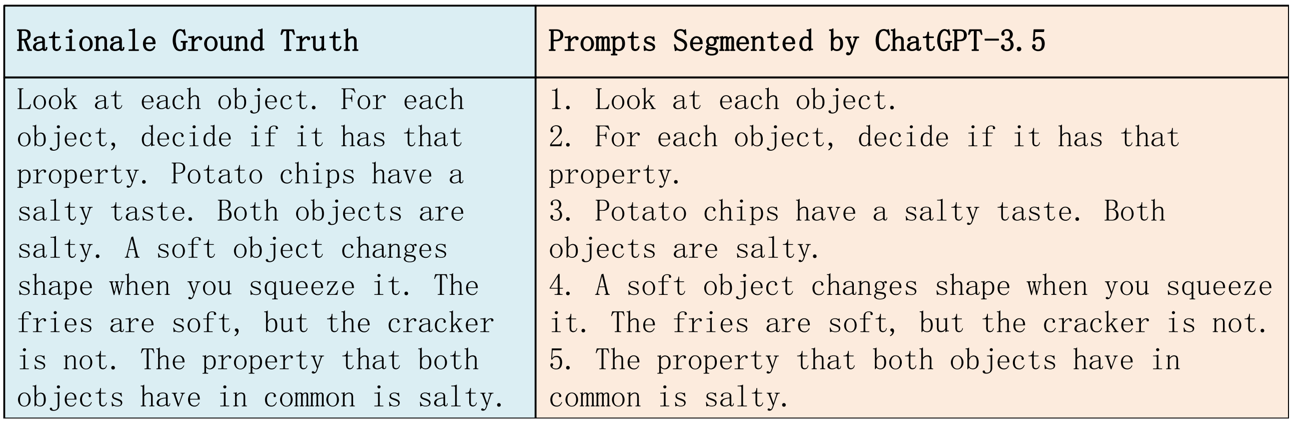
Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Construction and Optimization of Datasets and Databases

Appendix A.1. Optimization of Dataset

With the assistance of ChatGPT-3.5, we crafted short descriptions for each species object contained within the iNaturalist 2017 dataset's images. These descriptions were embedded into the context of the captions generated by BLIP [32] for each image. Consequently, the text descriptions for each image encompass not only the relationships between various objects within the image but also include the Latin names of the species and forestry ecological expertise, as illustrated in Figure 4.

For the ScienceQA dataset, we processed it using ChatGPT-3.5, as shown in Figure A1.



Rationale Ground Truth	Prompts Segmented by ChatGPT-3.5
Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.	1. Look at each object. 2. For each object, decide if it has that property. 3. Potato chips have a salty taste. Both objects are salty. 4. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. 5. The property that both objects have in common is salty.

Figure A1. Rationale ground truth splitting example for the ScienceQA dataset.

The purpose of this processing is to alleviate model hallucinations [62] because natural language expression is linear, and the semantics of words and sentences are constrained by context. Inference tasks are inherently difficult, and if the context spans are too long, it is inevitable that the reasoning process will become chaotic. To address this, we split longer rationale in the ScienceQA dataset into shorter sequences of sentences. This approach helps bridge the gap in model learning during reasoning. This step-by-step language model-based reasoning concept is inspired by chain-of-thought methodology [35].

Appendix A.2. Forestry Ecology Question Templates

After GPT-2 has learned in a self-supervised manner from a literature dataset, it struggles to perform question-answering tasks smoothly. It requires fine-tuning with a curated question database to meet the demands of these tasks. Our method begins with manually curating question-answer pairs, resulting in 1314 curated pairs. Subsequently, an evaluation algorithm is applied to automatically assess the model's responses. This process requires significant effort and professional support. However, based on our current knowledge, this step is indispensable. Although GPT-2 has assimilated information from the literature dataset, it does not grasp the logical relationships between question and answer contexts. It can extract information but fails to recombine internal knowledge from the literature into new relationships unless the context of the question aligns highly with what it has learned from the literature dataset. We constructed question templates by combining the three questions 'why, what, and how' with the four keywords hydrology, soil, climate, and ecology, to extract the knowledge learned by the model from the literature. Some example question templates are listed as follows:

1. Why is water important in forestry ecology?
2. What are the key factors influencing soil quality in forestry ecology?
3. How does climate change impact forestry ecology?
4. Why is biodiversity crucial in forestry ecology?
5. What are the interactions between water and soil in forestry ecology?
6. How do different climates affect forest ecosystems?
7. Why is understanding soil composition essential for forestry ecology?
8. What are the effects of climate variability on forest biodiversity?
9. How does water availability affect forest regeneration in forestry ecology?
10. Why is studying ecosystem dynamics important in forestry ecology?
11. What are the roles of soil nutrients in sustaining forest ecosystems?
12. How do ecological processes contribute to forest resilience under changing climatic conditions?

Appendix A.3. Literature Vector Database

The main sources for downloading literature are Web of Science, Google Scholar, and arXiv, spanning the years 1970 to 2020 and encompassing a total of 26,544 articles. Before being input into GPT-2, the textual data undergo several preprocessing steps. These include removing HTML tags, segmenting sentences, tokenizing words, and removing stop words and punctuation. The keywords used to index the literature dataset are shown in Table A1.

A total of 26,544 papers were selected through manual screening, representing the top 10 research areas. The distribution of papers in each field and their respective proportions are presented in Table A2.

Table A1. The Top 50 keywords in forest ecology articles.

Keywords	Sorted by Frequency, 5/Line
1–5	forest, diversity, conservation, dynamics, vegetation
6–10	biodiversity, patterns, growth, rain-forest, management
11–15	nitrogen, forests, soil, ecology, communities
16–20	carbon, climate change, ecosystems, disturbance, species richness
21–25	boreal forest, landscape, biomass, model, climate
26–30	fire, abundance, united-states, habitat, temperature
31–35	plants, organic matter, populations, decomposition, climate change
36–40	dispersa, responses, regeneration, tropical forest, land-use
41–45	habitat fragmentation, trees, fragmentation, forest soils, evolution
46–50	succession, deforestation, ecosystem, birds, population

Table A2. The top 10 research areas in forest ecology articles.

	Research Areas	Articles Number	Ratio (%)
1	Environmental Science Ecology	9352	35.23
2	Forestry	3949	14.88
3	Agriculture	2506	9.44
4	Plant Sciences	2408	9.07
5	Zoology	1941	7.31
6	Biodiversity Conservation	1802	6.79
7	Geology	1698	6.40
8	Meteorology Atmospheric Sciences	1096	4.13
9	Physical Geography	1036	3.90
10	Water Resources	756	2.85

To provide relevant literature responses in the question-answering context, we vectorized all 26,544 articles and established a literature vector database, with each vector having a dimensionality of 50,257. The process details are as follows:

1. The vocabulary of GPT-2 consists of 50,257 words;
2. We computed the TF-IDF (term frequency-inverse document frequency) [69] values for all real words in each article;
3. These values were arranged according to the positions of the corresponding real words in the vocabulary;
4. This process yielded the feature vectors for the respective articles.

Based on this, during human–machine questioning, literature lists can be pushed by calculating cosine similarity between the feature vector of the current context and the literature feature vectors in the database.

Appendix A.4. Details of the Expansion of the iNaturalist 2017 Dataset

We used the iNaturalist 2017 dataset updated on 15 February 2021. The dataset URL is "https://github.com/visipedia/inat_comp/tree/master/2017 (accessed on 18 August 2024)". For classification details, see Table 2, and for examples, see Table 3.

In order to generate text descriptions paired with images, we followed the pipeline of Laion COCO 600 M [39] to curate our Nature Conservation Image-text Pair Dataset (NACID) in four steps: (1) using BLIP L/14 to generate 40 captions for each image in iNaturalist dataset; (2) ranking them using Open AI CLIP L/14 to select the best 5 captions; (3) using Open AI RN50x64 CLIP model to select the best one; (4) using a small, fine-tuned T0 [70] model to roughly repair the grammar and punctuation of the text.

We obtained a dataset consisting of natural images and paired text descriptions, which are called captions. After that, we used the spaCy [71] method to recognize the predefined span types related to the categories of animals and plants. Then we followed the pipeline of entity name replacement [72] to further annotate the entities in captions with the fine-grained species names supported by the image classification ground truth of the iNaturalist dataset, such as *Heterotheca subaxillar*, *Ageratum houstonianum*, etc... Entity definitions are shown in Table A3, where AML represents animals, and ANT represents plant classification. Table 4 shows some samples of the final curated nature conservation image-text pair dataset.

Table A3. Applicable metadata for each entity type.

Entity Type	Applicable Types of Perturbable Spans
AML	<Animal-quantity> (e.g., a dog, two cats)
ANT	<Plant-quantity> (e.g., an apple, flowers)

References

1. Huang, H.G. Progress and perspective of quantitative remote sensing of forestry. *J. Beijing For. Univ.* **2019**, *41*, 1–14.
2. Zett, T.; Stratford, K.J.; Weise, F. Inter-observer variance and agreement of wildlife information extracted from camera trap images. *Biodivers. Conserv.* **2022**, *31*, 3019–3037. [CrossRef]
3. Abdusalomov, A.B.; Islam, B.M.S.; Nasimov, R.; Mukhiddinov, M.; Whangbo, T.K. An improved forest fire detection method based on the detectron2 model and a deep learning approach. *Sensors* **2023**, *23*, 1512. [CrossRef] [PubMed]
4. Yunusov, N.; Islam, B.M.S.; Abdusalomov, A.; Kim, W. Robust Forest Fire Detection Method for Surveillance Systems Based on You Only Look Once Version 8 and Transfer Learning Approaches. *Processes* **2024**, *12*, 1039. [CrossRef]
5. Newton, A. *Forest Ecology and Conservation: A Handbook of Techniques*; Oxford University Press: Cary, NC, USA, 2007.
6. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E5716–E5725. [CrossRef]
7. Gurevitch, J.; Curtis, P.S.; Jones, M.H. Meta-analysis in ecology. *Adv. Ecol. Res.* **2001**, *32*, 199–247.
8. Parmesan, C.; Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **2003**, *421*, 37–42. [CrossRef]
9. He, T.; Ding, W.; Cheng, X.; Cai, Y.; Zhang, Y.; Xia, H.; Wang, X.; Zhang, J.; Zhang, K.; Zhang, Q. Meta-analysis shows the impacts of ecological restoration on greenhouse gas emissions. *Nat. Commun.* **2024**, *15*, 2668. [CrossRef]
10. Benayas, J.M.R.; Newton, A.C.; Diaz, A.; Bullock, J.M. Enhancement of biodiversity and ecosystem services by ecological restoration: A meta-analysis. *Science* **2009**, *325*, 1121–1124. [CrossRef]
11. Koricheva, J.; Gurevitch, J.; Mengersen, K. *Handbook of Meta-Analysis in Ecology and Evolution*; Princeton University Press: Princeton, NJ, USA, 2013.
12. Ioannidis, J.P.; Greenland, S.; Hlatky, M.A.; Khoury, M.J.; Macleod, M.R.; Moher, D.; Schulz, K.F.; Tibshirani, R. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **2014**, *383*, 166–175. [CrossRef]
13. Huitema, B. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
14. Hunter, J.E.; Schmidt, F.L. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*; Sage: Newcastle upon Tyne, UK, 2004.
15. Lefebvre, C.; Glanville, J.; Briscoe, S.; Littlewood, A.; Marshall, C.; Metzendorf, M.I.; Noel-Storr, A.; Rader, T.; Shokraneh, F.; Thomas, J.; et al. Searching for and selecting studies. *Cochrane Handb. Syst. Rev. Interv.* **2019**, 67–107. . [CrossRef]
16. Fink, A. *Conducting Research Literature Reviews: From the Internet to Paper*; Sage Publications: Newcastle upon Tyne, UK, 2019.
17. Simsek, Z.; Fox, B.; Heavey, C. Systematicity in organizational research literature reviews: A framework and assessment. *Organ. Res. Methods* **2023**, *26*, 292–321. [CrossRef]
18. Parajuli, R.; Markwith, S.H. Quantity is foremost but quality matters: A global meta-analysis of correlations of dead wood volume and biodiversity in forest ecosystems. *Biol. Conserv.* **2023**, *283*, 110100. [CrossRef]
19. Akresh, M.E.; King, D.I.; McInvale, S.L.; Larkin, J.L.; D'Amato, A.W. Effects of forest management on the conservation of bird communities in eastern North America: A meta-analysis. *Ecosphere* **2023**, *14*, e4315. [CrossRef]

20. Liu, W.; Zhang, Z.; Li, J.; Wen, Y.; Liu, F.; Zhang, W.; Liu, H.; Ren, C.; Han, X. Effects of fire on the soil microbial metabolic quotient: A global meta-analysis. *Catena* **2023**, *224*, 106957. [CrossRef]
21. Stogiannis, D.; Siannis, F.; Androulakis, E. Heterogeneity in meta-analysis: A comprehensive overview. *Int. J. Biostat.* **2024**, *20*, 169–199. [CrossRef]
22. Coverdale, T.C.; Davies, A.B. Unravelling the relationship between plant diversity and vegetation structural complexity: A review and theoretical framework. *J. Ecol.* **2023**, *111*, 1378–1395. [CrossRef]
23. Urbano, F.; Viterbi, R.; Pedrotti, L.; Vettorazzo, E.; Movalli, C.; Corlatti, L. Enhancing biodiversity conservation and monitoring in protected areas through efficient data management. *Environ. Monit. Assess.* **2024**, *196*, 12. [CrossRef]
24. Zhu, J.J.; Yang, M.; Ren, Z.J. Machine learning in environmental research: Common pitfalls and best practices. *Environ. Sci. Technol.* **2023**, *57*, 17671–17689. [CrossRef]
25. Graham, E.B.; Averill, C.; Bond-Lamberty, B.; Knelman, J.E.; Krause, S.; Peralta, A.L.; Shade, A.; Smith, A.P.; Cheng, S.J.; Fanin, N.; et al. Toward a generalizable framework of disturbance ecology through crowdsourced science. *Front. Ecol. Evol.* **2021**, *9*, 588940. [CrossRef]
26. Meng, Y.; Liu, X.; Ding, C.; Xu, B.; Zhou, G.; Zhu, L. Analysis of ecological resilience to evaluate the inherent maintenance capacity of a forest ecosystem using a dense Landsat time series. *Ecol. Inform.* **2020**, *57*, 101064. [CrossRef]
27. Marshall, I.J.; Kuiper, J.; Wallace, B.C. RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 193–201. [CrossRef] [PubMed]
28. O'Mara-Eves, A.; Thomas, J.; McNaught, J.; Miwa, M.; Ananiadou, S. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Syst. Rev.* **2015**, *4*, 5. [CrossRef] [PubMed]
29. James, K.L.; Randall, N.P.; Haddaway, N.R. A methodology for systematic mapping in environmental sciences. *Environ. Evid.* **2016**, *5*, 7. [CrossRef]
30. Roy, D.; Alison, J.; August, T.; Bélisle, M.; Bjerge, K.; Bowden, J.; Bunsen, M.; Cunha, F.; Geissmann, Q.; Goldmann, K.; et al. Towards a standardized framework for AI-assisted, image-based monitoring of nocturnal insects. *Philos. Trans. R. Soc. B* **2024**, *379*, 20230108. [CrossRef] [PubMed]
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
32. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
33. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv* **2022**, arXiv:2208.10442.
34. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. Coca: Contrastive captioners are image–text foundation models. *arXiv* **2022**, arXiv:2205.01917.
35. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
36. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
37. Suadaa, L.H.; Kamigaito, H.; Funakoshi, K.; Okumura, M.; Takamura, H. Towards table-to-text generation with numerical reasoning. In Proceedings of the 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), Virtual, 1–6 August 2021; pp. 1451–1465.
38. Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The iNaturalist Species Classification and Detection Dataset-Supplementary Material. *Reptilia* **2018**, *32*, 1–3.
39. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image–text models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25278–25294.
40. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive learning of medical visual representations from paired images and text. *arXiv* **2020**, arXiv:2010.00747.
41. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
42. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2818–2829.
43. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
44. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
45. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

46. Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.W.; Zhu, S.C.; Tafjord, O.; Clark, P.; Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2507–2521.
47. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv* **2022**, arXiv:2205.12005.
48. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
49. Diao, Q.; Jiang, Y.; Wen, B.; Sun, J.; Yuan, Z. Metaformer: A unified meta framework for fine-grained recognition. *arXiv* **2022**, arXiv:2203.02751.
50. Touvron, H.; Vedaldi, A.; Douze, M.; Jegou, H. Fixing the train-test resolution discrepancy. In *Proceedings of the Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc. 2019. Available online: https://arxiv.org/abs/1906.06423 (accessed on 18 August 2024).
51. Song, Y.; Sebe, N.; Wang, W. On the eigenvalues of global covariance pooling for fine-grained visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3554–3566. [CrossRef]
52. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. Transfg: A transformer architecture for fine-grained recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Online, 22 February–1 March 2022; pp. 852–860.
53. Du, X.; Lin, T.Y.; Jin, P.; Ghiasi, G.; Tan, M.; Cui, Y.; Le, Q.V.; Song, X. SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
54. Li, S.; Gong, K.; Liu, C.H.; Wang, Y.; Qiao, F.; Cheng, X. MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5212–5221.
55. Juan, D.C.; Lu, C.T.; Li, Z.; Peng, F.; Timofeev, A.; Chen, Y.T.; Gao, Y.; Duerig, T.; Tomkins, A.; Ravi, S. Graph-rise: Graph-regularized image semantic embedding. *arXiv* **2019**, arXiv:1902.10814.
56. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–137.
57. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18166–18176.
58. Cheng, M.; Sun, Y.; Wang, L.; Zhu, X.; Yao, K.; Chen, J.; Song, G.; Han, J.; Liu, J.; Ding, E.; et al. ViSTA: Vision and scene text aggregation for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5184–5193.
59. Messina, N.; Stefanini, M.; Cornia, M.; Baraldi, L.; Falchi, F.; Amato, G.; Cucchiara, R. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. In Proceedings of the 19th International Conference on Content-based Multimedia Indexing, Graz, Austria, 14–16 September 2022; pp. 64–70.
60. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
61. Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv* **2023**, arXiv:2303.16199.
62. Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv* **2023**, arXiv:2302.00923.
63. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
65. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
66. Pianykh, O.S.; Langs, G.; Dewey, M.; Enzmann, D.R.; Herold, C.J.; Schoenberg, S.O.; Brink, J.A. Continuous learning AI in radiology: Implementation principles and early applications. *Radiology* **2020**, *297*, 6–14. [CrossRef] [PubMed]
67. Hadsell, R.; Rao, D.; Rusu, A.A.; Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends Cogn. Sci.* **2020**, *24*, 1028–1040. [CrossRef] [PubMed]
68. Chen, J.; Zhang, A. Hetmaml: Task-heterogeneous model-agnostic meta-learning for few-shot learning across modalities. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event, Australia, 1–5 November 2021; pp. 191–200.
69. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**, *28*, 11–21. [CrossRef]
70. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T.L.; Raja, A.; et al. Multitask prompted training enables zero-shot task generalization. *arXiv* **2021**, arXiv:2110.08207.

71. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. *spaCy: Industrial-Strength Natural Language Processing in Python*; Zenodo: Honolulu, HI, USA, 2020.
72. Yan, J.; Xiao, Y.; Mukherjee, S.; Lin, B.Y.; Jia, R.; Ren, X. On the Robustness of Reading Comprehension Models to Entity Renaming. *arXiv* **2021**, arXiv:2110.08555.

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.