Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI, the Arts and Creativity

KALE: An Artwork Image Captioning System Augmented with Heterogeneous Graph

Yanbei Jiang , Krista A. Ehinger , Jey Han Lau
School of Computing and Information Systems, The University of Melbourne
yanbeij@student.unimelb.edu.au, kris.ehinger@unimelb.edu.au, laujh@unimelb.edu.au

Abstract

Exploring the narratives conveyed by fine-art paintings is a challenge in image captioning, where the goal is to generate descriptions that not only precisely represent the visual content but also offer a in-depth interpretation of the artwork's meaning. The task is particularly complex for artwork images due to their diverse interpretations and varied aesthetic principles across different artistic schools and styles. In response to this, we present KALE (**K**nowledge-**A**ugmented vision-**L**anguage model for artwork **E**laborations), a novel approach that enhances existing vision-language models by integrating artwork metadata as additional knowledge. KALE incorporates the metadata in two ways: firstly as direct textual input, and secondly through a multimodal heterogeneous knowledge graph. To optimize the learning of graph representations, we introduce a new *cross-modal alignment loss* that maximizes the similarity between the image and its corresponding metadata. Experimental results demonstrate that KALE achieves strong performance (when evaluated with CIDEr, in particular) over existing state-of-the-art work across several artwork datasets. Source code of the project is available at https://github.com/Yanbei-Jiang/Artwork-Interpretation.

1 Introduction

Recently, research in the field of Artificial Intelligence (AI) has shown a growing interest in exploring the intersection between AI and Art. In the last few years, numerous initiatives have aimed to leverage AI technologies to make the domain of art more accessible and interpretable [Ma *et al.*, 2017; Gonthier *et al.*, 2018; Wynen *et al.*, 2018]. One such application is the generation of descriptions for visual arts, which is a case of image captioning. This task aims to automatically produce a short meaningful text given an input image. Beyond simple object and scene recognition, effective image captioning requires machines to understand the context and relationships among the elements within the image. Through appropriate analysis and extraction of high-level features from artwork images, generated descriptions could potentially convey implicit meanings that artists want to express and make artworks more accessible.

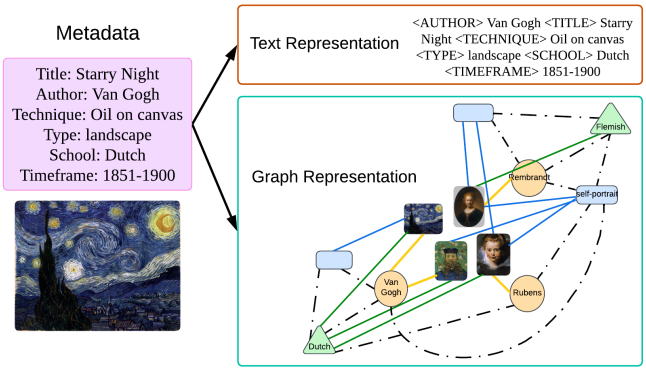

Figure 1: Artwork metadata are integrated through two ways, 1: Textual input 2: Knowledge graph input.

Generating captions for artwork images is a challenging task for several reasons. Firstly, unlike natural images, artwork images may lack clear entities, such as in the case of abstract art, making it difficult for models to extract useful information from just the images. Another significant challenge is dealing with the ambiguity and subjectivity inherent in the image. Artwork images often have multiple levels of interpretation, and captions may vary significantly depending on the observer's cultural background and artistic taste. A descriptive caption generator might say "some people under a tree in a park", but a better captioning system designed for artworks might say "a serene gathering in the shade that illustrates 19th-century pastoral life".

Recent studies, such as those by [Lu *et al.*, 2022], [Achlioptas *et al.*, 2021] and [Wu, 2022], employed Transformer-based encoder-decoder architectures. However, these models often lack extensive pre-training, limiting their diversity and effectiveness due to training on relatively small datasets. [Cetinic, 2021] utilizes a pre-trained model CLIP, which marks a significant advancement, but it struggles with captions that require an understanding of broader knowledge of art history. To address this, recent studies began to integrate external knowledge during training. [Bai *et al.*, 2021] proposed a framework that utilizes external knowledge from Wikipedia, where the model detects objects in artwork im-

ages and retrieves relevant information. However, this approach has its limitations, as artwork images do not always present salient objects for detection. [Sheng and Moens, 2019] take a different approach by incorporating artwork types into a CNN-LSTM model, but their model is constrained by its reliance on a single external data source.

To handle the above limitations, our solution is to incorporate supplementary data that provides broader knowledge beyond the image. The SemArt artwork dataset [Garcia and Vogiatzis, 2018] offers a valuable resource as it enriches each image with additional metadata. As shown in Figure 1, each image is associated with six metadata, which provide useful background information about the artworks. For example, the "School" metadata could be used to infer the artist's style, while "Type" tells us about the format of the artwork. To incorporate these supplementary data into an image captioning system, our first approach concatenates these metadata into a word sequence and feed them as additional input. Our second approach constructs a heterogeneous knowledge graph that integrates artwork metadata with the images to build continuous representations for the metadata. These continuous representations can then be injected as additional input. In summary, our main contributions are listed as follows:

- We propose KALE, an artwork image captioning system that extends the existing pre-trained vision-language model to the art domain. Through empirical evaluations, we have demonstrated that KALE largely outperforms existing artwork image captioning models across several artwork datasets.
- We design two ways to incorporate artwork metadata into our system. Firstly, by including all of them as additional textual inputs, and secondly, through the construction of a heterogeneous knowledge graph by treating each image and its associated metadata as distinct node types. This methodology leverages the strengths of heterogeneous graph structures and bridges the gap between visual and textual data in artwork analysis. The results suggest that the model can foster a better understanding of the narratives behind fine art pieces.
- KALE is trained with two objectives that maximize the likelihood of generating the ground-truth caption and the similarity between the image and its corresponding metadata in the knowledge graph.

2 Related Work

Pre-trained Vision-Language Model. Motivated by the success of pre-trained language models like BERT [Devlin *et al.*, 2019] in natural language processing, pre-trained vision-language models have attracted significant attention in the multi-modal domain and they have shown remarkable performance for image captioning task [Radford *et al.*, 2021; Wang *et al.*, 2021; Wang *et al.*, 2022; Li *et al.*, 2022; Li *et al.*, 2020]. Pre-training on large-scale datasets such as COCO [Lin *et al.*, 2014], Visual Genome [Krishna *et al.*, 2017], and Conceptual Captions [Sharma *et al.*, 2018] allows transferring knowledge to downstream tasks with limited data and enables the models to recognize and understand a broader range of objects and contexts. Typically, these models are composed of four main components: a vision backbone used to extract features from an input image, a language backbone to process an input text, a fusion encoder that captures the intricate interactions between visual and linguistic elements, and a language decoder to generate captions.

Knowledge Graph. In recent years, the concept of knowledge graphs has gained increasing attention in the field of artificial intelligence, as they can be used to represent a wide range of information, from simple facts and relationships to complex entities and events. Recent studies on incorporating knowledge graphs into image captioning systems demonstrate a significant enhancement. One approach involves using scene graphs to represent structural relationships in images [Yang *et al.*, 2019]. Other methods, like the one proposed by [Zhao and Wu, 2023], construct multi-modal knowledge graphs that associate visual objects with named entities to generate more informative and accurate captions. In the field of artwork analysis, [Garcia *et al.*, 2020] created an art-specific graph that connects paintings with their related attributes and incorporated it into cross-model retrieval task.

Heterogeneous Knowledge Graph. Unlike traditional graphs which focus on homogeneous nodes and edges, Heterogeneous Knowledge Graph (HKG) includes a variety of node and edge types, allowing for the representation of multifaceted data from different domains. For instance, in the context of artworks, nodes could represent images, artists, artwork types, schools, or historical periods. Edges, meanwhile, could denote relationships such as "created by", "belongs to" or "influenced by". This rich structuring could represent the art world and capture some complex historical, cultural, and stylistic factors. Heterogeneous Attention Networks (HANs) are a recent innovation leveraging the richness of HKGs [Wang *et al.*, 2019]. HANs apply the attention mechanism selectively across different types of nodes and relationships in an HKG. The output for HAN is a set of graph embeddings. These embeddings are crucial as they translate the entities and relations present in the graph into a low-dimensional, dense, and continuous vector space, which could be trained in an end-to-end manner. One of the key concepts in HKGs is the "meta-path", which is a sequence of relations defining a composite relationship among multiple types of entities. They enable the extraction of complex, higher-order relationships by traversing different types of nodes and edges in a sequence and capture a specific kind of interaction or relationship within the graph.

3 Proposed Methods

3.1 Heterogeneous Graph Construction

The construction of our heterogeneous graph commences with the definition of nodes and edges. In this graph, the diversity of node types is a key feature. We aim to create a multidimensional representation of the artwork and have a deeper analysis and interpretation through the graph. The nodes include:

- **Artwork Image:** Includes all the images in the training set, and represents the visual component of the artwork

and providing a link to each metadata node.

- **Author:** Represents the creators of the artwork, such as Vermeer and Van Gogh. Authors are key to understanding the stylistic and historical context of a piece.
- **Title N-grams:** To capture the essence of each artwork title, we include a range of N-grams as nodes – specifically 1-gram, 2-gram, and 3-gram, and select the most common ones, which represent the crucial keywords or phrases that characterize each artwork.
- **Title Cluster:** We further enrich the textual dimension by incorporating Title Clusters. We first use Sentence-Bert [Reimers and Gurevych, 2019] to process the titles and use k-means with cosine similarity to create these clusters.
- **Technique:** Describes the methods and materials used in creating the artwork, such as *oil on canvas*, which are crucial for understanding the artwork's texture and style. This metadata also include specific details about an artwork's dimensions, such as $167 \times 124cm$. However, they often do not contribute meaningful insights into the artwork's style, so we use regular expressions to filter out these dimension data.
- **Type:** Represents the genre or category of the artwork, such as *portrait* and *still-life*.
- **School:** Denotes the group the artwork is associated with, such as French and Spanish, offering cultural and historical relevance.
- **Timeframe:** Describes the era or period in which the artwork was created, such as 1650-1700, aiding in historical contextualization.

We decide to exclude the metadata **Date** as it typically provides very specific year information and so has limited utility in the context of caption generation.

As depicted in Figure 2, our approach results in a multi-layered graph structure, where the artwork images act as the central nodes and form the innermost layer of the graph. Branching outward from this core are the various types of metadata, each constituting its own layer, which could be directly linked to the central artwork nodes. Edges in this graph are designed to connect across different layers. Each artwork image node is connected to all its associated metadata (shown as coloured line) and edges are also established between different types of metadata if they belong to the same artwork (shown as dot-dash line). Note that there are no direct edges within layers but connections could be established through meta-paths. The meta-path enables indirect connections between one type of nodes, providing a means to uncover deeper insights and relationships in the graph. Here are some example meta-paths we defined in the graph:

- **Artwork-Author-Artwork:** This meta-path connects artworks through their authors. It can be used to explore the range and diversity within an individual author's body of work.
- **Type-Ngrams-Type:** This meta-path links art types through common themes found in artwork titles, suggesting shared or overlapping concepts between two types.
- **Artwork-Timeframe-Artwork:** This path connects different artworks based on the time period in which they were created, allowing for analysis of historical trends or evolution in art.

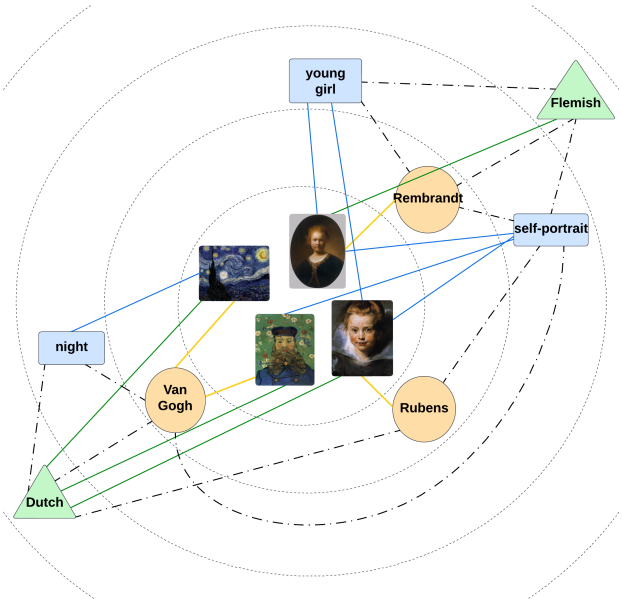

Figure 2: An example sub-graph of our multi-layer heterogeneous graph; only artwork, author (yellow), n-gram (blue) and school (green) nodes are included. For each type of nodes, it forms a layer. Solid colored lines denote image-metadata relationships, while dot-dashed lines represent inter-metadata linkages.

The last step is to create initial embeddings for nodes. We use ResNet50 [He *et al.*, 2016] to extract embeddings for images. For textual data such as Author, Title N-grams, and Technique, we use pre-trained FastText model [Joulin *et al.*, 2016]. For the Title Cluster nodes, we derive the embeddings by using the centroids of each cluster. For categorical data like Type, School, and Timeframe, we use one-hot encoding. The choice of pre-trained node embeddings is flexible in our architecture, and in future work it would be interesting to explore other pre-trained embeddings. In total, the resulting graph presents 28,796 nodes and 405,384 edges, with 20,310 images, 3,227 authors, 4,500 n-gram keywords, 100 clusters, 601 techniques, 26 schools, 22 timeframes and 10 types.

3.2 The KALE Model

At its core, KALE extends the existing pre-trained vision-language model and incorporates artwork metadata into the model through two ways: 1) as textual input, 2) through the inclusion of knowledge graph embeddings. Figure 3 depicts the general architecture for our KALE model, which has five main components: 1) Vision Encoder, 2) Text Encoder, 3) Graph Encoder, 4) Fusion Encoder, 5) Text Decoder. The general pipeline functions as follows:

Our input includes an image, artwork metadata, and a constructed heterogeneous graph with initial node embeddings.

The process begins with the image passing through the vision encoder, which yields the image embedding $\mathbf{v}_I$. Concurrently, the text encoder processes the artwork metadata to obtain the text embedding $\mathbf{v}_{\text{text}}$. Additionally, the graph is fed into the graph encoder, from which we extract the graph embedding for related nodes, denoted as $\mathbf{v}_{\text{graph}}$. These embeddings are then concatenated and input into the fusion encoder, which integrates the information from these three modalities into a fused embedding. Finally, the text decoder takes this fused embedding as input to generate captions. In the next few paragraphs, we introduce each of these components in detail.

Text Encoder. The text encoder is used to process and encode the metadata associated with each artwork as a textual input. Our approach involves concatenating various metadata elements into a single string, with each element separated by specially designed tokens. For instance, an author's name is preceded by an <AUTHOR> token, the title of an artwork by a <TITLE> token, and so on. Once concatenated, this string is then fed into a pre-trained language model, BERT [Devlin *et al.*, 2019]. This step results in the generation of text embeddings, which are vector representations capturing the semantic meaning of the metadata. Finally, we extract embeddings for only each of the special tokens and merge them to get the final textual representation. Formally,

$$\mathbf{v}_{\text{text}} = [\mathbf{e}_{<\text{AUTHOR}>}; \mathbf{e}_{<\text{TITLE}>}; \mathbf{e}_{<\text{TECHNIQUE}>}; \ldots] \quad (1)$$

where $\mathbf{e}_{<\ldots>}$ indicates the embedding obtained from BERT, and $[;]$ indicates concatenation.

Graph Encoder. The graph encoder is responsible for learning the embeddings for the nodes in the heterogeneous graph. The graph includes nodes representing different types of metadata, each with varying embedding sizes. To achieve uniformity in dimensionality, we first apply a linear layer followed by layer normalization to each node type in the graph. This step projects all node embeddings into a common dimension, referred to "Type-wise Feed Forward" in Figure 3. Mathematically, for a node of type t with initial embedding $\mathbf{v}_{\text{init}}^{(t)}$ from the graph, the transformation via a linear layer can be represented as:

$$\mathbf{v}_{\text{proj}}^{(t)} = \text{FFN}_t(\mathbf{v}_{\text{init}}^{(t)}) \quad (2)$$

Next, we use two Heterogeneous Attention Network (HAN) layers to process the graph. Initially, HAN focuses on node-level attention. This process involves computing attention coefficients for each node, taking into account its neighbors, highlighting the most significant connections based on node and edge types. Mathematically,

$$\alpha_{ij}^P = \text{softmax}\left(\sigma\left(\mathbf{a}_P^\intercal \cdot [\mathbf{W}_P\mathbf{v}_{\text{proj}_i} \| \mathbf{W}_P\mathbf{v}_{\text{proj}_j}]\right)\right) \quad (3)$$

where $\mathbf{W}_P$ is the weight matrix under meta-path P, $\mathbf{a}$ is the attention mechanism's learnable weight vector, and $\|$ denotes concatenation. The attention coefficients α_{ij}^P determine the importance of node j's features to node i. The node-level embeddings under a meta-path P, $\mathbf{v}_i^P$, are computed by aggregating these weighted features:

$$\mathbf{v}_i^P = \|_{k=1}^K \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^P(\mathbf{W}_P\mathbf{v}_{\text{proj}_j})\right) \quad (4)$$

where $\mathcal{N}(i)$ denotes the neighborhood of node i, σ represents an activation function and K is the number of heads in multi-head attention. HAN then extends this mechanism to a meta-path level, where it aggregates the node-level embeddings across different meta-paths or "relations" as different relations have their own parameters. Meta-path attention is similar to the node-level attention, the final graph embeddings, $\mathbf{v}_{\text{all}}$, are computed as:

$$e^p = \text{softmax}\left(\frac{1}{|S|}\sum_{i\in S}\left(\mathbf{q}^\intercal \cdot \tanh(\mathbf{W}\cdot\mathbf{v}_i^P + \mathbf{b})\right)\right) \quad (5)$$

$$\mathbf{v}_{\text{all}} = \sum_P e^P \cdot \mathbf{v}^P \quad (6)$$

where $\mathbf{q}$, $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters shared over all paths, S represents all the nodes in the graph and e^p represents the importance of meta-path P.

In the final stage, we extract the corresponding nodes in the graph based on the current input metadata and concatenate them to get the final graph embedding.

$$\mathbf{v}_{\text{graph}} = [\mathbf{v}_{\text{node}_1}; \mathbf{v}_{\text{node}_2}; \ldots], \text{where } \mathbf{v}_{\text{node}_i} \in \mathbf{v}_{\text{all}} \quad (7)$$

Note that during the test phase, sometimes the metadata may not be present in the training graph. To address this, for unseen Author, Title N-grams, Technique, we use FastText followed by the "Type-wise Feed Forward" layer in graph encoder to transform such metadata into embeddings. For Type, School, and Timeframe, we initialize to zero vectors.

Vision Encoder, Fusion Encoder, Text Decoder. In KALE the vision encoder, fusion encoder, and text decoder components are based on the architecture and weights of the pre-trained vision-language model, mPLUG, which achieved state-of-the-art performance in standard image captioning tasks [Li *et al.*, 2022]. As we are adapting vision-language models to to a new domain (i.e. from natural images to artwork images), we chose mPLUG given it was pretrainede on a rich set of image-text pairs which is likely to align well with our domain. Moreover, it uses a Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] which processes the input image by dividing it into a grid of regular patches, which is a more sensible approach for artwork images which often lack clear entities.

Multi-Task Training. We use multi-task learning to fine-tune our model. The first task utilizes the cross-entropy loss function, which is a standard approach in image captioning problems. Mathematically, this can be expressed as:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (8)$$

The second task introduces a cross-modal alignment loss. This loss function is designed to maximize the cosine similarity between the image embedding and all its corresponding metadata graph embeddings during training.

For the metadata graph embeddings $\mathbf{v}_{\text{graph}}$, which are obtained from the graph encoder, we pass them through a linear layer to project into the same space as the image. For the

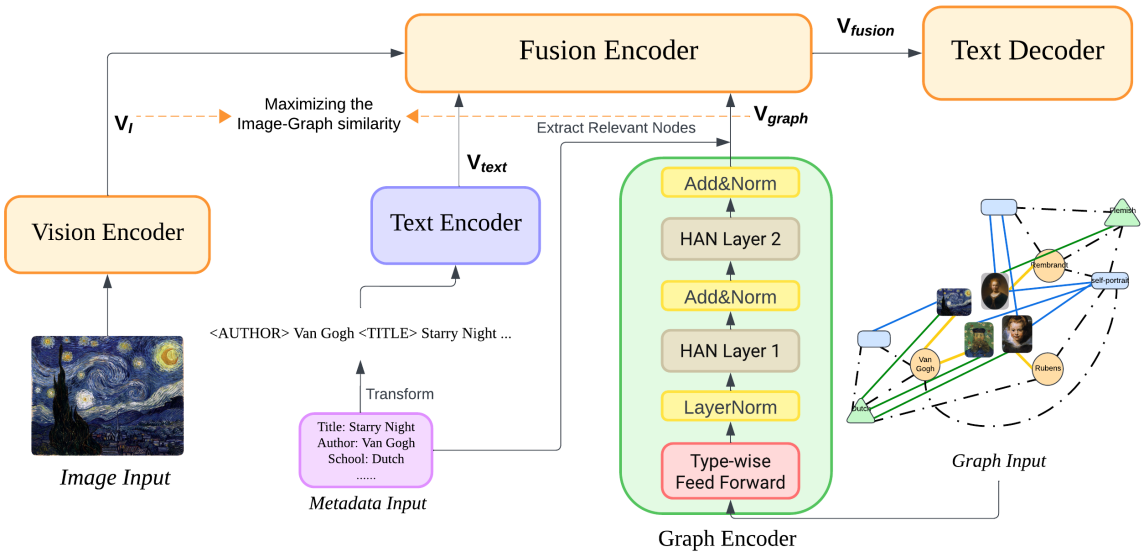

Figure 3: The architecture of our KALE model. There are five main components: (1) vision encoder; (2) text encoder; (3) graph encoder; (4) fusion encoder; and (5) text decoder. "Add&Norm" refers to residual connection [He *et al.*, 2016] and layer normalization [Ba *et al.*, 2016].

image embedding $\mathbf{v}_I$ from the vision encoder, we use a max pooling operation, which helps to reduce the dimensionality of the embeddings while preserving the most salient features. Formally,

$$\mathbf{v}_M = \text{FFN}(\mathbf{v}_{\text{graph}}) \tag{9}$$

$$\mathbf{v}_{I_{MP}} = \text{MaxPooling}(\mathbf{v}_I) \tag{10}$$

The loss function for this task can be represented as:

$$L_{CMA} = 1 - \frac{\mathbf{v}_{I_{MP}} \cdot \mathbf{v}_M}{\|\mathbf{v}_{I_{MP}}\| \|\mathbf{v}_M\|} \tag{11}$$

To combine the two loss functions, we introduce a balancing parameter β. The combined loss function can be expressed as:

$$L_{total} = (1 - \beta) \times L_{CE} + \beta \times L_{CMA} \tag{12}$$

4 Experiments and Results

4.1 Experiment Setup

Datasets and Competitors. Four artwork datasets are considered as benchmarks in this work, they are Artpedia, SemArt v1.0, SemArt v2.0 and ArtCap. Artpedia comprises a collection of 2,930 paintings from the 13th to the 21st century, each of which is associated with some textual descriptions and a corresponding title [Stefanini *et al.*, 2019]. SemArt dataset was first proposed by [Garcia and Vogiatzis, 2018] for cross-modal retrieval tasks, and contains European fine-art reproductions from the Web Gallery of Art. However, the descriptions of artworks are lengthy paragraphs, which do not align well with the image captioning task. To solve this, one recent study [Wu, 2022] separates each paragraph into single sentences, and labels them as visual and contextual sentences. The former describes the simple visual appearance of the artwork and the latter provides information about the painting's historical background. We call this as SemArt v1.0. [Bai *et al.*, 2021] proposed another more fine-grained way of categorising these sentences, based on their Form, Content and Context. Form deals with the visual composition, Content addresses the underlying meaning or subject matter, and Context provides the background of the artwork. We call this as SemArt v2.0 dataset. The last dataset ArtCap, contains 3605 paintings from WikiArt and the captions were manually collected by crowd-sourcing [Lu *et al.*, 2022].

We consider three previous state-of-the-art models as our baselines. [Wu, 2022] applies the Meshed-Memory transformer [Cornia *et al.*, 2020] to the artwork domain. We refer to this work as *Wu2022*. [Bai *et al.*, 2021] proposed a framework incorporating external knowledge by leveraging a knowledge retriever from Wikipedia and training a knowledge-filling module as a "fill in the blank" task to incorporate art information relevant to each painting. We refer to this work as *Bai2021*. For the last baseline, [Lu *et al.*, 2022] proposed a virtual-real semantic alignment training process through generating a virtual painting dataset via style transfer and training a painting feature extractor using the virtual dataset with a semantic alignment loss. We refer to this work as *Lu2022*.

Implementation details. For the vision encoder, fusion encoder and text decoder, we follow by default settings of mPLUG$_{large}$ model given in the open source code.[1]

For the heteregeneous graph construction, the cluster number of title embedding is set to 100, 1-gram of title is set to 2000, 2-gram is 1500 and 3-gram is 1000. For the optimizer, after parameter tuning on validation set, we use AdamW [Loshchilov and Hutter, 2018] optimizer with a weight decay of 0.02. The learning rate is first warmed up to 5e-5 for vision encoder, 1e-2 for graph encoder, and 1e-4 for other layers in the first 1000 iterations and decayed to 1e-5 following a cosine schedule. We use beam search decoding with beam width 5 and β is set to 0.2 to balance two loss functions.

[1]https://github.com/alibaba/AliceMind/tree/main/mPLUG

Dataset	Models	Evaluation Metrics							
		C	B-1	B-2	B-3	B-4	M	S	R
Artpedia	Wu2022	3.94	24.7	-	-	3.06	6.58	-	22.4
	KALE (w/o metadata)	11.7	29.9	15.0	7.95	4.77	8.02	5.49	22.4
	KALE (w/ metadata)	**23.4**	**32.6**	**17.7**	**10.9**	**7.48**	**9.33**	**7.68**	**23.7**
ArtCaps	Lu2022	15.21	52.89	33.11	20.42	12.57	15.38	8.39	36.15
	KALE (w/o metadata)	**36.51**	**59.78**	**41.24**	**28.56**	**20.14**	**19.39**	**12.06**	**42.22**
SemArt v1.0 (Visual)	Wu2022	6.93	19.20	-	-	3.24	6.28	-	21.90
	KALE (w/o metadata)	21.34	**29.58**	15.60	9.69	7.17	8.83	7.23	22.50
	KALE (w/ metadata)	**29.96**	28.68	**15.81**	**10.42**	**7.96**	**9.08**	**8.22**	**22.58**
SemArt v1.0 (Contextual)	Wu2022	-	-	-	-	-	-	-	-
	KALE (w/o metadata)	13.01	30.37	15.38	9.61	7.27	8.07	5.26	20.36
	KALE (w/ metadata)	**21.76**	**35.35**	**20.33**	**14.15**	**11.27**	**10.16**	**7.25**	**23.41**
SemArt v2.0	Bai2021	8.80	-	-	-	**9.10**	**11.4**	-	**23.1**
	KALE (w/o metadata)	13.60	25.85	13.75	8.78	6.70	7.48	5.97	19.86
	KALE (w/ metadata)	**20.7**	**27.7**	**15.7**	**10.8**	8.57	9.51	**7.31**	21.9

Table 1: Evaluation results on datasets Artpedia, ArtCaps, SemArt v1.0 and SemArt v2.0 over baselines. **Bold** representes the highest score. As we directly use the reported results for the competitors, some of the metrics are missing, we use "-" to indicate missing values.

4.2 Results and Analysis

Table 1 outlines the performance results on the four datasets Artpedia, ArtCaps, SemArt v1.0 and SemArt v2.0. We compare our model against three baseline models, Wu2022, Lu2022 and Bai2021. All models use the same train/validation/test split. Baseline numbers are values from the original publications.

Note that for SemArt v1.0 dataset, the author only conducts their experiments on visual sentences, so we break it into visual and contextual separately in the table. These models are evaluated over eight evaluation metrics, including CIDEr (C) [Vedantam *et al.*, 2015], BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), BLEU-4 (B-4) [Papineni *et al.*, 2002], METEOR (M) [Banerjee and Lavie, 2005], SPICE (S) [Anderson *et al.*, 2016] and ROUGE-L (R) [Lin, 2004].

We use "KALE (w/o metadata)" to denote the model without any metadata input, where the text and graph encoders are removed. In this case, it's an off-the-shelf pre-trained mPLUG fine-tuned on an artwork's dataset using standard generation cross-entropy objective.

On the other hand, "KALE (w/ metadata)" represents the model with both textual and graph representations. One exception is that for Artpedia dataset, the only metadata we have is artwork title, so we treat the artwork title as textual input and remove the graph encoder. As ArtCaps dataset lacks any metadata, we only use KALE without metadata approach for fine-tuning.

When we pre-process the SemArt v1.0 and v2.0 datasets, we notice some identical or highly similar image captions present in both the training set and validation/test set. Such duplication poses a risk of model overfitting, where the model might "memorize" the captions they have seen during training. Therefore, we remove such overlapping instances from the training dataset. Note that Wu2022 and Bai2021 did not do this, and so their reported performance may be an overestimate.

As evidenced by Table 1, our methodologies exhibit superiority over existing baselines. Overall, most of the best metric scores are achieved by our KALE model. By looking at the detailed scores for each metric, we find that CIDEr increases the most compared to baselines. For instance, our model achieves 23.4 CIDEr score on Artpedia dataset, which is over 5 times that of the competitor. On SemArt v2.0, our model also outperforms over 2 times than competitor, but for metrics like BlEU-4, METEOR and ROUGE, our model performs slightly worse than the competitor. Note that CIDEr offers insights into caption diversity, and this result shows that our model could generate more diverse captions. Later in our qualitative analysis, we will further verify this with some concrete examples. Looking at the performance of our KALE model without metadata, it still surpasses the Wu2022 model for Artpedia and SemArt v1.0 across all metrics. Such improvements demonstrate the effectiveness of pre-trained vision-language models in the art domain. The most important finding perhaps is the improvement of KALE leveraging metadata, which yields superior results against KALE without metadata across all metrics. This performance shows that metadata is a critical component for understanding artworks.

5 Discussion

5.1 Qualitative Analysis

Figure 4 depicts several random examples from the test set of our four benchmark datasets. In terms of content correctness, KALE is able to align its captions with the visual aspects of the images in most cases. For example, in image (d), KALE recognizes there is also a woman in the image. And interestingly, sometimes it can even recognize some elements that are not immediately present in the image. For instance, in image (g), KALE identifies the setting as a "French restaurant" (which is correct, based on the ground truth caption (GT). In terms of creativity, KALE's captions appear to be richer and more detailed than those of competitors. As illustrated

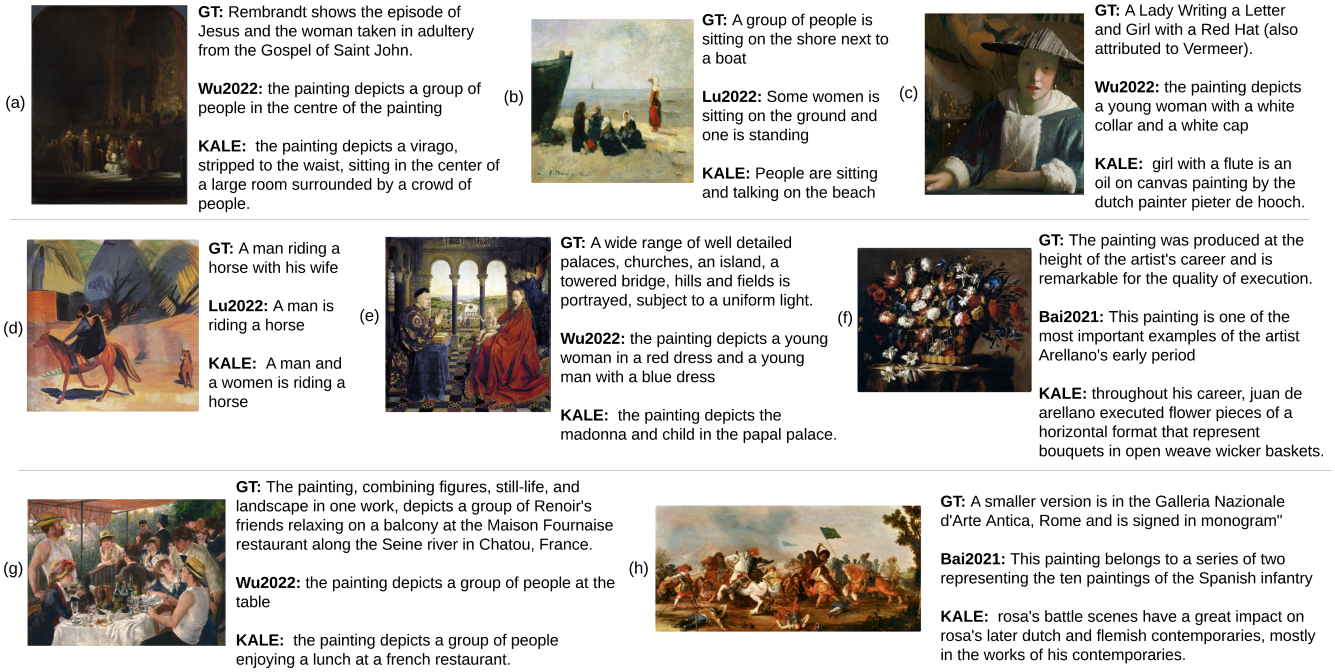

Figure 4: Some examples of generated captions by KALE, along with GT (ground truth) captions and competitor-generated captions.

Dataset	Model Variant	Evaluation Metrics			
		C	B-1	B-4	M
SemArt v1.0 (Visual)	KALE (w/o)	21.34	**29.58**	7.17	8.83
	KALE (T)	27.46	29.27	7.74	8.98
	KALE (T & G)	**29.96**	28.68	**7.96**	**9.08**
SemArt v1.0 (Contextual)	KALE (w/o)	13.01	30.37	7.27	8.07
	KALE (T)	20.35	34.50	10.78	9.86
	KALE (T & G)	**21.76**	**35.35**	**11.27**	**10.16**
SemArt v2.0	KALE (w/o)	13.60	25.85	6.70	7.48
	KALE (T)	12.35	**28.01**	6.03	7.89
	KALE (T & G)	**20.7**	27.7	**8.57**	**9.51**

Table 2: Performance impact of text and graph on KALE across SemArt v1.0 and v2.0 datasets. KALE (w/o) refers to the version without any metadata. KALE (T) refers to the version with only textual data input. KALE (T & G) refers to the version with both text and graph input.

in example (a), KALE portrays the scene as a "virago and stripped to the waist, surrounded by a crowd of people." In contrast, Wu2022 only describes it as a "group of people", which lacks substance and depth. However, the captions produced by KALE occasionally present inaccurate details about the artwork. For instance in example (c), KALE suggests that Pieter De Hooch is the artist of the painting rather than Vermeer. Overall, the integration of metadata appears to help generate higher quality captions, especially for images that have more background context. That said, compared to the ground truth captions there is arguably still quite a bit of gap, and so there is plenty of room for improvement and artwork interpretation is by no means a solved task.

5.2 Ablation Study

Impact of Text and Graph. This experiment assesses the comparative impact of using metadata as only textual input versus a combination of textual and graph inputs. As presented in Table 2, the text-only approach demonstrates a substantial improvement in performance over most of metrics compared to the version without metadata. This indicates the significant impact that textual metadata alone can make the model generate accurate and diverse captions. Further, when KALE was augmented with both textual and graph inputs, there was an additional enhancement in its performance, especially on metrics like CIDEr, BLEU-4 and METEOR, indicating the effectiveness of the knowledge graph. Interestingly, the knowledge graph integration showed more effectiveness on datasets like SemArt v1.0 Contextual and SemArt v2.0. These datasets are characterized by many contextual sentences that demand a deeper understanding of art, a requirement that the external knowledge provided by the graph is particularly well-suited to address.

6 Conclusion

In this work, we develop a novel artwork-specific image captioning system, KALE, that integrates external knowledge into the system through both text and heterogeneous graph. KALE is novel in that it captures the heterogeneity among images and several artwork attributes in the constructed graph. Results, both quantitative and qualitative, showed our method provides a better understanding of the narratives behind works of fine art.

References

[Achlioptas *et al.*, 2021] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021.

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *stat*, 1050:21, 2016.

[Bai *et al.*, 2021] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5422–5432, 2021.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[Cetinic, 2021] Eva Cetinic. Iconographic image captioning for artworks. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 502–516. Springer, 2021.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Garcia and Vogiatzis, 2018] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[Garcia *et al.*, 2020] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Contextnet: representation and exploration for painting classification and retrieval in context. *International Journal of Multimedia Information Retrieval*, 9:17–30, 2020.

[Gonthier *et al.*, 2018] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[He *et al.*, 2016] 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*, pages 770–778, 2016.

[Joulin *et al.*, 2016] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[Krishna *et al.*, 2017] 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.

[Li *et al.*, 2020] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[Li *et al.*, 2022] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[Loshchilov and Hutter, 2018] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[Lu *et al.*, 2022] Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 2022.

[Ma *et al.*, 2017] Daiqian Ma, Feng Gao, Yan Bai, Yihang Lou, Shiqi Wang, Tiejun Huang, and Ling-Yu Duan. From

part to whole: who is behind the painting? In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1174–1182, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[Sharma *et al.*, 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[Sheng and Moens, 2019] Shurong Sheng and Marie-Francine Moens. Generating captions for images of ancient artworks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2478–2486, 2019.

[Stefanini *et al.*, 2019] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20*, pages 729–740. Springer, 2019.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[Wang *et al.*, 2019] 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*, pages 2022–2032, 2019.

[Wang *et al.*, 2021] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[Wang *et al.*, 2022] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[Wu, 2022] Zixun Wu. Artwork interpretation. Master's thesis, University of Melbourne, 2022.

[Wynen *et al.*, 2018] Daan Wynen, Cordelia Schmid, and Julien Mairal. Unsupervised learning of artistic styles with archetypal style analysis. *Advances in Neural Information Processing Systems*, 31, 2018.

[Yang *et al.*, 2019] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019.

[Zhao and Wu, 2023] Wentian Zhao and Xinxiao Wu. Boosting entity-aware image captioning with multi-modal knowledge graph. *IEEE Transactions on Multimedia*, 2023.