LEVERAGING METAPATHS FOR LEARNING FROM KNOWLEDGE GRAPHS IN THE CONTEXT OF VISION-BASED CLASSIFICATION OF OBJECT STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-Shot Object State Classification (ZS-OSC) aims to recognize unseen object states without any visual training examples. Existing methods typically rely on Knowledge Graphs (KGs) to provide semantic information about states, but they often treat KGs as homogeneous, overlooking the rich relational knowledge encoded in their structure. We propose a novel approach to ZS-OSC[1] that leverages meta-paths to capture complex relationships between object states in a KG. Our method learns to project semantic information from the KG into the visual space via meta-path learning, generating discriminative visual embeddings for unseen state classes. To the best of our knowledge, this is the first work to utilize meta-paths for ZS-OSC. We conduct extensive experiments on four benchmark datasets, demonstrating the superior performance of our approach compared to SoTA zero-shot learning methods and a graph-based baseline. Our ablation study further provides insights into the impact of key design choices on the effectiveness of our method.

1 INTRODUCTION

Knowledge graphs (KGs) have become increasingly important in addressing various computer vision (CV) tasks, such as object classification (Marino et al., 2017), zero-shot recognition (Wang et al., 2018b) and Visual Question Answering (Krishna et al., 2017). This surge in KG utilization can be attributed to their ability to provide rich semantic information and contextual knowledge that can enhance the understanding of visual data. However, current approaches often under-utilize the full potential of KGs. Several factors contribute to this sub-optimal utilization. First, KGs used in Computer Vision (CV) are often large and contain irrelevant or erroneous information, which can introduce noise and hinder performance. Second, many methods fail to consider the diversity of edge types in KGs, treating all edges as homogeneous and overlooking valuable relational knowledge. This simplistic approach limits the ability to effectively exploit the rich semantic information embedded within KGs.

A number of approaches such as filtering mechanisms (Wang et al., 2014; Domingos & Richardson, 2007), ad-hoc KG construction (Dong et al., 2014; Gouidis et al., 2024) and random walking (Perozzi et al., 2014; Grover & Leskovec, 2016) were proposed as strategies against these shortcomings. An alternative approach focusing on the overcoming of these limitations concerned the concept of meta-paths (Sun et al., 2011; Dong et al., 2017). The utilization of meta-paths essentially involves the learning of the relative importance of the different paths between nodes within the KG via the assignment of weights to them. Meta-path-based approaches offer several advantages when applied to heterogeneous information networks (HINs). First, meta-paths enable the modeling of complex semantic relationships by explicitly defining sequences of node and edge types (Figure 1), making them particularly suitable for capturing the diverse interactions in multi-typed networks (Sun et al., 2011; Shi et al., 2014). This provides a significant advantage over traditional graph learning methods (Sun et al., 2011) , which treat all nodes and edges as homogeneous, thus failing to leverage the rich information embedded in HINs. Furthermore, meta-paths allow for task-specific traversal and filtering, enabling more accurate representations for applications such as link prediction, recom-

[1]The code can be found at https://anonymous.4open.science/r/Metapaths-7811/

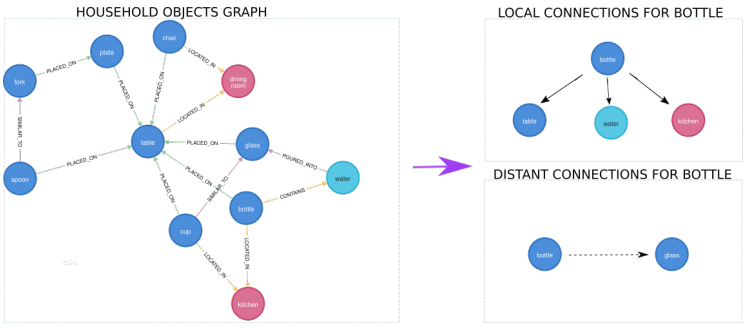

Figure 1: Unlike standard graph learning methods that prioritize local connections, meta-paths can capture distant relationships. In this toy example of a household objects graph, a meta-path can detect the stronger connection in specific contexts between “bottle” and “glass” (linked indirectly) than between ”bottle” and “table”, ”bottle” and “kitchen” and ”bottle” and “water” (directly linked).

mendation systems, and node classification. The ability to define domain-specific meta-paths also allows for more precise similarity measures, improving performance in tasks like clustering and search. Importantly, meta-paths enhance the interpretability of models by making the relationships between entities more transparent, providing domain experts with insights into how predictions are generated (Xiong et al., 2017).

Motivated by the great potential that the meta-paths learning seems to hold, this work attempts to explore the utilization of this approach in the context of the Object State Classification (OSC) task (Isola et al., 2015; Gouidis et al., 2022; Souček et al., 2022; Saini et al., 2023), which is a CV task attracting growing research attention over the last few years. OSC concerns the recognition of object states appearing in images and videos and is closely related to the more popular problems of Object Recognition and Action Recognition. OSC is an important problem whose solution is of significant impact. The recognition of object states and state changes is crucial for determining an object‘s condition and the interaction that was performed upon or could be performed in the future on it (Jamone et al., 2016). Moreover, the capacity for efficient OSC is of primary importance in AI systems that support tasks such as learning object affordances (Chuang et al., 2018), recognizing interactions (Wang et al., 2016b; Isola et al., 2015; Liu et al., 2017; Mancini et al., 2022), reasoning to achieve an object state change (Farhadi et al., 2009), recognizing the completion or failure of goals and recovery from possible mistakes during procedural activities (Schoonbeek et al., 2024) and many others. Meanwhile, large-scale video datasets (Grauman et al., 2022; Saini et al., 2023) concerning human-object interactions provide rich annotation data which are related to object state changes enabling the definition of new problems and the establishment of benchmarks and challenges related to object state detection and classification (Grauman et al., 2022).

This work addresses the challenging task of Zero-Shot Object State Classification (ZS-OSC), where the goal is to classify object states without any visual training examples. The key challenge lies in leveraging auxiliary non-visual information to enable the classification of these unseen states. Existing state-of-the-art methods typically utilize KGs as sources of structured semantic information (Gouidis et al., 2023), but they often treat KGs as homogeneous, potentially overlooking valuable relational information. We propose a novel approach that leverages meta-paths for more effective learning of zero-shot representations. Our method learns to project semantic information from a KG into the visual space via meta-path learning, generating visual embeddings for unseen state classes. To the best of our knowledge, this is the first method to utilize meta-paths in the context of Zero-Shot Visual Classification, and therefore ZS-OSC, through embedding generation.

Beyond their demonstrated utility in ZS-OSC, we posit that the generation of meta-path-based embeddings as a primary research objective holds significant promise. This direction offers several compelling advantages. First, meta-path-based embeddings can capture both local and global semantic relationships in heterogeneous networks, providing universal entity representations applicable across diverse tasks and domains without requiring task-specific fine-tuning. This fosters the development of multi-purpose, generalizable embeddings. Second, generating embeddings with a focus on representation quality can facilitate knowledge transfer across domains. Third, while cur-

rent embedding techniques, such as (Dong et al., 2017), are often task-specific, a framework that prioritizes the generation of meta-path embeddings as an independent objective would offer a generalized tool for analyzing heterogeneous information networks. This would mitigate the need for task-dependent tuning and enable researchers to investigate embeddings without being constrained by a particular task. Finally, since meta-paths effectively capture complex, higher-order relationships within networks, prioritizing the efficient generation of meta-path-based embeddings can lead to more compact and informative representations, optimizing storage and computational efficiency, particularly in large-scale heterogeneous networks.

This work makes the following key contributions:

- We introduce a novel method for generating embeddings by leveraging meta-paths within Graph Neural Networks (GNNs). This approach enables GNNs to effectively harness the rich information encoded in KGs. To the best of our knowledge, this is the first work to explore meta-path utilization for embedding generation in this context.
- We conduct a comprehensive ablation study to analyze the impact of various design choices and parameters on the performance of our proposed method. This analysis provides valuable insights into the interplay between meta-path learning and embedding generation.
- We perform an extensive experimental evaluation on four diverse datasets, comparing our method against established KG-based baselines and SoTA Large Pre-trained Models. The results demonstrate that our approach achieves superior performance by a significant margin.

2 Related Work

Meta-path Learning: Meta-paths, a fundamental concept in heterogeneous information networks (HINs) (Shi et al., 2016), have been widely studied in applications such as similarity search, recommendation systems, and link prediction. Meta-paths were introduced to capture complex relationships in HINs, enabling the study of object proximities and connectivity patterns. The work by Sun et al. (2011) proposed PathSim that used meta-paths to measure similarity between objects based on shared relationships, with success in applications like similarity search and clustering. Extending this, Shi et al. (2014) proposed HeteSim, which computes relevance between objects via meta-paths, incorporating directionality and node types for enhanced flexibility. Meta-paths have also enhanced recommendation systems in HINs. Yu et al. (2013) developed a collaborative filtering algorithm incorporating meta-path-based similarities between users and items, improving recommendation accuracy. Similarly, Wang et al. (2016a) used meta-path-based features in matrix factorization for item recommendation, leveraging HIN structure to model user-item interactions more effectively.

Learning optimal meta-path weights for specific tasks has been a key research area. Dong et al. (2017) introduced MetaPath2Vec, which learns node embeddings through meta-path-based random walks, showing improved performance in classification and clustering tasks. Fu et al. (2020) extended this with a scalable meta-path-guided graph neural network, learning meta-path importance for tasks in large-scale HINs. In link prediction, meta-paths have been used to predict missing or future links in networks. Liu et al. (2018) proposed a meta-path-based link prediction method, capturing complex node interactions and outperforming traditional algorithms in networks with multiple types of nodes. Recent advancements in graph neural networks (GNNs) have further improved meta-path-based link prediction. Zhang et al. (2019) introduced a heterogeneous graph neural network (HGNN) that integrates meta-path-based features into node aggregation, achieving better link prediction.

Sun & Han (2013) introduced the concept of meta-paths for mining HINs, capturing semantic relationships across data types. Automatic discovery methods for meta-paths were later explored by Meng et al. (2015) to address challenges in manually retrieving meta-paths. Ferrini et al. (2024) proposed a novel approach to enhance GNN accuracy through effective meta-path identification, while Noori et al. (2023) explored meta-paths for flexible similarity search in biological knowledge graphs. Additionally, Yun et al. (2022) discussed learning new graph structures using meta-paths, demonstrating their role in enhancing GNN performance. Recent studies have also focused on challenges such as automatic meta-path discovery (Huang et al., 2020), integration with deep learning models (Wang et al., 2019), and the application of meta-paths in dynamic HINs (Trivedi et al., 2019).

The main novelty of our work distinguishing it from the aforementioned works concerns the utilization of meta-paths in the context of embeddings generation. Although many actual works focusing on meta-paths utilize embeddings, they serve as a means for another goal, i.e. downstream task, such as link prediction or entity alignment. In contrast, in our work, the generation of embeddings is a final objective.

Object State and Attribute Classification: Visual attributes are commonly defined as visual concepts that are both machine-detectable and human-understandable (Duan et al., 2012). The prevailing approach to learning attributes mirrors that of object classes, where a convolutional neural network is trained with discriminative classifiers using annotated image datasets (Singh & Lee, 2016). However, existing labeled attribute datasets suffer from limitations such as smaller scale compared to object datasets, a restricted number of generic attributes, and limited category coverage (Lampert et al., 2009; Isola et al., 2015; Patterson & Hays, 2016; Yu & Grauman, 2017; Mancini et al., 2022). Furthermore, research specifically focusing on state classification remains limited (Gouidis et al., 2022), with most existing approaches relying on similar assumptions and techniques as those employed for attribute classification. This highlights a need for dedicated methods tailored to the unique challenges of state recognition.

Zero-shot Classification: Zero-shot learning (ZSL) has attracted significant attention due to its ability to address the practical challenge of classifying objects without any training examples (Xian et al., 2018a). This is particularly crucial in real-world scenarios where obtaining labeled data for every possible class is often infeasible. Several approaches have been proposed for zero-shot object classification, including semantic embedding-based methods (Wang et al., 2018a; Xian et al., 2018b), attribute-based methods (Lampert et al., 2014), generative models (Xian et al., 2018b; Changpinyo et al., 2016), and learning compatibility functions between image and class embeddings (Akata et al., 2015). Semantic embedding methods utilize compact semantic spaces or attribute sets to bridge the gap between seen and unseen object classes. Attribute-based methods leverage descriptive attributes to infer the class of unseen objects. Generative models synthesize images of unseen classes based on similarities to seen classes. Additionally, recent work has explored the use of knowledge graphs to capture semantic relationships between objects and facilitate ZSL (Kampffmeyer et al., 2019; Nayak & Bach, 2022). While ZSL has been extensively studied for object recognition, its application to Zero State Classification (ZSC) remains relatively unexplored. With the exception of Gouidis et al. (2023), which focuses exclusively on state classification without relying on prior knowledge about object classes, existing ZSC methods primarily address the Compositional Zero-Shot Learning (CZSL) variant of the problem. CZSL aims to generalize to unseen combinations of object and state primitives by learning their compositionality from the training set (Misra et al., 2017; Nagarajan & Grauman, 2018; Yang et al., 2020).

3 METHODOLOGY

This work introduces a novel approach for generating embeddings in heterogeneous graphs by synergistically combining KGs structures with meta-path-based GNNs. These generated embeddings are then employed for the task of ZS-OSC. Our methodology harnesses the strengths of both KGs embeddings and the rich relational information encapsulated by meta-paths. The core idea is to leverage a designated set of KG nodes as guides for meta-path learning. Specifically, we utilize the visual embeddings of these nodes as ground truth and train a Graph Transformer Network (GTN) to assign weights to different KG edge types. This is achieved by generating embeddings for the guide nodes that align with their visual embeddings. The GTN architecture is inspired by Yun et al. (2022), while the training procedure draws inspiration from Kampffmeyer et al. (2019); Gouidis et al. (2023). Figure 2 provides a general overview of our method.

3.1 PRELIMINARIES

Before proceeding with describing our method, it is necessary to present related backgrounds related to our work. Additionally, Table S1 presented in the supplementary section lists commonly used notations in this paper for quick reference.

Heterogeneous Graph: A heterogeneous graph is denoted as $G = (V, E, \phi, \psi)$ and is associated with a node type mapping function $\phi : V \rightarrow T_v$ and an edge type mapping function $\psi : E \rightarrow T_e$,

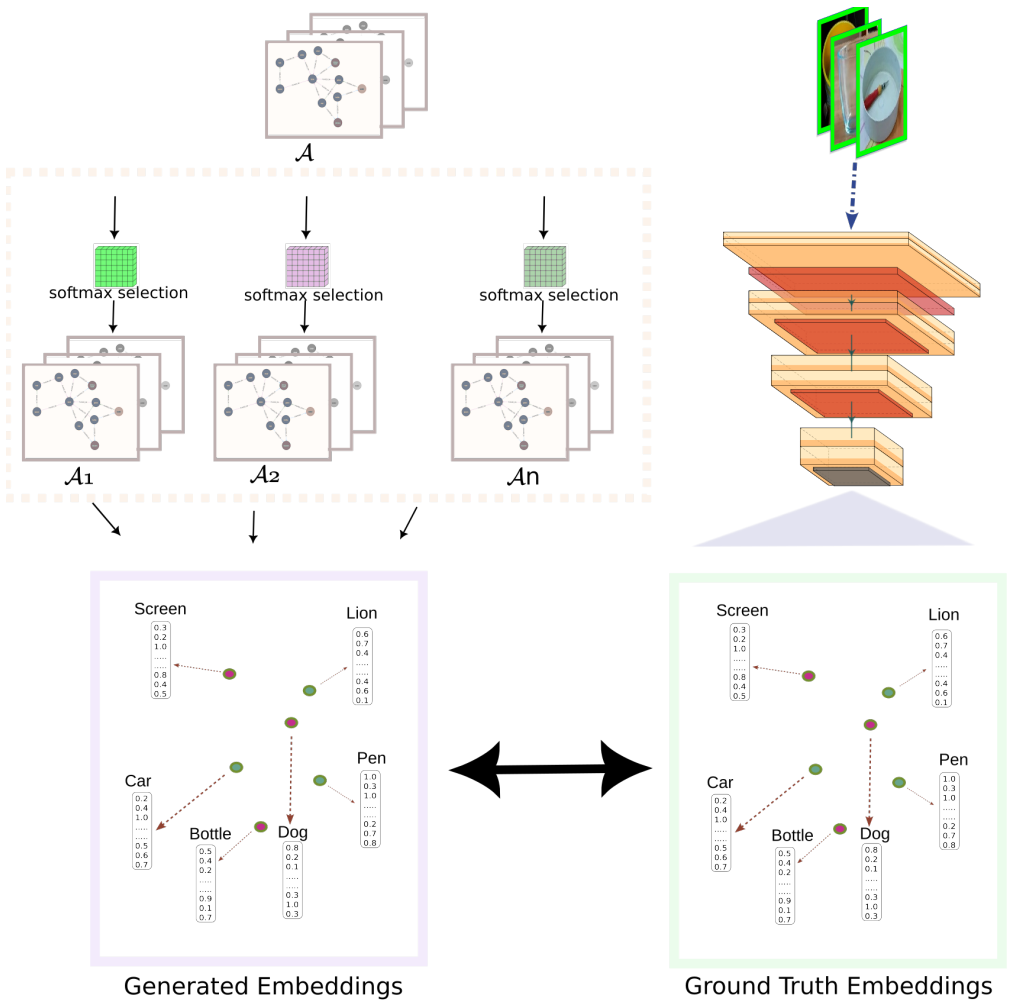

Figure 2: Overview of the proposed approach. A Graph Transformer Network (GTN) with a softmax selection mechanism assigns weights to different meta-paths. Meta-paths are aggregated and combined with node features to generate node embeddings. Visual ground truth embeddings for a set of guide nodes are produced using a pre-trained convolutional neural network. These embeddings guide the training of the GTN.

where T_v and T_e denote the set of nodes types and edges types respectively. A necessary requirement for a graph to be considered heterogeneous is to contained more than one type of nodes or more than one type of edges, i.e., $|T_v| + |T_e| > 2$.

Metapath: Within heterogeneous graphs, a meta-path represents a sequence of connected nodes traversed through diverse edge types. Formally, it can be defined as a path:

$$v_1 \xrightarrow{\psi(e_1)} v_2 \xrightarrow{\psi(e_2)} \ldots \xrightarrow{\psi(e_\ell)} v_{\ell+1},$$

where each edge e_ℓ in the path has a corresponding edge type $\tau_e(e_\ell)$, belonging to the set of all possible edge types T_e.

Metapath Instance: A metapath instance p of a metapath P in a heterogeneous graph is defined as a sequence of nodes that follows the schema defined by P.

Metapath-based Neighbor: For a node v and a metapath P in a heterogeneous graph, the set of metapath-based neighbors $N_P(v)$ is defined as the set of nodes connected to v via metapath instances

of P. A neighbor that is connected through two different metapath instances is treated as two distinct nodes in $N_P(v)$. If P is symmetric, $N_P(v)$ includes the node v itself.

Metapath-based Graph: For a given metapath P, the metapath-based graph G_P is the graph constructed using all metapath-based neighbor pairs from the original graph G. If P is symmetric, G_P is homogeneous.

Heterogeneous Graph Embedding: Given a heterogeneous graph $G = (V, E)$, with node attribute matrices $X_{A_i} \in \mathbb{R}^{|V_{A_i}| \times d_{A_i}}$ for each node type $A_i \in \mathcal{A}$, the goal of heterogeneous graph embedding is to learn d-dimensional node representations $h_v \in \mathbb{R}^d$ for each node $v \in V$, where $d \ll |V|$, capturing rich structural and semantic information from G.

3.2 META-PATH-BASED EMBEDDING GENERATION

Meta-paths represent semantic connections between different node types in a heterogeneous graph. GTNs automatically learn useful meta-paths by selecting and combining adjacency matrices of different edge types. This process allows the model to generate new graph structures that are useful and is used typically for downstream tasks such as node classification. In our case, the meta-paths learning is not mediated by a downstream task. Instead, meta-paths are learned via the generation of embeddings for the graph's nodes.

Following the notation introduced previously, we can use adjacency matrices, a different one for each edge type, to compute the different meta-paths. Specifically, the meta-path's adjacency matrix $A_P \in \{0, 1\}^{\mathbb{V} \times \mathbb{V}}$ of the graph is computed as:

$$A_P = \prod_{i=1}^{\ell} A_{t_i} \Longrightarrow A_P = A_{t_1} A_{t_2} \dots A_{t_\ell},$$

where A_{t_i} is the adjacency matrix corresponding to edge type t_i and l corresponds to the length of the meta-paths, with $a_{ij} = 1$ denoting a meta-path with length equal than l between nodes i and j and $a_{ij} = 0$ an absence of meta-path, respectively. In order to have also meta-paths with lengths less than l we add the identity matrix $I \in \mathbb{R}^{\mathcal{V} \times \mathcal{V}}$ to the adjacency matrices.

With the utilization of a soft selection mechanism, weights are assigned to the different types of edges. This technique enables the learning of the optimal combination of edge types for each meta-path. Specifically, this is achieved with a 1×1 convolution with softmax activation over the adjacency matrices of different edge types:

$$WA^{(k)} = \sum_{t=1}^{|T_e|} \alpha_t^{(k)} A_t,$$

where $\alpha_t^{(k)}$ is the learnt weight for edge type t at the k-th transformer layer. The resulting meta-path adjacency matrices are normalized using degree matrices D and applied to perform multi-hop convolutions. More than one softmax convolutions could be also applied, in which case the objective is that each channel learns a different representation and the overall combination of the different representations results in more a robust outcome.

With the learnt softmax weights, the meta-path's adjacency matrix $WA_P \in \mathbb{R}^{\mathbb{V} \times \mathbb{V}}$ of the graph is computed as:

$$WA_P = \prod_{i=1}^{\ell} A_{t_i} \Longrightarrow \sum_{t=0}^{|T_e|} \alpha_t^{(k)} A_t = \sum_{t=1}^{|T_e|} \alpha^{(0)} A_{t_0} \sum_{t=1}^{|T_e|} \alpha^{(1)} A_{t_1} \sum_{t=1}^{|T_e|} \alpha^{(2)} A_{t_2} \cdots \sum_{t=1}^{|T_e|} \alpha^{(l)} A_{t_l}.$$

After having computed the weighted adjacency matrix WA_P we can compute the aggregation meta-paths-based features matrix $\mathbb{X}_P \in \mathbb{R}^{|V_{A_i}| \times d_{A_i}}$ for the graphs nodes by multiplying the nodes features matrix matrices $\mathbb{X}_{A_i} \in \mathbb{R}^{|V_{A_i}| \times d_{A_i}}$ with WA_P:

$$\mathbb{X}_P = \mathbb{X}_A \times WA_P.$$

In order to generate embeddings for the graphs nodes we use a multi-layer Graph Neural Network which learns to project the nodes features into the target space:

$$\mathbb{E}_P = f_\theta(\mathbb{X}_P).$$

The weights corresponding to the softmax layers and the stacked layers of the GNN are updated based on the minimization of a L2 distance function $\mathcal{L}_\mathcal{E}$ between the generated nodes embeddings and the ground truth nodes embeddings for the set of nodes which serve as ground truths:

$$\mathcal{L}_\mathcal{E} = \frac{1}{2N} \sum_{n \in \mathcal{N}} \sum_{d \in \mathcal{D}} (\mathbb{E}_P - \tilde{\mathbb{E}}_P)^2,$$

where $\mathcal{D}$ is the dimension of the embeddings and $\mathcal{N}$ the number of the ground truth concepts, respectively.

3.3 Zero-Shot Object State Classification

To achieve zero-shot object-state classification, it is crucial to learn representations for the unseen target state classes. To this end, we construct a KG with state classes as nodes and utilize the GTN trained in the previous step for meta-path learning to generate embeddings for these target classes. These embeddings are then integrated into a visual classifier, specifically a pre-trained CNN, to serve as visual representations of the visually known state classes. Following established practices in transfer learning and zero-shot learning, we replace the final layer of the pre-trained CNN with these generated embeddings. To ensure compatibility, the dimensionality ($\mathcal{D}$) of the generated embeddings is set to match the dimension of the CN's last layer. This adaptation enables the CNN to effectively classify the target state classes in a zero-shot manner.

4 Experimental Evaluation

4.1 Implementation and evaluation issues

Implementation Details: The KG provided as input to the Graph Transformer Network (GTN) for meta-path learning was the WordNet hierarchy of the 1000 classes from the ImageNet1000 dataset (Russakovsky et al., 2015). These 1000 classes served as the ground truth set for training the GTN, which was trained for 200 epochs with five different learning rates (see Section 4.2 for details). The same KG was also used to generate the target object state classes. In experiments with multiple softmax channels, the generated embeddings were averaged. The convolutional neural network (CNN) used for zero-shot object-state classification (ZS-OSC) was ResNet-101 (He et al., 2016) pre-trained on the ImageNet1000 dataset.

Datasets: We utilized four publicly available datasets containing object state annotations: OSDD Gouidis et al. (2022), CGQA Mancini et al. (2022), MIT Isola et al. (2015), and VAW Pham et al. (2021). While OSDD is specifically designed for state detection, the other three are attribute datasets that include object states among their classes. We extracted subsets concerning exclusively object state classes. For OSDD and VAW, bounding boxes from the original images were extracted to create images suitable for the OSC task. Table S2 in the supplementary section provides details on the four datasets.

Metrics: Our evaluation follows the zero-shot method from (Purushwalkam et al., 2019). Following those guidelines we calculate accuracy per class and then average these instead of reporting overall accuracy. This ensures each class is equally important, regardless of its sample size.

4.2 Ablation Study

This ablation study investigates the optimal configuration of key parameters related to meta-path-based embedding generation, including: (a) the maximum length of meta-paths, (b) the number of softmax channels used for meta-path selection, and (c) the training learning rate.

Meta-paths Max Length	OSDD	CGQA-States	MIT-States	VAW-States
1	27.1	44.9	47.1	29.3
2	29.5	46.7	47.3	29.8
3	**31.3**	**47.6**	**48.7**	**32.1**
4	30.2	46.5	47.9	31.0

Table 1: Ablation results for maximum length of meta-paths. The number of softmax channels in the GTN was 3. The networks were trained with a learning rate equal to $1e-2$.

Number of Channels	OSDD	CGQA-States	MIT-States	VAW-States
1	27.2	44.3	46.2	28.5
2	28.9	46.1	46.4	29.7
3	**31.3**	**47.6**	**48.7**	**32.1**
4	30.5	46.9	47.0	30.7

Table 2: Ablation results for number of softmax channels. The maximum length of the meta-paths was 3. The networks were trained with a learning rate equal to $1e-2$.

Length of meta-path: Table 1 presents the results of varying the maximum length of meta-paths (1, 2, 3, and 4 hops). The best performance across all datasets was achieved with a maximum length of 3 hops. Performance generally improved with increasing meta-path length, likely because longer paths capture more global graph information. However, performance slightly decreased with a length of 4 hops, suggesting that excessively long paths might introduce noise, potentially by incorporating less relevant or spurious relationships. This finding highlights the importance of carefully selecting the appropriate meta-path length to balance information gain and noise reduction.

Number of Softmax Channels: Table 2 shows the impact of varying the number of softmax channels used for meta-path selection. The best performance was achieved with 3 channels, suggesting that multiple channels allow the GTN to learn diverse meta-path representations, thereby improving embedding quality. Increasing the number of channels allows the model to capture different aspects of the relationships encoded in the meta-paths, leading to richer and more informative embeddings. However, using too many channels, e.g., 4 in this case, might not provide further benefits and could potentially increase model complexity without a corresponding improvement in performance.

Training Learning Rate: Table 3 presents the results of using different learning rates (LR) during GTN training. The best performance was obtained with a learning rate of 1e-2. Smaller learning rates led to a significant performance drop, indicating the importance of this parameter for effective model training. This suggests that a learning rate of 1e-2 strikes a good balance between convergence speed and stability, allowing the model to effectively learn from the data without overshooting or getting stuck in local optima. The observed performance drop with smaller learning rates could be attributed to slower convergence and potential difficulties in escaping local optima.

4.3 Comparison to Competing Methods

This experiment had a two-fold objective. First, we compared our method against SoTA Large Pre-trained Models (LPMs) capable of ZS-OSC. We used six different prompts[2] related to the target states and report the mean average performance across all prompts, following standard convention. Second, we compared our approach to a graph-based method (Gouidis et al., 2023) specifically designed for ZS-OSC, which allows us to assess the impact of meta-paths on performance. For comparison with the baseline method (Gouidis et al., 2023), which relies on random walks, we used five different random seeds for initialization and report the mean performance over all seeds. The same KG was used as input to both the GCNs of the baseline method and our method. It is important to note that other ZSL methods, such as CZSL, are not applicable in this context, because they require information about object classes, which is not available in this zero-shot setting.

[2]The prompts are presented in the supplementary material.

LR	OSDD	CGQA-States	MIT-States	VAW-States
$5e-2$	29.2	46.2	47.5	30.1
$1e-2$	**31.3**	**47.6**	**48.7**	**32.1**
$5e-3$	28.9	45.1	47.0	28.5
$1e-3$	27.2	44.5	46.1	26.2
$5e-4$	26.1	43.8	44.5	24.6

Table 3: Ablation results for learning rate. The maximum length of the meta-paths was 3, and the number of softmax channels was 3.

Method	OSDD	CGQA-States	MIT-States	VAW-States
Baseline (Gouidis et al., 2023)	27.3	45.1	43.3	25.6
CLIP-RN101 (Radford et al., 2021)	22.5	**46.9**	39.3	28.0
CLIP-VITBP16 (Radford et al., 2021)	28.8	44.9	46.4	**30.1**
CLIP-VITLP14 (Radford et al., 2021)	28.4	43.4	**48.6**	27.9
ALIGN (Jia et al., 2021)	**29.5**	40.0	44.2	28.4
BLIP (Li et al., 2022)	13.3	26.0	27.2	16.1
Ours	**31.3**	**47.6**	**48.7**	**32.1**
Our improvement over the SoTA	1.8	0.7	0.1	2.0
Our improvement over the Baseline	4.0	2.4	5.4	6.5

Table 4: Results against competing methods. Bold/Blue text indicates best/2nd best performance.

Table 4 presents the results. Our method achieved the best performance across all datasets, outperforming both the baseline method and all LPMs. Notably, our method surpassed the baseline by a large margin (4.0%, 2.5%, 5.4%, and 6.5% for OSDD, CGQA, MIT, and VAW, respectively).This outcome highlights the effectiveness of meta-path learning compared to random walks for representation learning in graph structures. The superior performance of our method can be attributed to the ability of meta-paths to capture and exploit specific semantic relationships within the KG, leading to more informative and discriminative embeddings for ZS-OSC. In contrast, random walk techniques might not effectively capture these relationships, resulting in less effective representations.

Furthermore, the fact that our method outperforms SoTA LPMs (1.8%, 0.7%, 0.1%, and 2.0% for OSDD, CGQA, MIT, and VAW, respectively) demonstrates the potential of incorporating KG information and meta-path learning into ZS-OSC. LPMs, while powerful, might not fully capture the nuanced semantic relationships between objects and their states that are encoded in KGs. By leveraging meta-paths to extract and utilize this information, our method achieves a significant performance improvement.

5 Conclusion

This paper introduced a novel method for generating embeddings in heterogeneous graphs by leveraging meta-paths within a graph neural network framework. Our approach utilizes KGs structures and visual embeddings to guide the learning of meta-paths, enabling the generation of informative and discriminative embeddings. We demonstrated the effectiveness of our method in the context of ZS-OSC, achieving superior performance compared to state-of-the-art LPMs and a graph-based baseline. Our ablation study provided insights into the impact of key parameters on the performance of our method.

Future work will focus on several promising directions. First, we aim to explore the applicability of our method to other zero shot CV tasks such as semantic segmentation, image captioning and visual question answering to further evaluate its generalizability. Second, we plan to investigate the integration of different types of KGs such as multimodal KGs and KGs containg different type of nodes and explore the impact of KG quality on embedding generation. Finally, we intend to extend our approach to incorporate more complex meta-path structures and explore alternative graph neural network architectures for enhanced performance.

REFERENCES

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2927–2936, 2015. 4

Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5327–5336, 2016. 4

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 975–983, 2018. 2

Pedro Domingos and Matthew Richardson. Markov logic: A unifying framework for statistical relational learning. 2007. 1

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601–610, 2014. 1

Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 135–144. ACM, 2017. 1, 3

Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3474–3481. IEEE, 2012. 4

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 1778–1785, 2009. doi: 10.1109/CVPRW.2009.5206772. 2

Francesco Ferrini, Antonio Longa, Andrea Passerini, and Manfred Jaeger. Meta-path learning for multi-relational graph neural networks. In *Learning on Graphs Conference*, pp. 2–1. PMLR, 2024. 3

Xiaoyang Fu, Jiawei Zhang, Zitao Meng, and S Yu Philip. Scalable graph neural networks for heterogeneous networks. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1001–1010, 2020. 3

F. Gouidis, T. Patkos, A. Argyros, and D. Plexousakis. Detecting object states vs detecting objects: A new dataset and a quantitative experimental study. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, volume 5, pp. 590–600, 2022. 2, 4, 7, 15

Filipos Gouidis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis. Leveraging knowledge graphs for zero-shot object-agnostic state classification. *arXiv preprint arXiv:2307.12179*, 2023. 2, 4, 8, 9

Filippos Gouidis, Konstantinos Papoutsakis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis. Exploring the impact of knowledge graphs on zero-shot visual object state classification. *Proceedings Copyright*, 738:749, 2024. 1

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan

Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022. 2

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016. 1

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 7

Shaohua Huang, Xiao He, Ruifeng Jiang, Li Song, Xinyu Xie, and Hongyuan Ma. Meta-path guided heterogeneous graph neural network for intent recommendation. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, pp. 2249–2256, 2020. 3

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1383–1391, 2015. ISSN 10636919. doi: 10.1109/CVPR.2015.7298744. 2, 4, 7, 15

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1):4–25, 2016. 2

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021. 9

Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:11479–11488, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01175. 4

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 951–958. IEEE, 2009. 4

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorizationa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. ISSN 01628828. doi: 10.1109/TPAMI.2013.140. 4

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022. 9

Xiaolei Liu, Xiaodan Liang, Zhiqiang Li, Jie Tang, and Li Lin. Link prediction in heterogeneous networks using meta-paths. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 437–446, 2018. 3

Yang Liu, Ping Wei, and Song-Chun Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2924–2932, 2017. 2

Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning Graph Embeddings for Open World Compositional Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(c):1–15, 2022. ISSN 19393539. doi: 10.1109/TPAMI.2022.3163667. 2, 4, 7, 15

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2673–2681, 2017. 1

Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, and Wangda Zhang. Discovering meta-paths in large heterogeneous information networks. In *Proceedings of the 24th international conference on world wide web*, pp. 754–764, 2015. 3

Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1792–1801, 2017. 4

Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 169–185, 2018. 4

N. V. Nayak and S. H. Bach. Zero-shot learning with common sense knowledge graphs. *Transactions on Machine Learning Research (TMLR)*, 2022. 4

Ayush Noori, Michelle M Li, Amelia LM Tan, and Marinka Zitnik. Metapaths: similarity search in heterogeneous knowledge graphs via meta-paths. *Bioinformatics*, 39(5):btad297, 2023. 3

Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pp. 85–100. Springer, 2016. 4

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014. 1

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF CVPR*, pp. 13018–13028, June 2021. 7, 15

Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3593–3602, 2019. 7

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 9

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 7

N. Saini, H. Wang, A. Swaminathan, V. Jayasundara, B. He, K. Gupta, and A. Shrivastava. Chop amp; learn: Recognizing and generating object-state compositions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20190–20201, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.01852. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01852. 2

Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons van der Sommen, et al. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4365–4374, 2024. 2

Chuan Shi, Bin Hu, Xiaowei Zhao, and Philip S Yu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2479–2492, 2014. 1, 3

Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1): 17–37, 2016. 3

Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pp. 753–769. Springer, 2016. 4

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Multi-Task Learning of Object State Changes from Uncurated Videos. 2022. URL http://arxiv.org/abs/2211.13500. 2

Yizhou Sun and Jiawei Han. Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Science and Technology*, 18(4):329–338, 2013. 3

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 992–1000. ACM, 2011. 1, 3

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Dyngem: Deep embedding method for dynamic heterogeneous information networks. In *Proceedings of the 2019 World Wide Web Conference*, pp. 1132–1143, 2019. 3

Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1225–1234. ACM, 2016a. 3

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The World Wide Web Conference*, pp. 2022–2032, 2019. 3

Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions $\sim$ Transformations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pp. 2658–2667. IEEE, jun 2016b. ISBN 9781467388504. doi: 10.1109/CVPR.2016.291. URL http://ieeexplore.ieee.org/document/7780660/. 2

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6857–6866, 2018a. ISSN 10636919. doi: 10.1109/CVPR.2018.00717. 4

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6857–6866, 2018b. 1

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014. 1

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018a. 4

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5542–5551, 2018b. 4

Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 565–574, 2017. 2

Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10248–10256, 2020. 4

Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5570–5579, 2017. 4

Xin Yu, Zhaochun Ren, Yizhou Sun, Quanquan Gu, Zhiqiang Luo, and Dawei Chen. Collaborative filtering with entity similarity regularization in heterogeneous information networks. In *IJCAI International Joint Conference on Artificial Intelligence*, pp. 21–29, 2013. 3

Seongjun Yun, Minbyul Jeong, Sungdong Yoo, Seunghun Lee, S Yi Sean, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks: Learning meta-path graphs to improve gnns. *Neural Networks*, 153:104–119, 2022. 3, 4

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 793–803. ACM, 2019. 3

6 SUPPLEMENTARY MATERIAL

This section presents tables that due to space limitation were omitted from the main body of the paper. Specifically, Table S1 shows the notation symbols that are used, Table S2 presents the details about the datasets used in the experimental evaluation and Table S3 shows the prompts that were using for the LPMs respectively.

Symbol	Description
$G = (V, E)$	Heterogeneous graph
V	Set of nodes in G
E	Set of edges in G
$\phi : V \rightarrow T_v$	Node type mapping function
$\psi : E \rightarrow T_e$	Edge type mapping function
T_v	Set of node types
T_e	Set of edge types
$\|T_v\|$	Number of node types
$\|T_e\|$	Number of edge types
P	Meta-path
p	Meta-path instance
v_i	Node in G
e_l	Edge in G
$\tau_e(e_l)$	Edge type of edge e_l
$N_P(v)$	Set of metapath-based neighbors of node v under meta-path P
G_P	Metapath-based graph derived from meta-path P
X_A	Node feature matrix
$\mathcal{A}$	Adjacency matrix
$\mathcal{A}_\mathcal{P}$	Meta-path adjacency matrix
$\mathbb{E}_P$	Meta-path based generated embedding of node v
$\tilde{\mathbb{E}}_P$	Ground truth embedding of node v
$\mathcal{D}$	Dimensionality of node embeddings
$\mathcal{N}$	Number of ground truth concepts
$\mathcal{L}_\mathcal{E}$	L2 distance loss function

Table S1: Notation Table

Dataset	**Train**	**Val**	**Test**	**States**	**Objects**	**VOSC**	**TOSC**	**S\O**
OSDD (Gouidis et al., 2022)	6,977	1,124	5,275	9	14	35	126	2.36
CGQA-states (Mancini et al., 2022)	244	46	806	5	17	41	75	1.71
MIT-states (Isola et al., 2015)	170	34	274	5	14	20	70	1.57
VAW (Pham et al., 2021)	2,752	516	1,584	9	23	51	207	2.61

Table S2: Details about the four image datasets utilized in this work. Train/Val/Test: Number of Training/Validation/Testing Images. States: Number of State classes, Objects: Number of Object classes. VOSC/TOSC: Valid/Total Object-State combinations. S\O: Average number of states than an Object can be situated in.

	Prompt
1	An image of a {} object
2	The object in the image is {}
3	The state of the object in the image is {}
4	The object in the image is currently {}
5	An image of a object in a state of {}
6	The scene depicts a object that appears to be {}

Table S3: Prompts used for the Large Pretained Models.