electronics

MDPI

Article

Towards Cognition-Aligned Visual Language Models via Zero-Shot Instance Retrieval

Teng Ma [1], Daniel Organisciak [2], Wenbao Ma [1,*] and Yang Long [3]

[1] School of Humanities and Social Science, Xi'an Jiaotong University, Xi'an 710049, China; matengsax@gmail.com
[2] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK; d.organisciak@gmail.com
[3] Department of Computer Sciences, Durham University, Durham DH1 3LE, UK; yang.long@durham.ac.uk
* Correspondence: smiling-ma@163.com

Abstract: The pursuit of Artificial Intelligence (AI) that emulates human cognitive processes is a cornerstone of ethical AI development, ensuring that emerging technologies can seamlessly integrate into societal frameworks requiring nuanced understanding and decision-making. Zero-Shot Instance Retrieval (ZSIR) stands at the forefront of this endeavour, potentially providing a robust platform for AI systems, particularly large visual language models, to demonstrate and refine cognition-aligned learning without the need for direct experience. In this paper, we critically evaluate current cognition alignment methodologies within traditional zero-shot learning paradigms using visual attributes and word embedding generated by large AI models. We propose a unified similarity function that quantifies the cognitive alignment level, bridging the gap between AI processes and human-like understanding. Through extensive experimentation, our findings illustrate that this similarity function can effectively mirror the visual–semantic gap, steering the model towards enhanced performance in Zero-Shot Instance Retrieval. Our models achieve state-of-the-art performance on both the SUN (92.8% and 82.2%) and CUB datasets (59.92% and 48.82%) for bi-directional image-attribute retrieval accuracy. This work not only benchmarks the cognition alignment of AI but also sets a new precedent for the development of visual language models attuned to the complexities of human cognition.

Keywords: large visual language models; zero-shot instance retrieval; cognition alignment

check for updates

Citation: Ma, T.; Organisciak, D.; Ma, W.; Long, Y. Towards Cognition-Aligned Visual Language Models via Zero-Shot Instance Retrieval. *Electronics* **2024**, *13*, 1660. https://doi.org/10.3390/electronics13091660

Academic Editor: Arkaitz Zubiaga

Received: 9 March 2024
Revised: 16 April 2024
Accepted: 16 April 2024
Published: 25 April 2024

1. Introduction

The advent of large-scale Artificial Intelligence (AI) models has marked a transformative era in computational learning, with their unprecedented capacity for data processing and pattern recognition shaping the trajectory of technological advancement. As these behemoths of AI continue to burgeon, their integration into diverse societal sectors underscores a critical need. Cognition alignment ensures that AI models not only perform tasks efficiently but also reflect the intricacies of human thought processes. Cognition-aligned models promise to deliver more intuitive interactions, enhance decision-making compatibility, and foster trust, as their operational logic mirrors the cognitive frameworks humans use to understand, reason, and contextualise. In essence, aligning AI with human cognition is not merely a technical aspiration but the foundation for the harmonious coexistence of AI systems and their human counterparts in an increasingly automated world.

The notion of cognition alignment in AI is deeply rooted in the rich soil of cognitive psychology and constructivist theory. Cognitive psychology, a discipline that develops from understanding mental processes, posits that human cognition is a complex interplay of various mental activities. This field has long been fascinated with how people perceive, remember, think, speak, and solve problems. Constructivist theory complements this by suggesting that learners construct knowledge through an active learning process rather than absorbing information passively. It emphasises the learner's critical role in making sense of new information by linking it to prior knowledge and experiences stored in memory.

Cognition alignment theory extrapolates these principles to the realm of AI. It advocates for the design of AI systems that 'learn' as much as humans do—by connecting new data to pre-existing knowledge frameworks and by abstracting underlying principles through reflection. The theory underscores the importance of AI systems being able to not only recall information but also apply it to novel situations, predict future scenarios, and adaptively learn from experiences. This approach ensures that AI can engage with tasks in a way that is reminiscent of human problem-solving and decision-making processes, with the flexibility and creativity that are hallmarks of human cognition.

In essence, cognition alignment theory in AI calls for the development of intelligent systems that go beyond pattern recognition and data analysis. It seeks to create AI that can understand the context of data, draw upon a wealth of experiences (real or simulated), make inferences, and anticipate future needs or actions—much like a human would when faced with new and complex challenges. It is a pursuit to bridge the gap between human and machine learning processes to create AI that not only computes but comprehends.

Zero-shot learning (ZSL) refers to a classification problem where the learning algorithm must correctly classify objects or data points that it has not seen during training. It is a machine learning technique where the model is expected to infer information about unseen classes by only learning about seen classes, usually through some form of transfer learning or by exploiting commonalities between classes. In ZSL, the model typically uses attributes or descriptions of objects to make these inferences. For example, if a model is trained on a dataset of animal images that includes various seen classes like 'tiger', 'elephant', and 'horse', it might later be tested on its ability to recognise an 'antelope', which it has never seen before, by using learnt attributes such as 'four legs', 'hooves', and 'horns'.

ZSL emerges as an ideal testbed for the study of cognition alignment precisely because it encompasses many of the challenges and intricacies of replicating human cognitive processes in AI systems. The paradigm of ZSL fundamentally relies on the ability of neural networks to engage in visual perception, not by merely recognising patterns through brute computational power but by understanding and extrapolating concepts in the absence of explicit prior examples.

This learning paradigm calls upon advanced knowledge representation techniques that are vital to human cognition, such as the identification of visual attributes, the interpretation of free text, and the application of knowledge encapsulated in ontological taxonomies. Furthermore, zero-shot learning utilises similes and exemplars that are inherently tied to the way humans draw analogies and learn from abstract examples. These methodologies are cornerstones in the extraction and emulation of human cognitive strategies, allowing AI to go beyond simple task execution to demonstrate an understanding of context.

As shown in Figure 1, in traditional ZSL research, ZSL models are designed to match images to human cognition. Human cognition is represented by free text or attributes that can guide the ZSL model to recognise images from unseen classes. An LLM explores human cognition from online information. We ask the LLM to create the same representation as human cognition, such as free-text descriptions or visual attributes. We can then use ZSL to measure whether the LLM's cognition can align well with human cognition so that the ZSL model still recognises the same images from unseen classes. The cognition alignment between an LLM and a human is then measured by the final ZSL recognition rate. The neural network is tasked with the challenge of transferring its learnt knowledge to entirely new classes, samples, tasks, or domains that it never encountered during training. The success of this transfer hinges on the model's alignment with human cognitive representations—its ability to generalise and apply abstract principles to new and unseen data. Hence, ZSL does not just assess an AI's learning efficiency; it evaluates the AI's cognitive congruence with human thought patterns. It is in this rigorous testing of generalisation capabilities that the true measure of cognition alignment is found, making zero-shot learning a prime candidate for advancing our understanding of how AI can not only mimic but also meaningfully engage with human cognitive processes.

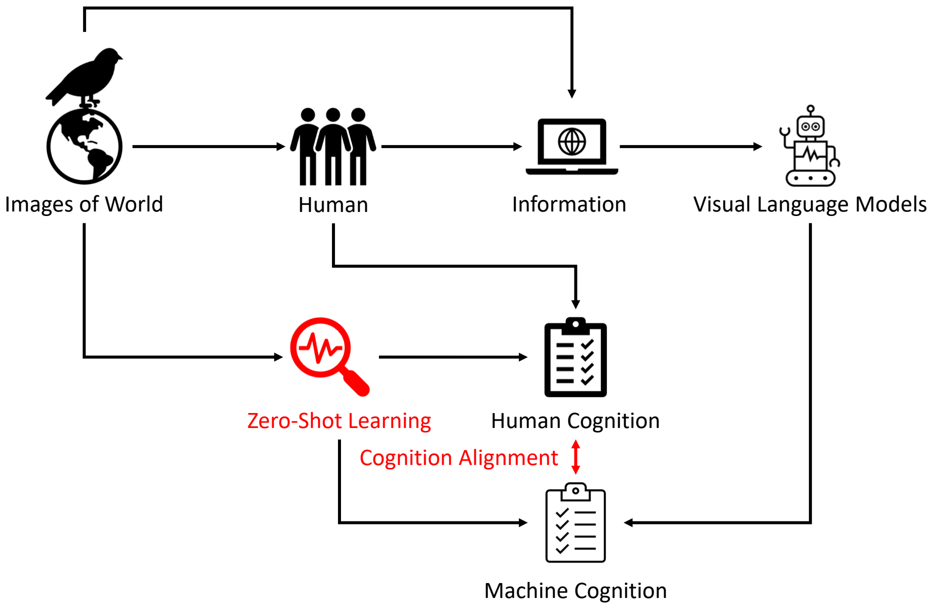

Figure 1. The main aim of this paper is to use visual perception to measure the alignment between human and AI cognition.

While traditional ZSL is usually concerned with classification tasks, Zero-Shot Instance Retrieval (ZSIR) [1] is about retrieving particular instances of data that match a given description or query without having seen examples of that specific category or instance during training. It is a more specific task where the model needs to understand and match a complex query to instances of an unseen category. The difference lies in the output and the nature of the task: ZSL is about classifying an instance into a category not seen during training, while ZSIR is about retrieving all relevant instances that match a zero-shot query, even when the model has not been trained with any examples from the category of the query. Essentially, ZSL is about 'what' an object is, and ZSIR is about finding 'where' or 'which' objects fulfil the criteria described in the query.

ZSIR requires the model to have a sophisticated understanding of attributes and their relationships, as it may need to retrieve specific instances based on descriptions that involve unseen combinations of attributes. This is a more complex task since the model must deal with a more nuanced space of attributes and must be able to rank instances in terms of their relevance to a query. Our contributions are summarised as follows:

- We introduce a novel framework that utilises Zero-Shot Instance Retrieval (ZSIR) as a method to study and analyse the cognitive alignment of large visual language models. This approach allows us to simulate and evaluate how AI interprets and processes visual information in a manner that parallels human cognitive abilities, particularly in scenarios where the model encounters data it has not been explicitly trained to recognise.
- A key innovation of our research is the development of a unified similarity function specifically designed to quantify the level of cognitive alignment in AI systems. This function provides a metric that correlates the AI's interpretations with human-like cognition, offering a quantifiable measure of the AI's ability to align its processing with human thought patterns.
- The effectiveness of our proposed similarity function was thoroughly tested through extensive experiments on the SUN and CUB datasets. Our results demonstrate that the function is versatile and robust across different forms of knowledge representation,

including visual attributes and free text generated by large AI models. This versatility is critical, as it reflects the level of cognition alignment between humans and AI.

- Our experiments not only establish the validity of the proposed similarity functions but also showcase the enhanced performance of our model in the context of ZSIR tasks. The model demonstrates superior capabilities compared to existing state-of-the-art models on both the SUN (92.8% and 82.2%) and CUB datasets (59.92% and 48.82%) for image-to-attribute and attribute-to-image retrieval accuracy.

2. Related Work

The recent surge in research on cognition-aligned Large Language Models (LLMs) reflects a growing interest in developing AI systems that can reason, understand, and interact in ways that align with human cognitive processes. This literature review provides insights from key papers in this domain and discusses related zero-shot learning (ZSL) techniques that are related to our work. The reviewed studies are summarised in Table 1.

Table 1. Summary of literature on cognition-aligned LLMs and zero-shot learning.

Authors	Year	Focus Area	Key Contributions
Xu et al. [2]	2023	Cognition alignment in LLMs	Proposing a risk taxonomy and policy framework for aligning LLMs with human preferences.
Wang et al. [3]	2023	Enhancing reasoning in LLMs	Introducing the AFT paradigm to improve the reasoning capabilities of LLMs.
Lester et al. [4]	2021	Cognition alignment in LLMs	Emphasising the alignment of textual neural representations with cognitive language processing signals.
Xu et al. [5]	2023	Cultural specificity in LLMs	Exploring the values of Chinese LLMs for cultural alignment.
Sengupta et al. [6]	2023	Linguistic alignment in LLMs	Developing Arabic-centric LLMs.
Zhang et al. [7]	2023	Cross-lingual alignment	Bridging cross-lingual alignment through interactive translation.
Wang et al. [8]	2023	Emotional intelligence in LLMs	Assessing emotional intelligence crucial for effective communication.
Bhardwaj et al. [9]	2023	Safety alignment in LLMs	Proposing red-teaming techniques for safety alignment.
Liu et al. [10]	2023	LLM alignment surveys	Discussing key dimensions crucial for assessing LLM trustworthiness.
Gu et al. [11]	2023	Application in ZSL	Focusing on zero-shot NL2SQL generation combining pre-trained language models with LLMs.
Kirk et al. [12]	2023	Personalisation in LLMs	Discussing the personalisation of LLMs within societal bounds.
Petroni et al. [13]	2019	Capability of LLM to recall knowledge	Investigating the LLM as an unstructured knowledge base.

2.1. Cognition-Aligned AI

Cognition alignment has become an emerging trend in the AI research community. The sharp-rising intelligence capacity of LLMs has caused both technical and social concerns. Xu et al. (2023) [2] addressed the challenges in aligning LLMs with human preferences, proposing a risk taxonomy and policy framework for personalised feedback. Their work underscores the complexity of aligning LLMs with diverse human values and preferences.

Another trend in cognition alignment studies focuses on enhancing reasoning in LLMs. Wang et al. (2023) [3] and Lester et al. (2021) [4] focused on improving the reasoning capabilities of LLMs. An Alignment Fine-Tuning (AFT) paradigm was introduced to address the Assessment Misalignment problem in LLMs, enhancing their reasoning abilities. Lester et al.'s work on CogAlign emphasised the alignment of textual neural representations with cognitive language processing signals, highlighting the importance of cognitive alignment in LLMs.

For broader concerns and applications in the linguistic and social domains, the papers by Xu et al. (2023) [5], Sengupta et al. (2023) [6], and Zhang et al. (2023) [7] explored the alignment of LLMs in specific cultural and linguistic contexts. Xu et al. focused on the values of Chinese LLMs. Sengupta et al. developed Arabic-centric LLMs, and Zhang et al. bridged cross-lingual alignment through interactive translation. Another research direction for the cognition alignment of AI focuses on the emotional and safety guarantees of AI models. Wang et al. (2023) [8] and Bhardwaj et al. (2023) [9] explored the emotional intelligence of LLMs and their safety alignment. Wang et al. assessed LLMs' emotional intelligence, crucial for effective communication, while Bhardwaj et al. proposed red-teaming techniques for safety alignment.

Considering that cognition alignment in the contexts of LLMs and AI is still a new research topic that just started gaining interest at the beginning of 2023, there are only a few survey papers summarising the progression and challenges in this domain. Liu et al. (2023) [10] and Petroni et al. (2019) [13] provided comprehensive surveys on LLM alignment and their potential as knowledge bases, respectively. Liu et al. discussed key dimensions crucial for assessing LLM trustworthiness. In contrast, Petroni et al. explored the capability of LLMs to store and recall factual knowledge in 2019. By comparing the differences between and the progression of the two survey papers from 2019 and 2023, we found that one of the topics of cognition alignment related to our work aims to apply the aligned cognition representation to improve the performance of AI training and machine learning. Gu et al. (2023) [11] and Kirk et al. (2023) [12] presented application-specific advancements in LLMs. Gu et al. focused on zero-shot NL2SQL generation, combining pre-trained language models with LLMs, while Kirk et al. discussed the personalisation of LLMs within societal bounds.

As a short summary, this review highlights the diverse approaches and challenges in aligning LLMs with human cognition, values, and preferences. From enhancing reasoning capabilities to addressing cultural specificity and emotional intelligence, these studies collectively contribute to the development of more aligned, effective, and ethically sound LLMs. In line with our research focus in this paper, we also explore a new paradigm that can use well-aligned AI cognition to seamlessly improve the efficiency of human ontological engineering, i.e., brainstorming for conceptualisation; data collection; labelling and tagging of class embeddings, descriptions, and attributes; annotation via crowd-sourcing approaches; validation of the ontological structure via theoretical analysis and discussion; etc. In contrast to all of the existing work mentioned above, our unique contribution in this paper is the introduction of the ZSL task ZSIR as a quantitative measurement for the level of cognition alignment between AI and humans. This is considered to be a bilateral reciprocal benefit. For one thing, it is crucial to understand how well AI LLMs are aligned with human cognition so that the data annotation and interpretation work can be reliably handed over to the machine. Otherwise, the poisoned, biased cognition of the LLM can exaggerate the risk when it is applied to downstream supervised learning tasks. From another perspective, AI with well-aligned cognition can efficiently improve the model performance in downstream tasks. At such an early stage, our work aims to establish a healthy paradigm that both assesses the level of cognition alignment and applies the method to improve the downstream task, e.g., ZSL image recognition.

2.2. Zero-Shot Learning

Zero-shot learning (ZSL) has undergone significant advancements, marked by key contributions that have shaped its current state. The journey began with Larochelle et al. [14], who introduced the concept of zero-data learning, proposing a method to learn new tasks without training data, a foundational idea in ZSL. The first mention of ZSL was made by Palatucci et al. [15], who explored semantic output codes in ZSL, demonstrating how semantic information could be used to recognise classes unseen during training. Lampert et al. [16] then introduced an attribute-based approach to detecting unseen object classes, a seminal work that showed the effectiveness of shared attributes in identifying novel objects, which popularised ZSL paradigms in the computer vision domain.

The integration of deep learning with semantic embeddings was significantly advanced by Frome et al. [17] with the DeViSE model, which combined visual and textual information for ZSL. Norouzi et al. [18] further contributed by combining multiple semantic embeddings, improving the accuracy of ZSL models. The concept of synthesised classifiers, which was crucial for generalising from seen to unseen classes, was introduced by Changpinyo et al. [19].

A comprehensive evaluation of various ZSL approaches was presented by Xian et al. [20], establishing a benchmark for future research in the field. This was crucial for understanding the strengths and weaknesses of different ZSL methodologies. Liu et al. [21] addressed the problem of generalised zero-shot learning, where the test set contains both seen and unseen classes, proposing a deep calibration network to balance the learning between these classes.

Wang et al. [22] conducted a detailed survey that provided an extensive overview of the methodologies, datasets, and challenges in ZSL, offering insights into the state of the field. This survey was instrumental in summarising the progress and directing future research efforts. In recent years, the focus of ZSL has shifted towards more complex and realistic scenarios. This includes the integration of unsupervised and semi-supervised techniques, the use of generative models to synthesise features for unseen classes, and the exploration of cross-modal ZSL. These advancements aim at improving the scalability, robustness, and practical applicability of ZSL models.

Zero-shot learning has been applied to various downstream tasks, each marked by key milestone papers that have significantly advanced the field. Khandelwal et al. [23] proposed a simple yet effective method for zero-shot detection and segmentation, outperforming more complex architectures. This work was pivotal in demonstrating the effectiveness of straightforward approaches in ZSL for object detection and segmentation. Chen et al. [24] described a vision-based method for analysing excavators' productivity using zero-shot learning. This method identifies activities of construction machines without pre-training, showcasing the practical application of ZSL in real-world scenarios. Díaz et al. [25] presented a novel zero-shot prototype recurrent convolutional network for human activity recognition via channel state information. This method enhances cross-domain transferability, a crucial aspect of ZSL in activity recognition. Nag et al. [26] designed a transformer-based framework, TranZAD, for zero-shot temporal activity detection. This framework streamlines the detection of unseen activities, demonstrating the potential of transformers in ZSL.

Zero-Shot Instance Retrieval via Dominant Attributes [1] is a methodology that reflects the core strengths and challenges in the development of cognitively aligned LLMs. It is a novel approach to semantic searching in the context of zero-shot learning. This paradigm is particularly relevant for measuring the cognitive alignment of Large Language Models (LLMs). Firstly, the paper's focus on semantic searching aligns well with the cognitive capabilities of LLMs, which rely on understanding and processing semantic information. Secondly, the use of dominant attributes in zero-shot retrieval mirrors the way that LLMs leverage contextual cues and attributes to generate responses, making it a fitting method to evaluate their cognitive alignment. Thirdly, the approach emphasises affordability, which is crucial in making advanced semantic searching techniques more accessible, a goal that aligns with the democratisation efforts in AI and LLMs. In addition, the zero-shot aspect of the retrieval process is akin to the generalisation capabilities of LLMs, making it an ideal testbed to assess how well these models can adapt to new, unseen data while maintaining semantic coherence. In contrast to ZSIR model research, this paper focuses on introducing ZSIR to measure cognition alignment via the visual perception task of image retrieval. Both synthetic attributes and class descriptions are explored and compared with the human cognition representation. To the best of our knowledge, this is the first-ever paradigm that can quantitatively measure cognition alignment between AI and humans via visual perception tasks and ontological engineering.

3. Methodology

Our key contributions are reflected and illustrated in the framework shown in Figure 2. First, our work introduces LLMs to autonomous ontological engineering, which improves the efficiency of human labour by over 1500 times. Second, our ZSIR model, with the selected base model and similarity functions, can measure cognition alignment between AI and humans via ontological engineering. In addition, we found that hybrid AI–human cognition can be reflected by improved ontological engineering and, in turn, can improve the ZSIR model performance. In traditional zero-shot learning (ZSL), $\mathcal{X}$ denotes the visual space, where each instance $x \in \mathcal{X}$ represents a visual instance. Correspondingly, $\mathcal{Y}$ represents the label space for seen classes, and $\mathcal{Y}^u$ represents the label space for unseen classes, where $\mathcal{Y} \cap \mathcal{Y}^u = \emptyset$. The objective of ZSL is to learn a mapping function $f : \mathcal{X} \to \mathcal{Y}^u$ that can accurately associate unseen visual instances with their corresponding category label. Because the distribution between $\mathcal{Y}$ and $\mathcal{Y}^u$ is disjoint, the association between the two domains is required. $\Phi(.)$ and $\Psi(.)$ denote the perception and cognition functions needed to process visual features in $\mathcal{X}$ and labels in $\mathcal{Y}$, respectively. As collecting category attributes requires considerable human cognitive labour, it is infeasible to collect instance-level annotations for large datasets. In other words, existing ZSL can only map an unseen instance to a category, while ZSIR requires retrieving a specific instance of the category. Our approach adopts both human- and AI-generated attribute representations $\mathcal{A} = \Psi(\mathcal{Y})$ so that each attribute dimension $a \in \mathcal{A}$ corresponds to a data-driven feature that captures explicit and latent attributes pertinent to cognitive alignment.

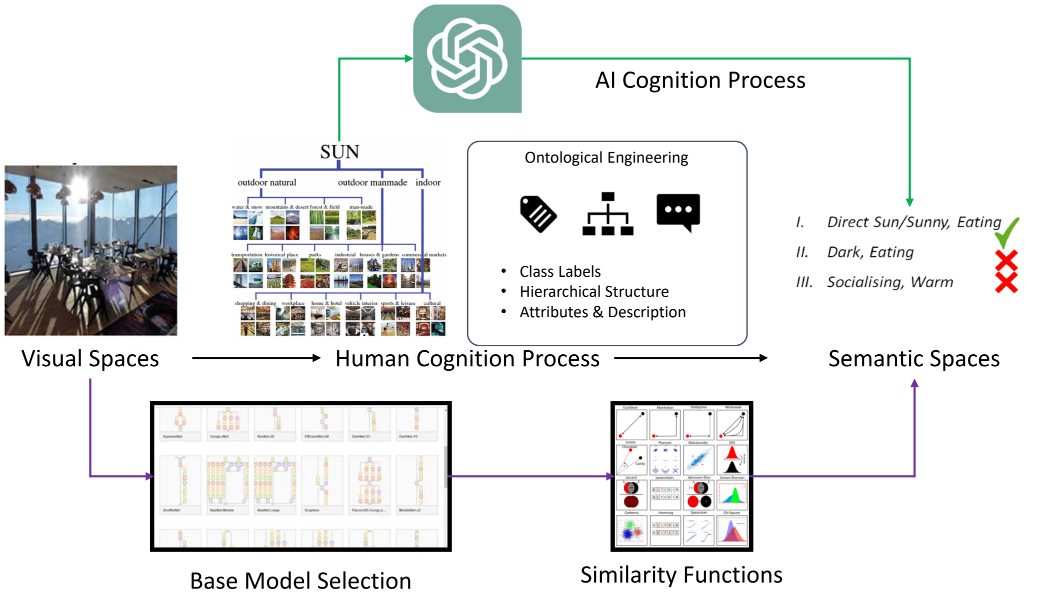

Figure 2. Our key contributions are reflected and illustrated in the framework.

To measure human–AI cognitive alignment, we introduce a cognitive alignment function $\mathcal{L} : \Phi(\mathcal{X}) \times \Psi(\mathcal{Y}) \to \mathbb{R}$, which measures the degree of alignment between the AI's data-driven representation of a visual instance and human cognitive processes. The function $\mathcal{L}$ assesses how closely the AI's output for an unseen instance x aligns with human-like cognition, e.g., visual attributes, free texts, and knowledge graphs, when classifying it into an unseen class y^u.

The methodology concentrates on optimising the cognitive alignment function $\mathcal{L}$ by adjusting the mapping from both visual and attribute spaces to the latent instance attribute, ensuring that AI's interpretation of visual data not only aligns with human cognition in recognising unseen classes but also adheres to the cognitive processes that humans employ in categorising and understanding visual stimuli.

3.1. Cognition Representation

In addressing the challenge of cognitive alignment using the realm of ZSL, our approach begins with the fundamental premise that visual stimuli serve as a common informational foundation. We operate under the assumption that both AI systems and humans perceive the same visual information, yet their methods of processing and interpreting this information lead to divergent cognitive representations. Traditional human cognition in ZSL research encompasses a variety of forms, including visual attributes, free-text descriptions, and ontological taxonomies such as similes. A critical challenge in this context is the unification of these diverse cognitive representations into a coherent framework that AI can understand and utilise while minimising the need for extensive manual labour typically required for such tasks.

To bridge this gap, we propose the integration of Large Language Models (LLMs) like ChatGPT to facilitate the automated generation of labels and annotations $\mathcal{A}$. These models can provide a scalable and efficient means of translating the rich paradigms of human cognitive representations into a format that AI systems can process. In this paper, we consider the two most frequently used paradigms as a comparison:

- Automated Attribute Generation: LLMs can be used to automatically generate descriptive attributes for visual data. Similar to human attributes [16], this paradigm provides a structured and detailed attribute set that mirrors human perception.
- Free-Text Description Synthesis: LLMs can be employed to create comprehensive free-text descriptions of visual stimuli represented by word embeddings [1]. These narratives offer a deeper, more contextual understanding of the images, akin to how humans might describe them.

Through these methods, we aim to significantly reduce the manual labour involved in the annotation process while ensuring that the AI system's understanding of visual information aligns closely with human cognitive processes. This approach not only enhances the cognitive alignment of AI models but also paves the way for more intuitive and human-like AI interactions and interpretations in the field of ZSL.

For example, we estimate the total hours that humans might take to build up the "SUN attribute database". The process can be broken down into three main stages: (1) Developing a Taxonomy of 102 Discriminative Attributes: This initial stage involves crowd-sourced human studies. The complexity here depends on the methodology (e.g., surveys, workshops) and the level of agreement required to finalise the list. For estimation, we can assume that this stage requires several rounds of surveys and analysis. An initial setup, literature review, and preparation phase are assumed: 40 h (5 work days). Each round of survey and analysis is expected to take 20 h. At least 3 rounds are assumed for a robust taxonomy: $3 \times 20 = 60$ h. (2) Building the SUN Attribute Database: This involves annotating over 700 categories and 14,000 images. Each image needs to be reviewed and annotated with relevant attributes of the established taxonomy. The time to annotate one image can vary significantly, but we can just assume an average of 2 min per image. The total time for one annotator is estimated to be 14,000 × 2 min = 28,000 min. (3) Annotation by Three Human Annotators: The total effort will be multiplied by three, as each image is annotated by three different people to ensure accuracy and consistency. The total annotation time is $467 \times 3 = 1401$ h. So, the estimated total time would be approximately 100 + 1401 = 1501 h. This is a rough estimation, and the actual time may vary based on the efficiency of the process, the complexity of the images, and the proficiency of the annotators.

While an LLM can significantly reduce the time to build up class associations, hierarchies, and attribute annotations, we are curious whether the cognition of LLM can align well with the time-consuming ontological engineering processed by human annotators. In this paper, we consider the following paradigms, which have been widely adopted in previous ZSL research:

- Class Embedding: An automatic description is provided by AI for each given class.
- AI-Revised Human Attributes: By incorporating the class names and human-designed attributes, AI revises the attribute list and makes the words more related to visual perception for the image retrieval task.
- AI-Generated Attributes: AI creates attributes that are associated with the class names without any constraints.
- ZSL-Contextualised AI Attributes: Based on the AI-generated attributes, the prompting further constrains the task for ZSL purposes to focus on improving the visual perception association and generalisation for unseen classes and instances.

AI-revised human attributes aim to demonstrate whether the AI model can enhance human-designed attributes using its own cognition. The new list combines specific elements that are more directly applicable to individual scenes, such as 'Traffic Intensity' or 'Flora Types', which were aspects highlighted in the human-generated list. While maintaining specificity, these attributes are still broad enough to apply across various scenes, unlike some of the very niche attributes in the human-generated list. The attributes balance physical characteristics (e.g., 'Rock Formations', 'Weather Elements'), emotional or atmospheric qualities (e.g., 'Emotional Atmosphere', 'Safety Perception'), and functional aspects (e.g., 'Commercial Features', 'Conservation Efforts'). The list includes attributes related to human experiences and activities, reflecting the way that people interact with and perceive different environments. Attributes related to sensory experiences (e.g., 'Aroma Characteristics', 'Acoustic Qualities') are included, emphasising the multisensory nature of human scene perception. By combining specificity, broad applicability, and a balance of different types of descriptors, this AI-revised human attribute list aims to offer a more comprehensive and nuanced framework for scene classification than the human-generated list. It acknowledges the complexity of scenes and the multifaceted ways in which they can be understood.

While the first baseline is more constrained by human cognition inputs, e.g., class descriptions and human-designed attributes, the two other baselines provide more freedom for AI to incorporate its own cognition to create task-specific attributes. AI-generated attributes create a free attribute list using only given class names. This baseline can best reflect the true AI cognition based on concept-level associations. However, our cognition alignment approach is based on the assumption that both AI and humans aim to describe the same visual perception. The validation of cognition alignment is based on whether the multi-sourced cognition representation by human and AI can lead to accurate image retrieval in the ZSIR task. Therefore, in the final proposed paradigm, the prompting information constrains the AI to create more visual-specific attributes, and the list should be applied to both seen and unseen classes to test the generalisation of the association.

The final proposed ZSL-contextualised AI attributes focus on improving the generalisation, visual discriminators, balance between abstract and concrete levels, relevance to scene understanding, and compatibility with downstream visual–semantic modelling. These attributes are broad enough to be applicable across a wide range of scenes, which is essential for zero-shot learning, where the model needs to generalise from seen to unseen classes. The attributes are chosen for their potential to be visually discriminative. They capture key aspects of scenes that can distinguish one class from another. The list balances abstract qualities (like 'tranquil' or 'bustling') with concrete, visually identifiable features (like 'wooden' or 'mountainous'). This mix is crucial for a model that needs to understand and categorise scenes it has not been explicitly trained on. Attributes are relevant to understanding and describing scenes, which is the primary goal of the classifier. They cover a range of aspects, including material properties, environmental characteristics, and human-made vs. natural elements. These attributes are conducive to the creation of visual–semantic models, as they can be easily linked to visual data and semantic descriptions, forming a bridge between the visual appearance of a scene and the language-based descriptors. This revised list is designed to optimise the effectiveness of zero-shot learning models in classifying images by focusing on attributes that are both descriptive and discriminative, enhancing the model's ability to make accurate predictions on unseen data.

3.2. Latent Instance Attributes Discovery

As shown in Figure 2, for both ZSL and ZSIR, it is essential to establish class associations so that the training set of seen class samples can be generalised to the unseen domain. Visual spaces need to be projected into semantic spaces created by either human annotators or AI models, as mentioned above. Class embedding is the baseline approach that has been widely used in text-based ZSL methods. The generated class descriptions are then encoded by traditional word embedding, the intermediate BERT model, and GPT3 LLMs.

Once class embedding is achieved, ZSL models extract visual features using base models and learn to project them into semantic spaces via similarity functions. As shown in Figure 3, ZSIR is different from ZSL since the task requires differentiating instances in the same class, while ZSL aims to reduce the intra-class distances and enhance the inter-class distances. During the training stages, each class (of ABC) has a class-level attribute provided by either humans or AI (ChatGPT). LIAD aims to discover instance-specific attributes with both an orthogonal constraint and prototype grouping. Using the trained P1 and P2 networks, a cognition alignment score can be obtained during the test phase. ZSIR generalisation ensures that the test is applied to new classes from an unseen distribution so that the overall retrieval performance can better reflect the cognition alignment. Therefore, the semantic representation of attributes or word embeddings in ZSIR needs to reflect the detailed differences between instances in the same class. Although the SUN and CUB datasets provide both class-level and instance-level attributes, it is a very challenging task for AI-generated approaches. We follow the paradigm of Latent Instance Attributes Discovery [1], which is formalised as follows:

$$\min_{P_1,P_2} \mathcal{L}_1(\mathcal{X}P_1 - \mathcal{V}) + \mathcal{L}_2(\mathcal{A}P_2 - \mathcal{V}),\ s.t.\ \mathcal{V}^\top \mathcal{V} = I, \tag{1}$$

where $\mathcal{L}_1(\mathcal{X}P_1 - \mathcal{V})$ and $\mathcal{L}_2(\mathcal{A}P_2 - \mathcal{V})$ are the loss functions to learn a mapping from the visual space and attribute space to a shared latent space to discover the instance-level visual–semantic attributes. The latent space is constrained by an orthogonal projection so that the discovered attributes in $\mathcal{V}$ are uncorrelated. Each dimension of the discovered attributes can be considered an independent visual–semantic vocabulary formally written as follows, which ensures that each latent attribute dimension $\mathbf{v}^i, \mathbf{v}^j \in [\mathbf{v}^1, ..., \mathbf{v}^k]$ is compact and not redundant. k is a hyper-parameter that controls the dimension of the latent space:

$$< \mathbf{v}^i, \mathbf{v}^j >= \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Note that the cardinality $|\mathcal{X}| = N$ equals the sample size of the images, but the cardinality $|\mathcal{A}|$ is the number of categories. Therefore, we would expect a reduced rank from the visual space to the latent space and an increased rank from the attribute space to the latent space. In other words, for each attribute provided by either humans or LLMs, there are richer image examples to support the concept. In this paper, we introduce a prototype grouping (PG) method to (1) encourage more diverse prototypes of each visual–semantic attribute to be learnt and (2) encourage inter-class association so that the ZSIR generalisation to an unseen domain can be achieved. Firstly, to discover the intrinsic relationship between training samples $\mathbf{x}_i, \mathbf{x}_j \in [\mathbf{x}_1, ..., \mathbf{x}_N]$, we construct a graphical adjacency matrix $\mathcal{S} \in \mathbb{R}^{N \times N}$ for $\mathcal{X}$:

$$\mathcal{S}_{ij} = \max(0, \delta(\mathbf{x}_i^\top \mathbf{x}_j) - \epsilon), \tag{3}$$

where δ is a non-linear mapping function that can keep high-similarity responses while eliminating low-similarity responses to ensure that connected neighbours have strong intrinsic associations. ϵ is a threshold hyper-parameter that ensures that the neighbourhood connection in $\mathcal{S}$ is stronger so that the learnt prototypes can eliminate outliers. The property of δ is shown in Figure 4.

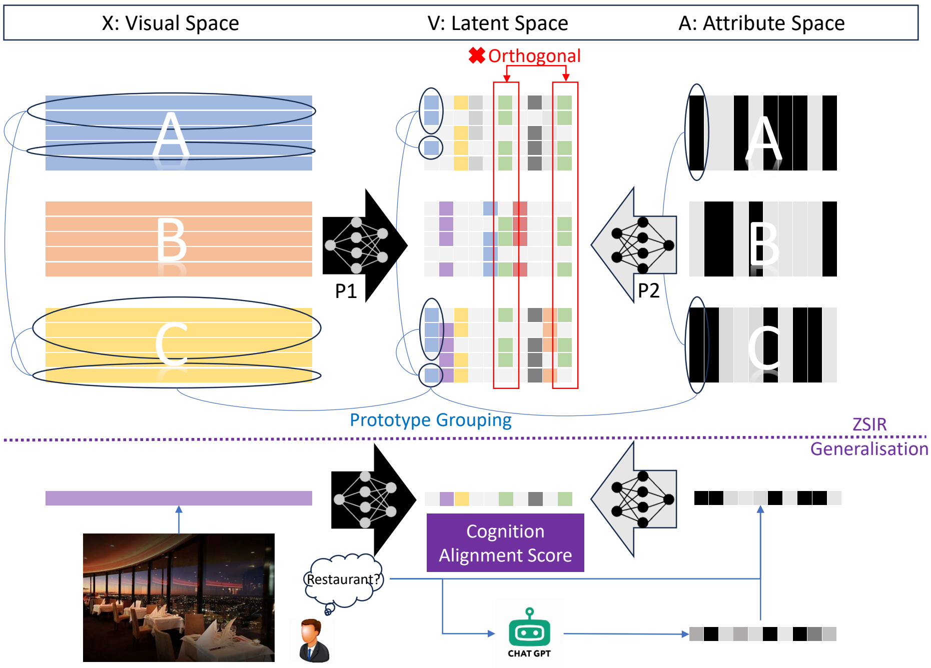

Figure 3. Cognition alignment involves training and test stages. Orthogonal Projection (in red) reduces redundant dimensions in the latent space. Prototype grouping (blue circles) selects only representative samples rather than putting the whole classes A and C to learn the attributes.

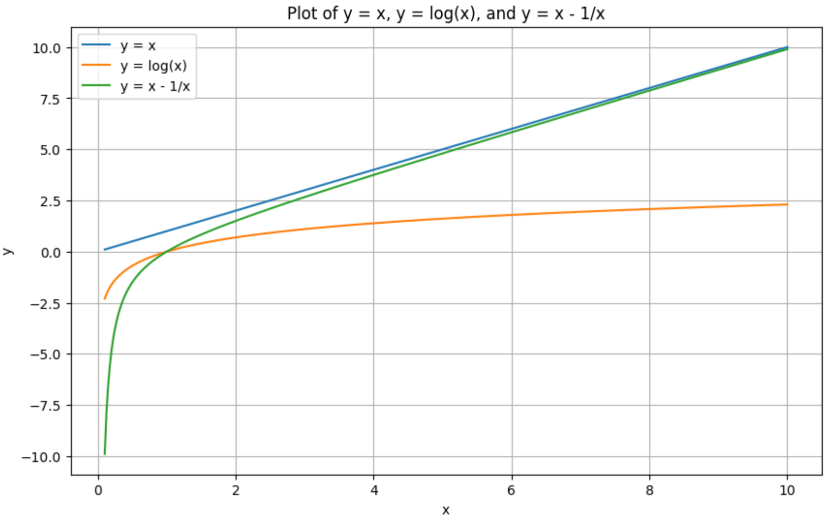

Figure 4. Property of non-linear mapping function δ compared with other functions.

Similar to the visual space, we apply the same adjacency matrix approach to the attribute space. As illustrated in Figure 3, $\mathcal{A}$ is a low-rank matrix since the rank equals the number of classes C, which is much smaller than the sample size N. As a result, the adjacency matrix will have a block of connections for the same class samples, and the inter-class associations are also reflected at the category level. Finally, Latent Instance Attributes Discovery consists of two mapping functions as follows:

$$\begin{cases} \mathcal{V}_X = \mathrm{softmax}(\hat{\mathcal{D}}_X^{-\frac{1}{2}}\hat{\mathcal{S}}_X\hat{\mathcal{D}}_X^{-\frac{1}{2}}\mathcal{X}P_1) \\ \mathcal{V}_A = \mathrm{softmax}(\hat{\mathcal{D}}_A^{-\frac{1}{2}}\hat{\mathcal{S}}_A\hat{\mathcal{D}}_A^{-\frac{1}{2}}\mathcal{A}P_2) \end{cases}, \tag{4}$$

where, for both domains $\mathcal{X}$ and $\mathcal{A}$, $\hat{\mathcal{S}} = \mathcal{S} + I$ is the enhanced adjacency matrix by the identity matrix I, and $\hat{\mathcal{D}}_{ii} = \sum_j \hat{\mathcal{S}}_{ij}$ is the degree matrix with values on the matrix diagonal and zeros elsewhere. $\hat{\mathcal{D}}^{-\frac{1}{2}}\hat{\mathcal{S}}\hat{\mathcal{D}}^{-\frac{1}{2}}$ is the normalised adjacency matrix so that the graphical condition can be applied to the projections from visual and attribute spaces $P_1 \in \mathbb{R}^{d_x \times k}$ and $P_2 \in \mathbb{R}^{d_a \times k}$ to provide the prototype grouping condition. d_x and d_a are the dimensions of raw visual and attribute spaces $\mathcal{X}$ and $\mathcal{A}$, and k is the dimension of the shared latent visual attribute space $\mathcal{V}$.

3.3. Cognition Alignment via ZSIR

Using both the orthogonality and PG constraints, we can project the visual perception $\mathcal{X}$ and semantic cognition $\mathcal{A}$ into the shared latent space to achieve cognition alignment (CA). The equation serves to ensure our essential promise that different cognition representations in the attributes are aligned with the same visual stimuli in $\mathcal{X}$.

$$\mathcal{L}_{CA} = \frac{1}{2N}\|\mathcal{V}_X - \mathcal{V}_A\|_F^2. \tag{5}$$

Optimisation Strategy: Solving the above Equation (5) is a dynamic NP-hard problem because either visual or attribute projection to the latent space is unknown. In this paper, we propose an alternating optimisation strategy, which is summarised in Algorithm 1.

Algorithm 1 LIAD optimisation for ZSIR cognition alignment

```
Input: Visual features of training images $\mathcal{X} = \Phi(\mathrm{imgs}^s)$; attributes of seen classes $\mathcal{A} = \Psi(\mathcal{Y})$;
    test images from unseen classes $\mathcal{X}^u = \Phi(\mathrm{imgs}^u)$ with the attributes $\mathcal{A}^u = \Psi(\mathcal{Y}^u)$.
Output: Gallery and query instances $\mathcal{V}_{X^u}$ and $\mathcal{V}_{A^u}$.
Initialise: $P_1$ and $P_2$;
While $\mathcal{L}_1$ and $\mathcal{L}_2$ not converge:
    $\mathcal{V} \leftarrow \mathcal{V}_X = \mathrm{softmax}(\hat{\mathcal{D}}_X^{-\frac{1}{2}}\hat{\mathcal{S}}_X\hat{\mathcal{D}}_X^{-\frac{1}{2}}\mathcal{X}P_1)$;
    for iter ∈ 0, 1, ..., MaxIter:
        $P_2 \leftarrow \min_{P_2} \mathcal{L}_2 := \frac{1}{2N}\|\mathcal{V}_A - \mathcal{V}\|_F^2 + \|\mathcal{V}_A^\top\mathcal{V}_A - I\|_F^2$;
    $\mathcal{V} \leftarrow \mathcal{V}_A = \mathrm{softmax}(\hat{\mathcal{D}}_A^{-\frac{1}{2}}\hat{\mathcal{S}}_A\hat{\mathcal{D}}_A^{-\frac{1}{2}}\mathcal{A}P_2)$;
    for iter ∈ 0, 1, ..., MaxIter:
        $P_1 \leftarrow \min_{P_1} \mathcal{L}_1 := \frac{1}{2N}\|\mathcal{V} - \mathcal{V}_X\|_F^2 + \|\mathcal{V}_X^\top\mathcal{V}_X - I\|_F^2$;
Return: $\mathcal{V}_{X^u}$ and $\mathcal{V}_{A^u}$ according to Equation (4).
```

To calculate the cognition-alignment score via ZSIR, the process involves evaluating the system's ability to correctly match queries with their corresponding instances in a gallery, where both queries and gallery instances belong to unseen classes $\mathcal{Y}^u$. This evaluation is conducted under two distinct scenarios: attributes to image (A2I) and image to attributes (I2A). In the A2I scenario, the system is provided with a set of attributes as the query. The objective is to accurately retrieve the visual instance in the image gallery that best matches these attributes. Conversely, in the I2A scenario, the system is given a query image and must predict its identity by matching it to the exact attribute instance in the gallery. For both scenarios, the initial step involves inferring the full instance attribute vectors from the given class attributes, along with the visual features of the query, and then projecting them into the orthogonal space. The retrieval process occurs within this space, where the system attempts to find the closest match between the query's projected representation and the projected representations of instances in the gallery.

The cognition-alignment score is derived from the accuracy of these retrieval tasks. It quantifies the system's proficiency in aligning its data-driven representations (inferred attribute vectors and visual features) with human cognitive processes (dominant attributes and visual identity). High accuracy in retrieval, reflected in a high cognition-alignment score, indicates effective alignment, demonstrating the system's capability to generalise and accurately interpret unseen classes based on the cognitive congruence between its learnt representations and human-understandable attributes.

4. Experimental Results

4.1. Experimental Setup

Datasets: The assessment of our approach was conducted using two established benchmarks for ZSIR: the SUN dataset introduced by Patterson et al. [27] and the CUB dataset presented by Wah et al. [28], with detailed results presented in Table 2. Both SUN and CUB are fine-grained tasks, where SUN contains 717 classes of scene images, while CUB has 200 classes of birds. The visual features leveraged in our study are derived from [29]. For the purpose of word embedding, our study utilised the conventional GoogleNews-vectors-negative 300 [1] , which underwent training on a segment of the Google News dataset, encompassing approximately 100 billion words. Our methodology adheres to the traditional splits between seen and unseen classes typical in zero-shot learning (ZSL) frameworks [20], with an emphasis placed on evaluating ZSIR capabilities. In scenarios involving image-to-attribute (I2A) and attribute-to-image (A2I) conversions, attributes and images of unseen instances are interchangeably utilised as gallery and query sets.

Table 2. Key statistics of the experimental datasets. For both datasets, the attributes were provided by a human annotator at the instance (ins.) level with either continuous values (cont.) or binary values (bin.). # indicates the number.

Dataset	SUN	CUB
# of instances	14,340	11,788
# of attributes	102	312
seen/unseen splits	707/10	150/50
attribute type	ins. + cont.	ins. + bin.
# of total concepts	819	512
unseen gallery size	200	2933

Evaluation Methodology: The primary metric for our evaluation is the hit rate, e.g., the accuracy (%) of instance retrieval which assesses whether a given query's corresponding instance can be retrieved within the top ranks. To provide a comprehensive overview, we calculated the average hit rates across various classes, reflecting the general performance trend.

Implementation Details: We employed a cross-validation approach for all hyperparameters within LIAD on the training dataset. Given the absence of attribute usage during training, we introduced a five-fold cross-validation strategy tailored for the CA challenge. This involves initially determining $\mathcal{V}$ across the entire training dataset, which represents the dominant attributes' inferred outputs, denoted by $\mathcal{V}_A$. The training classes were subsequently segmented into five groups. For each group, we calculated a new pair of projections, P_1 and P_2, utilising the remaining four groups. The obtained P_1 was then applied to map visual instances from the validation group to $\mathcal{V}_X$. The retrieval performance was assessed by comparing $\mathcal{V}_A^u$ and $\mathcal{V}_X^u$ for unseen classes.

4.2. ZSIR Main Results

Table 3 presents a comprehensive evaluation of our proposed method for Zero-Shot Instance Retrieval (ZSIR) using human attributes, benchmarked against both baseline and state-of-the-art approaches on the SUN Attribute and CUB datasets. Our method significantly outperforms existing methods across all ranks, showcasing its effectiveness in ZSIR tasks. Baseline methods such as DAP, ALE, ESZSL, LatEm, and LIAD show varying degrees of success, with LIAD previously leading with scores up to 28.7% at Rank1 and 86.2% at Rank50. In comparison, our approach not only surpasses these baselines but also demonstrates superior performance to additional methods like CCA and the Siamese Network, particularly noted in the more challenging scenario of retrieving images based on attributes (A2I) and vice versa (I2A).

Table 3. Main ZSIR results (shown as hit rate accuracy from Rank1 to Rank50) using human attributes compared to state-of-the-art approaches. The first- and second-half sections demonstrate I2A and A2I retrieval, respectively.

	SUN Attribute Dataset					**CUB Dataset**				
Methods	@Rank1	@Rank5	@Rank10	@Rank20	@Rank50	@Rank1	@Rank5	@Rank10	@Rank20	@Rank50
DAP [16]	7.5	18.8	34.9	48.5	61.2	3.80	5.82	12.61	17.92	24.25
ALE [30]	14.8	29.6	47.5	64.2	78.4	7.81	18.23	22.52	30.74	38.72
ESZSL [31]	19.9	38.8	56.2	69.7	82.8	15.28	20.34	25.88	38.21	40.72
LatEm [32]	25.3	38.4	62.8	70.1	85.2	17.42	24.82	32.48	40.96	46.81
LIAD [1]	28.7	42.2	68.5	72.8	86.2	19.82	27.53	36.20	44.12	48.83
CCA	8.3	18.2	33.2	56.2	63.2	7.63	11.32	18.89	27.53	28.76
Siamese Network	12.8	22.5	40.2	57.2	69.8	8.52	12.42	18.92	28.42	30.79
Ours (orthogonal only)	26.6	38.2	58.8	65.2	79.9	17.72	26.85	29.97	37.72	40.12
Ours (PG only)	28.9	44.6	69.7	74.4	87.7	20.28	28.82	38.83	46.62	50.53
Ours	**35.5**	**49.8**	**71.0**	**79.9**	**92.8**	**25.52**	**32.74**	**48.85**	**52.88**	**59.92**
DAP [16]	8.8	19.2	32.6	44.7	52.5	5.42	8.82	14.27	16.82	22.36
ALE [30]	12.2	26.7	43.0	61.5	72.2	12.87	16.43	24.50	29.98	34.71
ESZSL [31]	18.8	34.2	49.1	66.2	76.9	14.31	17.40	23.65	36.48	39.22
LatEm [32]	17.3	36.4	58.8	67.6	80.8	15.82	20.26	29.48	36.25	43.82
LIAD [1]	18.7	37.7	61.9	70.2	78.8	18.61	26.62	32.81	39.42	44.28
CCA	13.8	27.4	44.5	62.8	70.7	10.43	14.52	18.85	25.58	30.76
Siamese Network	15.5	30.2	49.9	58.8	69.4	11.13	18.82	24.95	31.10	37.74
Ours (orthogonal only)	17.2	35.2	58.8	64.9	72.2	17.72	24.32	28.81	35.52	39.98
Ours (PG only)	18.9	38.1	63.2	73.2	79.1	19.21	27.78	37.75	42.29	46.62
Ours	**20.5**	**40.2**	**65.5**	**75.8**	**82.2**	**28.21**	**30.87**	**39.92**	**44.97**	**48.82**

A key innovation in our method is the introduction of the prototype grouping (PG) technique, which significantly enhances the diversity of prototypes for each visual–semantic attribute and strengthens inter-class associations. This is evident from the performance leap observed when comparing our method's orthogonal-only and PG-only variants to the combined approach. Specifically, our full method achieves remarkable improvements, reaching up to 35.5% at Rank1 and 92.8% at Rank50 for the SUN Attribute dataset, outperforming the PG-only variant's 28.9% at Rank1 and 87.7% at Rank50, and the orthogonal-only variant's 26.6% at Rank1 and 79.9% at Rank50.

These results underscore the efficacy of our method in generalising ZSIR to unseen domains through enhanced attribute representation and cognitive alignment. The prototype grouping method, in particular, stands out as a pivotal advancement, enabling more nuanced and contextually rich retrieval outcomes that closely align with human cognitive processes. This breakthrough underscores the potential of our approach in bridging the gap between AI-driven visual recognition and human-like understanding. The evaluation of our methods ensures reliable alignment between visual and cognition spaces and the method's ability to generalise to unseen classes. Qualitative results in Table 4 and the ablation study are discussed as follows.

Table 4. Comparison of human attributes, AI-revised human attributes, AI-generated attributes, and ZSL-contextualised AI.

Human Attributes	AI-Revised Human Attributes	AI-Generated Attributes	ZSL-Contextualised AI Attributes
'sailing/boating'	Open Space	Natural	Natural
'driving'	Enclosed Space	Man-made	Man-Made
'biking'	Natural Landscape	Indoor	Indoor
'transporting things or people'	Man-made Structures	Outdoor	Outdoor
'sunbathing'	Urban Environment	Urban	Urban
'vacationing/touring'	Rural Setting	Rural	Rural
'hiking'	Water Presence	Modern	Bright
'climbing'	Vegetation Density	Historical	Dim
'camping'	Color Palette	Spacious	Colorful
'reading'	Textural Qualities	Cramped	Monochrome
'studying/learning'	Lighting Conditions	Bright	Spacious
'teaching/training'	Weather Elements	Dim	Cramped
'research'	Architectural Style	Colorful	Populated
'diving'	Historical Context	Monochrome	Deserted
'swimming'	Modern Elements	Busy	Vegetated
'bathing'	Artistic Features	Tranquil	Barren
'eating'	Functional Aspects	Populated	Watery
'cleaning'	State of Maintenance	Deserted	Dry
'socializing'	Population Density	Greenery	Mountainous
'congregating'	Noise Level	Barren	Flat
'waiting in line/queuing'	Movement Dynamics	Waterbody	Forested
'competing'	Activity Presence	Dry	Open
'sports'	Cultural Significance	Mountainous	Enclosed
'exercise'	Geographical Features	Flat	Architectural
'playing'	Seasonal Characteristics	Forested	Naturalistic
'gaming'	Time of Day	Open	Ornate
'spectating/being in an audience'	Material Dominance	Enclosed	Simple
'farming'	Symmetry	Architectural	Cluttered
'constructing/building'	Asymmetry	Naturalistic	Minimalistic
'shopping'	Spaciousness	Ornate	Artistic
'medical activity'	Clutter	Simple	Functional
'working'	Tranquility	Cluttered	Symmetrical
'using tools'	Bustle	Minimalistic	Asymmetrical
'digging'	Accessibility	Artistic	Traditional
'conducting business'	Seclusion	Functional	Contemporary
'praying'	Safety Perception	Symmetrical	Luxurious
'fencing'	Risk Elements	Asymmetrical	Modest
'railing'	Sensory Stimuli	Traditional	Cultivated
'wire'	Emotional Atmosphere	Contemporary	Wild
'railroad'	Privacy Level	Luxurious	Paved
'trees'	Connectivity	Modest	Unpaved
'grass'	Isolation	Cultivated	Vibrant
'vegetation'	Ecological Elements	Wild	Muted
'shrubbery'	Industrial Presence	Paved	Textured
'foliage'	Commercial Features	Unpaved	Smooth
'leaves'	Educational Aspects	Vibrant	Reflective
'flowers'	Recreational Facilities	Muted	Matte
'asphalt'	Religious Symbols	Textured	Elevated
'pavement'	Cultural Diversity	Smooth	Ground-level
'shingles'	Historical Monuments	Reflective	Aerial

4.3. Ablation Study

In our ablation study, we meticulously analysed the impact of two critical components of our framework: orthogonal projection and prototype grouping (PG). This analysis is grounded in the comparative performance of our method against established baselines, as delineated in the results table.

Effect of Orthogonal Projection: The influence of orthogonal projection on Zero-Shot Instance Retrieval (ZSIR) performance is evident when comparing the outcomes of Canonical Correlation Analysis (CCA), the Siamese Network, LIAD, and our method with orthogonal projection only. CCA, which focuses on extracting correlation information between visual features and attributes without imposing specific constraints, offers a foundational comparison point. The Siamese Network, leveraging a deep architecture based on triplet contrastive learning, aims to minimise distances within classes while maximising distances between classes, offering a nuanced approach to learning separable feature spaces. LIAD, incorporating an orthogonal constraint alongside augmented attributes, introduces a structured approach to aligning visual and semantic spaces. Our method, when employing orthogonal projection exclusively, demonstrates a marked improvement over these baselines, underscoring the efficacy of orthogonal constraints in enhancing cognitive alignment and retrieval accuracy. Specifically, the orthogonal-only variant of our method outperforms CCA and the Siamese Network across all ranks, indicating orthogonal projection's pivotal role in achieving more discriminative and well-aligned feature representations.

Effect of Prototype Grouping: The prototype grouping (PG) mechanism's contribution is highlighted through a comparison between our method's PG variant and the baseline approaches. The PG approach is designed to foster more diverse and representative prototypes for each visual–semantic attribute, thereby facilitating better generalisation to unseen classes through enhanced inter-class associations. The results table reveals that our method with PG significantly surpasses the performance of all baseline methods, including the orthogonal projection variant. This superiority is particularly pronounced at higher ranks, suggesting that PG effectively captures the complex underlying structures of the data, enabling the more accurate retrieval of unseen instances. The comparison underscores PG's critical role in bridging the semantic gap between visual features and attributes, thereby bolstering the model's zero-shot retrieval capabilities.

The results provide a quantitative testament to the individual and combined impacts of orthogonal projection and prototype grouping. Notably, our method, which integrates both components, outperforms all other approaches, consistently achieving the highest retrieval accuracy across the board. This comprehensive performance boost, observed across different datasets and ranking metrics, attests to the synergistic effect of orthogonal projection and PG in refining the model's ability to navigate the complex landscape of ZSIR. The orthogonal projection's role in structuring the feature space, coupled with PG's enhancement of prototype diversity and inter-class connectivity, culminates in a robust framework that adeptly aligns AI's cognitive processes with human-like understanding. These findings not only validate the proposed components' effectiveness but also pave the way for future explorations into optimising ZSIR frameworks for improved cognitive alignment and retrieval performance.

4.4. Cognition Alignment Analysis

The analysis of the results in Figure 5 for ZSIR on the SUN and CUB datasets provides insightful observations into the performance of various AI approaches in comparison to human attributes. This analysis is pivotal for understanding the cognitive alignment between AI-generated attributes and human perception in the context of ZSIR tasks.

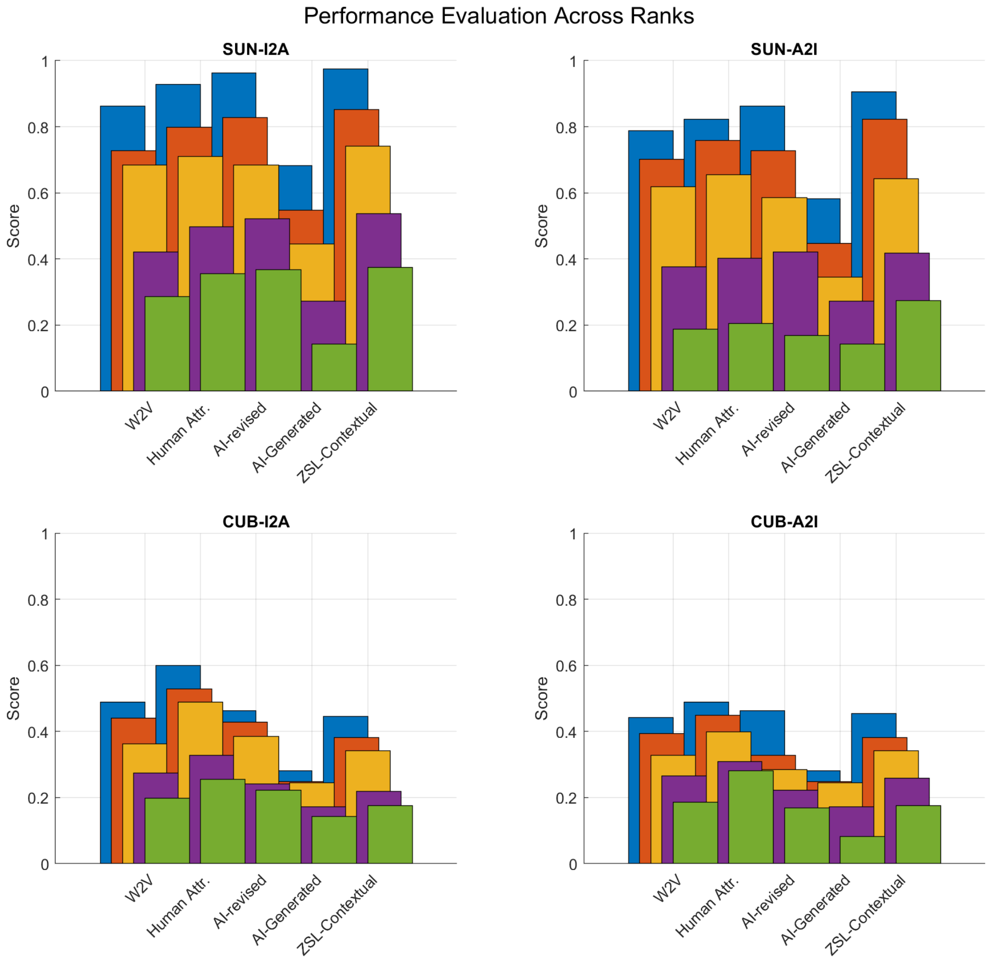

Figure 5. Cognition alignment reflected by the performance evaluation across ranks (green, purple, yellow, orange, and blue indicate retrieval rate 1, 5, 10, 20 and 50.

4.5. Observations and Discussion

W2V Word Embedding: The Word2Vec (W2V) embeddings exhibit stable performance across both tasks (image to attributes and attributes to image), slightly trailing behind the results achieved using human attributes. This consistency underscores the robustness of W2V embeddings in capturing semantic relationships, albeit with a marginal gap in cognitive alignment compared to human-derived attributes.

AI-Revised Attributes' Performance: The AI-revised attributes, while maintaining the conceptual framework of attributes defined by human experts, show an interesting dichotomy in performance. On the SUN dataset, these revised attributes outperform both W2V embeddings and human attributes, suggesting a closer alignment with AI's visual–semantic understanding for this dataset. Conversely, on the CUB dataset, their performance dips below that of both W2V and human attributes. This variation highlights the context-dependent effectiveness of AI revisions, particularly struggling with the fine-grained classification required by the CUB dataset. Moreover, Table 5 demonstrates the ability that the AI can explain the rationale of each assigned attribute value based on common sense and human-understandable language. This may become an efficient paradigm for future data annotation approaches.

AI-Generated Attributes' Limitations: The approach based on generating attributes and their values entirely through AI results in the lowest performance across all tasks and datasets. This outcome points to a significant misalignment in the AI's generation process with the specific demands of visual-semantic learning and zero-shot generalisation.

The lack of proper scoping using prompt engineering in attribute generation and value assignment critically hampers the effectiveness of this method.

Table 5. Augmented attribute example: AI-assigned class–attribute association values and the rationale explained by AI as to why the association makes sense.

Attributes	Value	Association with the Class Abbey as an Example
Natural	0	(Abbeys are man-made structures, though they may be situated in natural settings.)
Indoor	1	(Abbeys typically have significant indoor areas.)
Outdoor	1	(They also have outdoor components like courtyards.)
Urban	0	(Generally, abbeys are in rural or secluded settings, but some may be in urban areas.)
Bright	0	(Traditional abbeys might have a dimmer, more solemn interior.)
Colorful	0	(Abbeys often have a more muted, monastic color scheme.)
Spacious	1	(They usually have spacious interiors like chapels and halls.)
Populated	0	(Abbeys are often associated with tranquility and seclusion.)
Vegetated	1	(Many abbeys have gardens or are located in green settings.)
Watery	0	(Unless located near a water body, which is not typical for all abbeys.)
Mountainous	0	(This is location-dependent.)
Forested	0	(Again, location-dependent.)
Open	1	(They often have open courtyards.)
Enclosed	1	(Enclosed structures like cloisters are common.)
Architectural	1	(Abbeys are known for their distinctive architecture.)
Ornate	1	(Many abbeys are ornately decorated, especially older ones.)
Simple	0	(Abbeys are typically not simple in design.)
Artistic	1	(Abbeys often contain artistic elements like stained glass.)
Symmetrical	1	(Many have symmetrical architectural designs.)
Modest	1	(Abbeys are often modest in terms of luxury.)
Cultivated	1	(Gardens or cultivated lands are common.)
Paved	1	(Pathways and internal floors are typically paved.)
Textured	1	(Stone walls, woodwork, etc.)
Elevated	0	(Dependent on the specific location.)
Underground	0	(Some abbeys may have crypts or basements.)
Foggy	0	(Location-specific.)
Daytime	1	(Abbeys are typically functional during the day.)
Weathered	1	(Many abbeys are old and show signs of weathering.)
Secluded	1	(Abbeys are often in secluded locations.)
Quiet	1	(Associated with quietude and peace.)
Cool	1	(Stone buildings often have a cool interior.)

ZSL-Contextualised Attributes Success: When the LLM of ChatGPT is informed about the ZSL and image retrieval tasks, the resulting ZSL-contextualised attributes significantly improve performance, particularly on the SUN dataset. This improvement indicates a higher degree of cognitive alignment, as the attributes and their values are more precisely tailored to the tasks at hand.

Challenges with Fine-Grained Classification: Despite the advancements in AI approaches, human attributes remain the superior cognitive representation for the CUB dataset, which demands extensive expert knowledge for fine-grained bird classification. The ZSL-contextualised model and W2V embedding exhibit similar performance in this domain, underscoring the challenge that AI faces in matching human expertise in highly specialised tasks.

The comparative analysis of AI approaches against human attributes in ZSIR tasks reveals critical insights into the cognitive alignment of AI with human perception. While AI-revised and AI-generated attributes show potential under certain conditions, they also highlight the limitations of current AI methodologies in fully grasping the nuances of visual–semantic relationships, especially in specialised domains like fine-grained classification. The success of ZSL-contextualised attributes on the SUN dataset opens promising avenues for enhancing cognitive alignment through task-aware attribute generation.

5. Conclusions

In this study, we have introduced a groundbreaking paradigm that leverages ZSIR to delve into the cognitive alignment of large visual language models with human cognitive processes. Our approach, focused on a novel unified similarity function, marks a significant stride in understanding how AI systems interpret and process visual information in scenarios involving previously unseen data. The rigorous evaluation of our framework across the CUB and SUN datasets has not only validated the effectiveness of our similarity function but also highlighted its adaptability across various knowledge representations, including visual attributes and textual descriptions generated by AI models. Our findings underscore the potential of our method to serve as a benchmark for cognitive alignment in AI, demonstrating superior performance in ZSIR tasks compared to existing state-of-the-art approaches on both the SUN (92.8% and 82.2%) and CUB datasets (59.92% and 48.82%) for image-to-attribute and attribute-to-image retrieval accuracy. This research contributes to the broader goal of developing AI technologies that can seamlessly integrate with human-centric applications, ensuring that AI systems can interpret and respond to the world in ways that mirror human thought and understanding. The AI-annotated attributes significantly reduced the time cost compared to human approaches. The AI also provided human-understand explanation about the rationale of each assigned attribute value. This finding may have strong implication for future high-level data annotation industry.

Several avenues for future research emerge from our findings. First, exploring the application of our unified similarity function across a wider array of datasets and in more diverse scenarios could further validate its robustness and versatility. Additionally, integrating our approach with other forms of knowledge representation, such as video or audio data, could offer deeper insights into the cognitive alignment of AI across different sensory modalities. Another promising direction involves refining the similarity function to accommodate dynamic learning environments, where AI systems continuously adapt to new information in a manner akin to human learning. Finally, investigating the ethical implications of cognition-aligned AI systems and their impact on society will be crucial as these technologies become increasingly prevalent in everyday life. Through these future endeavours, we aim to advance the field of AI towards more intuitive, human-like understanding and interaction with the world, fostering the development of ethical and cognitively aligned AI systems.

Author Contributions: Conceptualisation, T.M. and W.M.; methodology, T.M. and Y.L.; software, T.M. and Y.L.; formal analysis, D.O., T.M. and Y.L.; writing—original draft preparation, T.M.; writing—review and editing, D.O., T.M. and Y.L.; visualisation, T.M.; supervision, W.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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

References

1. Long, Y.; Liu, L.; Shen, Y.; Shao, L. Towards affordable semantic searching: Zero-shot retrieval via dominant attributes. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018 ; Volume 32.
2. Bereska, L.; Gavves, E. Taming Simulators: Challenges, Pathways and Vision for the Alignment of Large Language Models. *arXiv* **2023**, arXiv:2308.01317.
3. Wang, X.; Li, X.; Yin, Z.; Wu, Y.; Liu, J. Emotional Intelligence of Large Language Models. *arXiv* **2023**, arXiv:2307.09042.
4. Ren, Y.; Xiong, D. CogAlign: Learning to Align Textual Neural Representations to Cognitive Language Processing Signals. *arXiv* **2021**, arXiv:2107.06354.
5. Xu, G.; Liu, J.; Yan, M.; Xu, H.; Si, J.; Zhou, Z.; Yi, P.; Gao, X.; Sang, J.; Zhang, R.; et al. CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility. *arXiv* **2023**, arXiv:2307.09705.
6. Sengupta, N.; Sahu, S.K.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Marshall, W.; Gosal, G.; Liu, C.; Chen, Z.; et al. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv* **2023**, arXiv:2308.16149.

7. Zhang, S.; Fang, Q.; Zhang, Z.; Ma, Z.; Zhou, Y.; Huang, L.; Bu, M.; Gui, S.; Chen, Y.; Chen, X.; et al. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models. *arXiv* **2023**, arXiv:2306.10968.
8. Wang, P.; Li, L.; Chen, L.; Song, F.; Lin, B.; Cao, Y.; Liu, T.; Sui, Z. Making Large Language Models Better Reasoners with Alignment. *arXiv* **2023**, arXiv:2309.02144.
9. Bhardwaj, R.; Poria, S. Red-Teaming Large Language Models using Chain of Utterances for Safety-Alignment. *arXiv* **2023**, arXiv:2308.09662.
10. Liu, Y.; Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Xiong, D. Large Language Model Alignment: A Survey. *arXiv* **2023**, arXiv:2308.05374.
11. Gu, Z.; Fan, J.; Tang, N.; Zhang, S.; Zhang, Y.; Chen, Z.; Cao, L.; Li, G.; Madden, S.; Du, X. Interleaving Pre-Trained Language Models and Large Language Models for Zero-Shot NL2SQL Generation. *arXiv* **2023**, arXiv:2306.08891.
12. Kirk, H.R.; Vidgen, B.; Röttger, P.; Hale, S.A. Personalisation within Bounds: A Risk Taxonomy and Policy Framework for the Alignment of Large Language Models with Personalised Feedback. *arXiv* **2023**, arXiv:2303.05453.
13. Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.H.; Riedel, S. Language Models as Knowledge Bases? *arXiv* **2019**, arXiv:1909.01066.
14. Larochelle, H.; Erhan, D.; Bengio, Y. Zero-data learning of new tasks. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, IL, USA, 13–17 July 2009.
15. Palatucci, M.; Pomerleau, D.; Hinton, G.E.; Mitchell, T.M. Zero-shot learning with semantic output codes. In Proceedings of the Neural Information Processing Systems 2009, Vancouver, BC, USA, 7–10 December 2009.
16. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
17. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.A.; Mikolov, T. DeViSE: A deep visual-semantic embedding model. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013.
18. Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.S.; Dean, J. Zero-shot learning by convex combination of semantic embeddings. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
19. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized classifiers for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
20. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 2251–2265. [CrossRef] [PubMed]
21. Liu, W.; Long, M.; Wang, J.; Jordan, M.I. Generalized zero-shot learning with deep calibration network. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018.
22. Wang, W.; Zheng, V.W.; Yu, H.; Miao, C. A survey on zero-shot learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *10*, 1–33. [CrossRef]
23. Khandelwal, S.; Nambirajan, A.; Siddiquie, B.; Eledath, J.; Sigal, L. Frustratingly Simple but Effective Zero-shot Detection and Segmentation: Analysis and a Strong Baseline. *arXiv* **2023**, arXiv:2302.07319.
24. Chen, C.; Xiao, B.; Zhang, Y.; Zhu, Z. Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition. *Autom. Constr.* **2023**, *104*, 104702. [CrossRef]
25. Díaz, G.; Sobron, I.; Eizmendi, I.; Landa, I.; Velez, M. CSI-Based Cross-Domain Activity Recognition via Zero-Shot Prototypical Networks. *arXiv* **2023**, arXiv:2312.07076.
26. Nag, S.; Goldstein, O.; Roy-Chowdhury, A.K. Semantics Guided Contrastive Learning of Transformers for Zero-shot Temporal Activity Detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023.
27. Patterson, G.; Xu, C.; Su, H.; Hays, J. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *Int. J. Comput. Vis.* **2014**, *108*, 59–81. [CrossRef]
28. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-UCSD Birds-200-2011 Dataset*; Technical Report CNS-TR-2011-001; California Institute of Technology: Pasadena, CA, USA, 2011.
29. Zhang, Z.; Saligrama, V. Zero-shot learning via semantic similarity embedding. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
30. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for attribute-based classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
31. Romera-Paredes, B.; Torr, P.H.S. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 7–9 July 2015.
32. Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; Schiele, B. Latent embeddings for zero-shot classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

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.