ImageSet2Text: Describing Sets of Images Through Text

Piera Riccio[*1] [†], **Francesco Galati**[*2], **Kajetan Schweighofer**[3], **Noa Garcia**[4], **Nuria M Oliver**[5]
[1]University of Amsterdam
[2]Independent Researcher
[3]Johannes Kepler University Linz
[4]The University of Osaka
[5]ELLIS Alicante
p.riccio@uva.nl

Abstract

In the era of large-scale visual data, understanding collections of images is a challenging yet important task. To this end, we introduce `ImageSet2Text`, a novel method to automatically generate natural language descriptions of image sets. Based on large language models, visual-question answering chains, an external lexical graph, and CLIP-based verification, `ImageSet2Text` iteratively extracts key concepts from image subsets and organizes them into a structured concept graph. We conduct extensive experiments evaluating the quality of the generated descriptions in terms of accuracy, completeness, and user satisfaction. We also examine the method's behavior through ablation studies, scalability assessments, and failure analyses. Results demonstrate that `ImageSet2Text` combines data-driven AI and symbolic representations to reliably summarize large image collections for a wide range of applications.

Code — https://github.com/ellisalicante/ImageSet2Text
Datasets — https://github.com/ellisalicante/ImageSet2Text/tree/main/data
Extended version — https://arxiv.org/pdf/2503.19361

1 Introduction

The analysis of large-scale image sets is essential to uncover visual patterns that isolated samples fail to reveal (Deng et al. 2025). For example, describing a dataset of historical newspaper photographs can disclose stylistic trends over time or systematic biases (such as underrepresented demographics or recurring stereotypical depictions), which only become apparent when analyzing the images collectively. To make such analyses scalable to large image collections, automated tools are indispensable. While visualizations help identify broad patterns (Manovich 2012), automatically generated textual summaries can transform overwhelming visual collections into *interpretable* knowledge.

A variety of application domains would benefit from automatic summarization of image collections, including assistive technologies (Bigham et al. 2010; Gurari et al. 2020), cultural analytics (Shen, Efros, and Aubry 2019; Cetinic 2021; Manovich 2020), bias detection (Khosla et al. 2012), socio-economics (Jean et al. 2016; Dubey et al. 2016; Deng et al. 2025), exploratory data analysis (Boiman and Irani 2007), and training data transparency (Gebru et al. 2021) — whose need is intensified by emerging AI regulations (Parliament and the Council of the European Union 2024). In addition, in explainable AI, dataset-level insights have been found to be valuable for influential sample analysis and data segmentation (Park et al. 2023; Shah et al. 2023; Chung et al. 2019; d'Eon et al. 2022; Eyuboglu et al. 2022).

However, while large vision-language models have made significant progress in the last few years (Achiam et al. 2023; Bai et al. 2025), their application to describing large-scale image collections remains limited. Current approaches typically handle either individual images for captioning (Hossain et al. 2019; Vinyals et al. 2015; Xu 2015) or descriptions of small curated sets (Chen et al. 2018; Alayrac et al. 2022; Li et al. 2023a; Yao, Wang, and Jin 2022), failing to address the challenges of summarizing large visual datasets. This limitation remains unresolved due to fundamental technical challenges in processing multiple visual inputs at once (Dunlap et al. 2024; Deng et al. 2025).

In this paper, we propose `ImageSet2Text`, a method for generating natural language descriptions[1] (Pi et al. 2024) of large collection of images, as shown in Fig. 1. `ImageSet2Text` leverages multimodal large language models (LLMs) through an iterative visual question answering (VQA) process, which combines hypothesis formulation-verification with external knowledge to extract comprehensive insights from image collections. Inspired by recent concept bottleneck models (CBMs) (Koh et al. 2020; Tan, Zhou, and Chen 2024; Chattopadhyay, Chan, and Vidal 2024), `ImageSet2Text` extends this interpretable approach based on intermediate open-set concept prediction beyond classification tasks.

The descriptions generated by `ImageSet2Text` are obtained by means of an iterative process, consisting of two phases: (a) *Guess what is in the set*, and (b) *Look and keep*. To ensure scalability, a small subset of images is randomly selected in each iteration. In *Guess what is in the set*, an

[*]These authors contributed equally.
[†]Work conducted while at ELLIS Alicante, Spain.

[1]We refer to captions as "short pieces of text" (Cambridge Dictionary 2025), while descriptions are longer and more detailed.

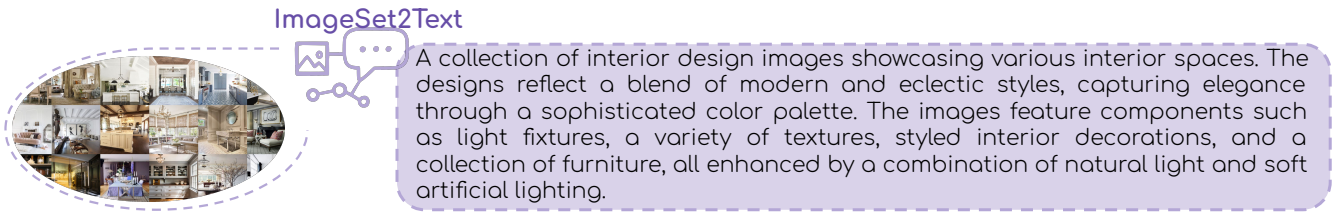

Figure 1: `ImageSet2Text` generates detailed and nuanced descriptions from large sets of images. We report an exemplary generated descriptions for a group of images (Sharma et al. 2018).

LLM-based VQA process identifies prevalent visual elements within the subset, and an external lexical graph is integrated to formulate hypotheses about the original set of images. In *Look and keep*, the hypotheses are verified using contrastive vision-language (CVL) embeddings (Radford et al. 2021) to assess their consistency across the entire image set. Verified hypotheses are joined into a *concept graph* and used to seed the next iteration. The process terminates when the concept graph cannot be updated anymore, at which point a final description of the entire image set is generated based on the accumulated information.

In extensive experiments (see Fig. 2), we evaluate `ImageSet2Text`´s descriptions according to their: 1) *accuracy*, via a large-scale group image captioning experiment; 2) *completeness*, by means of an image sets comparison task; and 3) *user satisfaction*, through a user study. We also evaluate its behavior via: 4) an *ablation study*, 5) a *scalability estimate*, and 6) an *analysis of failure cases*. Our results indicate that the generated descriptions accurately capture the visual content of large collections, offering rich detail and human-friendly readability, highlighting the potential of `ImageSet2Text` for diverse applications.

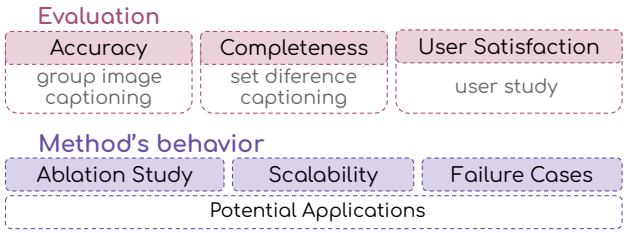

Figure 2: `ImageSet2Text`'s evaluation considers three key properties of the descriptions (accuracy, completeness and user satisfaction) and analyzes the method's behavior through ablations, scalability estimates and failure cases analysis, showing great versatility for potential applications.

2 Related Work

Image Captioning generates short textual descriptions of images by recognizing objects, attributes, scenes, and their relationships (Hossain et al. 2019). Early methods relied on deep learning for feature extraction (Vinyals et al. 2015; Xu 2015), while recent approaches leverage LLMs. For example, ChatCaptioner integrates VQA with chat logs for iteratively refinement (Zhu et al. 2023), and Mao et al. (2024) proposes the creation of context-aware, user-specific captions. The importance of context has been highlighted in image captioning for Art History, where different interpretations can lead to different descriptions (Bai, Nakashima, and Garcia 2021; Cetinic 2021; Lu et al. 2024).

Group-Image Captioning extends single-image captioning to small sets (typically $2 - 30$ images) by identifying shared patterns in the images (Chen et al. 2018; Alayrac et al. 2022; Li et al. 2023a; Yao, Wang, and Jin 2022). Proposed methods include modeling temporal relationships among images (Wang et al. 2019), analyzing pairwise differences (Chang and Ghamisi 2023; Kim et al. 2021; Park, Darrell, and Rohrbach 2019), and comparing target and reference groups of images (Li et al. 2020). Scene graphs have also been used to model and summarize the relationships between visual elements (Phueaksri et al. 2024, 2023), while LLMs appear to be promising for both individual and small-group captioning (Achiam et al. 2023). Benchmarks focus on spatial, semantic, and temporal aspects of small groups of images (Meng et al. 2024) or on evaluating large vision-language models in multi-image question answering (Liu et al. 2024). Despite the variety of approaches, the main limitation remains scaling to larger groups containing hundreds or thousands of images (Phueaksri et al. 2023).

Understanding Collections of Images remains challenging despite its importance in today's world of large-scale visual data. Existing approaches, such as textual PCA (Hupert, Schwartz, and Wolf 2022), concept-level prototypes (Doersch et al. 2015; Van Noord 2023), color-based statistical analysis (Torralba and Efros 2011), and set-level classification (Wang et al. 2022) fail to produce interpretable textual descriptions. Steps towards bridging this gap include set difference captioning (SDC) (Dunlap et al. 2024) and temporal change detection in urbanist images (Deng et al. 2025). In this paper, we advance beyond comparative approaches by proposing a novel method for generating comprehensive textual descriptions of large image collections.

Foundation Models are increasingly used for complex vision-language tasks. In addition to SDC (Dunlap et al. 2024), querying a VQA model through an LLM has been used to iteratively improve image and video captions (Chen et al. 2023; Zhu et al. 2023), to detect biases in text-to-image generation (D'Incà et al. 2024), and to evaluate text-to-image generation faithfulness (Hu et al. 2023). `ImageSet2Text` extends this paradigm by integrating multiple foundation models to describe image collections, contributing to emerging research (Deng et al. 2025).

3 ImageSet2Text

ImageSet2Text generates textual descriptions of image sets that highlight the common visual elements present in most of the images. As shown in Fig. 3, it leverages external prebuilt components, including an LLM, a VQA model, a lexical graph, and a CVL model, to construct an intermediate concept graph, which then serves as the basis for generating the description. Given a set of N images $\mathcal{D} = \{x_1, \ldots, x_N\}$, ImageSet2Text automatically generates a textual description d that summarizes the visual elements in $\mathcal{D}$. This is achieved by constructing an intermediate concept graph represented as a list of verified hypotheses $\mathcal{G}_c = \{h^1_\checkmark, ..., h^T_\checkmark\}$, where each $h^\tau_\checkmark$ is a triplet $\langle s, p, o\rangle$ identifying a subject s, a predicate p, and an object o that capture the key visual elements and their relations in $\mathcal{D}$. To build $\mathcal{G}_c$, ImageSet2Text follows an iterative process with T iterations depicted in Fig. 3, with the following steps:

1. **Initialization** ($\tau = 0$): It starts by defining the first subject $s :=$ '*image*' and a list of three candidate predicates, $P :=$ {'*content*', '*background*', '*style*'}. The initial $\mathcal{G}^0_c$ contains only '*image*' as root node.
2. **Iterations** ($\tau = 1, ..., T-1$), composed of two phases:
   (a) **Guess what is in the set** – A random subset of images $\mathcal{S} \subset \mathcal{D}$, with $|\mathcal{S}| = M \ll N$, is analyzed to hypothesize a full triplet $\langle s, p, o\rangle$ of elements present in $\mathcal{D}$, where M is a predefined parameter.
   (b) **Look and keep** – The formulated hypothesis is verified on $\mathcal{D}$. If confirmed, it is used to update $\mathcal{G}^\tau_c$.
3. **Termination** ($\tau = T$): After convergence at $\tau = T$, a coherent textual description d is generated from the final graph representation $\mathcal{G}_c = \mathcal{G}^T_c$.

Next, we describe the two phases of the iterations.

Guess What Is in the Set

The first phase generates a hypothesis from the random sample $\mathcal{S}$ of M images for later verification on the full set $\mathcal{D}$. Let τ denote the current time step. From the current graph $\mathcal{G}^\tau_c$, ImageSet2Text selects the first encountered closest leaf node to the root node as the subject s, along with a random candidate predicate $p \in P$ appropriate for s. Fig. 3 depicts an illustrative example where the image set $\mathcal{D}$ contains desert photos. At iteration $\tau = 5$, the subject selected is $s =$ '*desert*' and the predicate assigned is $p =$ '*color*'.

VQA. An LLM is prompted to ask a specific question about the images in $\mathcal{S}$ depending on the values of s and p. In the given example, with $s =$ '*desert*' and $p =$ '*color*', the resulting question is: *What colors can be observed in the desert landscape depicted in this image?*

The question is posed to a VQA model for each of the M images $x_i \in \mathcal{S}$, resulting in a set of answers $\mathcal{A} = \{a_1, a_2, \ldots, a_M\}$, where a_i denotes the answer for x_i.

Hypothesis Formulation. Given $\mathcal{A}$, each iteration verifies whether the predicate p can expand $\mathcal{G}^\tau_c$. To achieve this, the LLM is prompted to perform two tasks: **1) summarization**, *i.e.*, condense $\mathcal{A}$ into a single hypothesis h_0, which is formulated as a triplet where the subject s and predicate p are given, and the object o_0 has to be derived from $\mathcal{A}$; and **2) completion**, *i.e.*, suggest possible expansions of $\mathcal{G}^\tau_c$ from o_0 as a list of new predicates to append to P. Following our example, $\mathcal{A}$ would yield hypothesis h_0 and continuations P:

$$h_0 = \langle \text{'}\textit{desert}\text{'}, \text{'}\textit{color}\text{'}, \text{'}\textit{gold}\text{'} \rangle$$
$$P \leftarrow \{\text{'}\textit{shade}\text{'}, \text{'}\textit{brightness}\text{'}\}$$

Hypothesis Expansion. Since h_0 is derived from the subset $\mathcal{S}$, it may not generalize well to the full image set $\mathcal{D}$ due to sampling bias. To mitigate this, ImageSet2Text creates a set $\mathcal{H} = \{h_0, h_1, \ldots, h_k\}$ of hypotheses ordered from the most general (h_k) to the most specific (h_0), where $h_k \supset h_{k-1} \supset \cdots \supset h_0$. To generate $\mathcal{H}$ from h_0, ImageSet2Text relies on a given lexical graph $\mathcal{G}_l$. Let $\mathcal{G}_l = (V, R)$ be a directed graph, where V is the set of lexical entries and R represents semantic relations between nodes in V. For any given node $v \in V$ (*e.g.*, '*gold*'), its parent node denotes a more general concept, or hypernym (*e.g.*, '*yellow*'), and vice versa: '*gold*' is a hyponym of '*yellow*', *i.e.*, a more specific lexical concept. In addition, two nodes $v_1, v_2 \in V$ are sibling nodes if they share the same parent node, but correspond to different lexical concepts (*e.g.*, '*gold*' and '*gamboge*', both hyponyms of '*yellow*').

The set $\mathcal{H}$ is obtained by traversing upward in the knowledge hierarchy of $\mathcal{G}_l$ by a maximum number of steps δ. Given a hypothesis $h_i = \langle s, p, o_i\rangle$, its generalization h_{i+1} is created as $h_{i+1} = \langle s, p, o_{i+1}\rangle$, where $o_i = \text{parent}^i(o_0)$ and the parent function is the operation of moving to the hypernym of a lexical entry in $\mathcal{G}_l$. In the ongoing example, the hypotheses in $\mathcal{H}$ follow the hierarchy:

$$h_0 = \langle \text{'}\textit{desert}\text{'}, \text{'}\textit{color}\text{'}, \text{'}\textit{gold}\text{'} \rangle$$
$$h_1 = \langle \text{'}\textit{desert}\text{'}, \text{'}\textit{color}\text{'}, \text{'}\textit{yellow}\text{'} \rangle$$
$$h_2 = \langle \text{'}\textit{desert}\text{'}, \text{'}\textit{color}\text{'}, \text{'}\textit{chromatic color}\text{'} \rangle$$

Look and Keep

Next, all hypotheses $h_i \in \mathcal{H}$ are verified on the full image set $\mathcal{D}$ and the concept graph updated accordingly.

Verification. ImageSet2Text evaluates each $h_i \in \mathcal{H}$ against the entire image set $\mathcal{D}$ by leveraging the zero-shot classification capabilities of a CVL, following a one-vs-all classification problem, where positive and negative examples are generated for a given hypothesis h_i, drawing from $\mathcal{G}_l$. Let $\mathcal{H}^+_i$ denote the set of other hypotheses that support h_i, which are constructed by substituting the object o_i with its hyponyms in $\mathcal{G}_l$; and let $\mathcal{H}^-_i$ denote the set of hypotheses that contradict h_i, which are constructed by substituting o_i with its sibling nodes in $\mathcal{G}_l$. Note that the supporting set $\mathcal{H}^+_i$ is expanded to include h_i itself. In the given example, with $s =$'*desert*' and $p =$'*color*', for $h_1 = \langle s, p, \text{'}\textit{yellow}\text{'}\rangle$, supporting hypotheses replace '*yellow*' with its hyponyms o_h, while contradicting ones use sibling nodes o_s:

$$\mathcal{H}^+_1 = \{\langle s, p, o_h\rangle \mid \forall o_h \in \{\text{'}\textit{gold}\text{'}, \text{'}\textit{gamboge}\text{'}, ...\}\}$$
$$\mathcal{H}^-_1 = \{\langle s, p, o_s\rangle \mid \forall o_s \in \{\text{'}\textit{red}\text{'}, \text{'}\textit{orange}\text{'}, ...\}\}$$

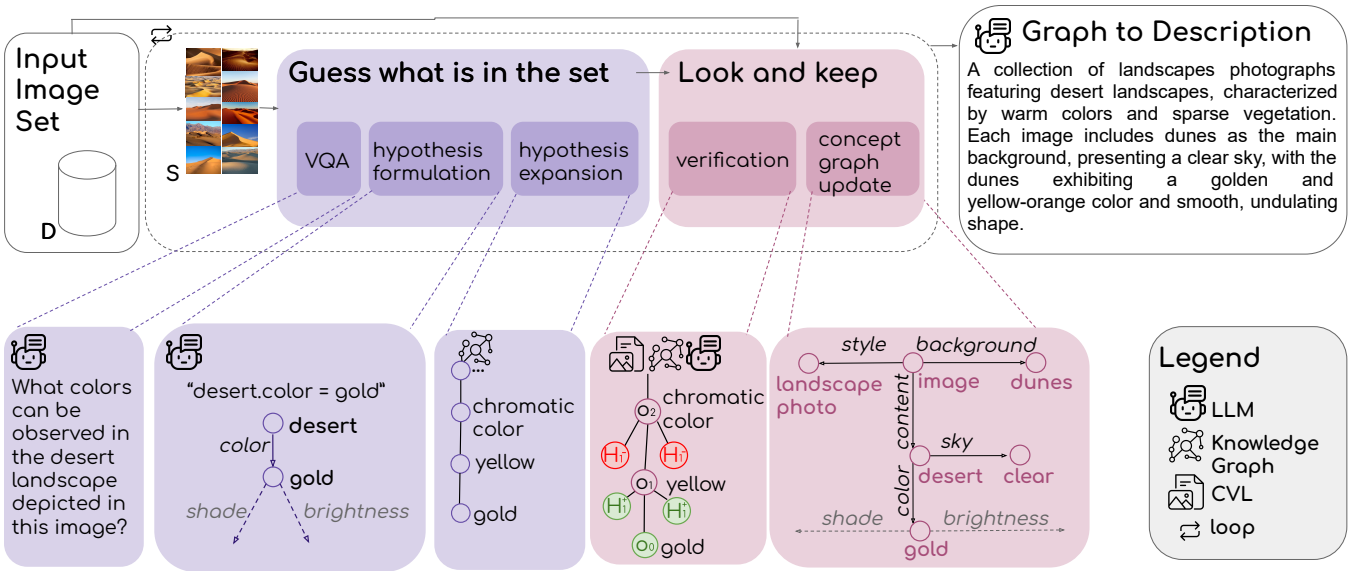

Figure 3: Overview of `ImageSet2Text`, considering an example set from the PairedImageSets datasets (Dunlap et al. 2024). The figure shows how the different modules of the iterative process allow inferring information from the input image set, eventually generating a nuanced textual description.

Next, all images in $\mathcal{D}$, along with each o_h and o_s are projected one-by-one into the CVL latent space and L2-normalized, yielding the sets of embeddings $\mathcal{E}_{\mathcal{D}}$, $\mathcal{E}_{\mathcal{H}_i^+}$, and $\mathcal{E}_{\mathcal{H}_i^-}$, respectively. A weighted k-Nearest Neighbors (kNN) classifier is then applied using the embeddings from $\mathcal{E}_{\mathcal{H}_i^+}$ as positive examples and those from $\mathcal{E}_{\mathcal{H}_i^-}$ as negative examples, with cosine similarity serving as the weighting metric. In particular, the kNN classifies each image $x_j \in \mathcal{D}$ as supporting the hypothesis h_i if its corresponding embedding $e_j \in \mathcal{E}_{\mathcal{D}}$ is labeled as positive, and as contradicting h_i otherwise. As a result, the hypothesis h_i is rejected if it is not verified on at least a predefined minimum portion α of the images in $\mathcal{D}$. Since hypotheses follow a hierarchical structure, if a hypothesis h_i is rejected, then any more specific hypothesis $h_{i-1} \subset h_i$ is also unverified. This follows from the logical implication that $h_i \Rightarrow h_{i+1}$.

The hypotheses are verified in a general-to-specific manner to ensure semantic consistency in the CVL embedding space. If a highly specific hypothesis is prematurely tested without confirming its general category first, there is a risk of making comparisons in an embedding subspace that is not reliable (Radford et al. 2021). Similarly, the set $\mathcal{H}_i^-$ is used because computing cosine similarities solely within $\mathcal{H}_i^+$ would not provide a reliable basis for evaluating the validity of a hypothesis. Comparisons against negative examples are necessary to establish a meaningful criterion for hypothesis acceptance (Chattopadhyay, Chan, and Vidal 2024).

Concept Graph Update. At the end of the verification process, let $h_\checkmark = \langle s, p, o_\checkmark \rangle$ be the most specific hypothesis $h_i \in \mathcal{H}$ that is verified; $h_\checkmark$ is appended to the graph $\mathcal{G}_c^{\tau+1}$ and the list of still-pending predicates P from o_0 is carried over to the next iteration (as illustrated in Fig. 3). The object $o_\checkmark$, now a leaf node in $\mathcal{G}_c^{\tau+1}$, is thus eligible to become the subject s in a next iteration.

Stopping Conditions

An iteration in `ImageSet2Text` interrupts if any of the following conditions occurs: **1)** the VQA module flags a question as invalid (*e.g.*, unsafe, inappropriate, unrelated to the content of the image) for at least a predefined number of images θ in $\mathcal{S}$; **2)** no hypothesis in $\mathcal{H}$ is verified for $\mathcal{D}$; **3)** the updated graph $\mathcal{G}_c^{\tau+1}$ adds no new information when compared to $\mathcal{G}_c^{\tau}$ as per an LLM evaluation. The entire iterative process ends when: **1)** no further graph expansion is possible, *i.e.*, all existing nodes in $\mathcal{G}_c^{\tau}$ have been explored; or **2)** a certain number ϵ of consecutive iterations are discarded according to the previously mentioned criteria. Once the iterative process ends, any pending predicate in P is discarded, and the final textual description d is generated directly from $\mathcal{G}_c = \mathcal{G}_c^T$ using the LLM, as illustrated in Fig. 3.

4 Evaluation

We report the results of evaluating the generated descriptions (qualitative examples available in App. B.2) according to three properties, as depicted in Fig. 2.

Implementation Details. We use GPT-4o-mini (Achiam et al. 2023) as the LLM (including for VQA), Open-CLIP ViT-bigG-14 (Ilharco et al. 2021) as the CVL, and WordNet (Miller 1995) as the external lexical graph. Each random subset $\mathcal{S}$ contains $M = 10$ images; hypotheses are rejected if their verification rate falls below $\alpha = 0.8$ with $k = 1$; iterations terminate upon encountering $\theta = 10$ invalid images; traversing the knowledge hierarchy is limited to $\delta = 2$ steps; and the maximum consecutive discarded iterations for stopping is $\epsilon = 5$. Further details are provided in App. A.

Accuracy

Accuracy is evaluated through the task of group image captioning. This task tests whether models can identify and describe common visual elements in a group of images, making it an ideal proxy for this property of the descriptions.

Datasets. We did not find publicly available benchmarks for group image captioning of either small (up to 30 images) or large (up to thousands of images) image sets. Hence, we created two new benchmark datasets for this part of the evaluation (details in App. B.1):

- **GroupConceptualCaptions**: Images with the same caption from the Conceptual Captions dataset (Sharma et al. 2018) are grouped. Each group's caption serves as ground truth, providing concise single-sentence descriptions of the main visual content. This dataset contains 116 groups with a total of $23,412$ images.
- **GroupWikiArt**: WikiArt artworks (Tan et al. 2019) with identical style, genre, and artist, are grouped, and the group captions are derived from these attributes. This dataset allows to evaluate nuanced, abstract concepts emerging from shared artistic interpretations within groups, and requires examining multiple images to capture group similarities. The dataset contains 105 groups and $53,707$ images.

Models. Given the lack of public models for group image captioning, we compare `ImageSet2Text` with four state-of-the-art vision-language models: BLIP-2 (Li et al. 2023b), LLaVA-1.5 (Liu et al. 2023), GPT-4o (Achiam et al. 2023), and Qwen2.5-VL (Bai et al. 2025). As these are designed for single-image captioning, we adapt them to group captioning through three strategies: (1) prompting image grids of varying sizes in a single image, (2) averaging image embeddings before caption generation (for open-source models), and (3) summarizing individual captions into a group caption using GPT-4o. Further details are available in App. B.2 and App. B.7. Tests on multi-image settings are reported in App. B.6. For `ImageSet2Text`, we generate captions by summarizing the full description into a single short sentence using GPT-4o, as detailed in App. B.3.

Metrics. Accuracy is quantified by measuring the semantic alignment between generated descriptions and ground-truth captions on both GroupConceptualCaptions and GroupWikiArt. Results are reported as the average rank across seven standard captioning metrics: four model-free,[2] two model-based,[3] and one reference-free.[4] Full evaluation details can be found in App. B.4 and B.5.

Results. As shown in Fig. 4, `ImageSet2Text` achieves the **best** performance on both GroupConceptualCaptions and GroupWikiArt, with average ranks of 2.50 and 7.57, respectively, outperforming all baselines. In contrast, the performance of the baselines is not consistent across datasets:

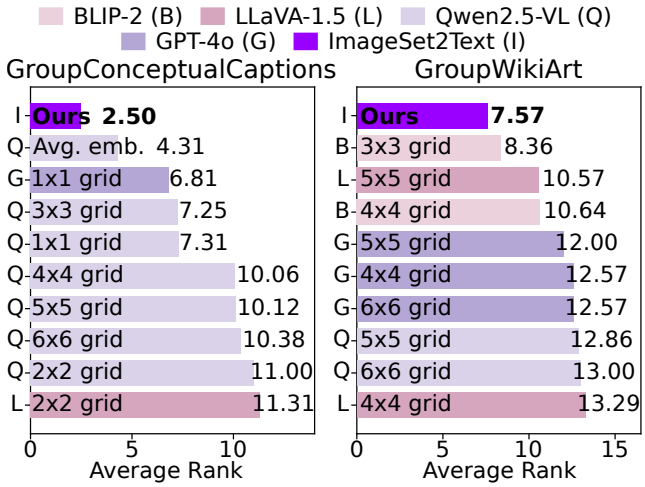

Figure 4: Accuracy as average rank across seven metrics on GroupConceptualCaptions (left) and GroupWikiArt (right). Only top ten methods shown; full results in App. B.5.

while Qwen2.5-VL is the second-best model in GroupConceptualCaptions, BLIP-2, LLaVA-1.5, and GPT-4o exhibit stronger results in GroupWikiArt. Metric-wise analysis (detailed in App. B.5) reveals that `ImageSet2Text` performs particularly strong on model-based and reference-free metrics, which better capture semantic content and align more closely with human judgment (Zhang et al. 2020).

Completeness

Completeness is evaluated by means of the Set Difference Captioning (SDC) task, which consists of identifying differences between two image sets, $\mathcal{D}_A$ and $\mathcal{D}_B$. To perform this comparison, we rely on the concept graphs generated by `ImageSet2Text` for each set. If the concepts are incomplete or vague, they would lead to failures in detecting differences between the sets. Hence, SDC serves as an effective proxy for evaluating the completeness of the underlying information used to generate the final descriptions—and, by extension, the completeness of the descriptions themselves.

Dataset. We adopt the methodology and dataset (PIS) described in (Dunlap et al. 2024). The PIS dataset consists of 150 contrastive set pairs (with $30,000$ images in total) spanning easy, medium, and hard distinctions. Consistent success in this task across varying difficult levels indicates that the description captures a range of relevant details necessary to differentiate $\mathcal{D}_A$ from $\mathcal{D}_B$.

Models. The PIS dataset was introduced together with VisDiff (Dunlap et al. 2024), a proposer-ranker framework in which an LLM-based proposer suggests potential differences between image sets, and a ranker evaluates and ranks them using CLIP embeddings. The proposer relies on captions created with BLIP-2 from subsets of the two original datasets to identify differences between the two image sets. We compare VisDiff with `ImageSet2Text` by replacing the BLIP-2 captions with the concept graph representations produced by `ImageSet2Text`, keeping the rest

[2]CIDEr-D, SPICE, METEOR, ROUGE-L.
[3]BERTScore F1, LLM-as-a-judge.
[4]CLIPScore.

Method	Category	Acc@1	Acc@5
VisDiff	Easy	0.88	**0.99**
	Medium	0.75	0.86
	Hard	0.61	0.80
`ImageSet2Text`	Easy	**0.90**	**0.99**
	Medium	**0.77**	**0.89**
	Hard	**0.66**	**0.82**

Table 1: Completeness evaluation on the PIS dataset. The best performance per difficulty category is shown in bold.

of the framework unchanged (details in App. C.1 and C.3).

Metrics. We use standard metrics: acc@1 and acc@5.

Results. Results are shown in Table 1. Using `ImageSet2Text` improves the performance with respect to the original VisDiff in all metrics except for acc@5 on the easy sets, where both methods achieve near-perfect accuracy (0.99). Importantly, `ImageSet2Text` consistently improves accuracy on the medium and hard sets, showing that its richer semantic representations enable better identification of subtle differences between the image sets (*e.g.*, distinguishing between cars with metallic paint and matte paint). Indeed, the lack of details in the BLIP-2 captions lead to a set of failure cases in VisDiff that are mostly addressed with `ImageSet2Text` (see App. C.2).

User Satisfaction

While the previous experiments measure accuracy and completeness, they do not assess the overall quality of the descriptions, which we evaluate in a user study conducted on a random subset of the PIS dataset with 233 users.

Methodology. We sample 60 PIS image sets (20 per difficulty level) and, for each, we show participants 16 images in a 4×4 grid alongside a description (see App. D). Participants are asked to rate the *clarity, accuracy, detail, flow,* and *overall satisfaction* of the descriptions on a five 5-point Likert-scale. To ease the interpretation of the user feedback and given that there is a lack of alternative methods to create descriptions of large image sets, we generate 10 control descriptions using ChatGPT (see App. D for examples), divided into three categories:

- **Control accuracy** (3 descriptions): well-written but intentionally inaccurate descriptions.
- **Control detail** (3 descriptions): descriptions that reference the correct visual content but lack details.
- **Control clarity and flow** (4 descriptions): factually correct, detailed but with low coherence descriptions.

Note that comparing against the captions described in Section 4 would be inappropriate due to length and detail mismatches that could bias results (Grice 1975; Kahneman 2011). Also note that the control descriptions are not performance baselines: they isolate specific qualities, allowing for controlled comparisons in alignment with human-centered evaluation best practices (Rogers 2023).

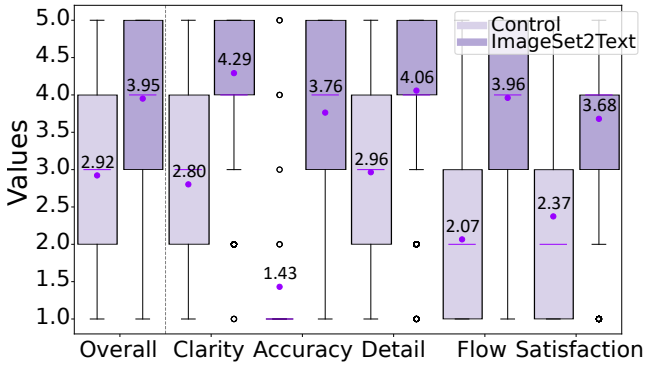

Figure 5: User study results. For control, values are of those designed to assess clarity, accuracy, detail, and flow, respectively, whereas overall and satisfaction are averaged across all control descriptions. Purple bars indicate medians, dots show means with the actual values reported in the figure.

A total of 233 qualifying participants[5] were recruited via Prolific, each evaluating 7 descriptions (6 `ImageSet2Text`, 1 control) and performing 2 attention tests. 198 participants successfully completed the study, yielding $16 - 22$ evaluations per description. The average task completion time was 8 minutes, compensated at $12/hour. The process was fully anonymized, and no personal information was collected.

Results. Results are shown in Fig. 5. `ImageSet2Text` descriptions are consistently rated favorably, not only in absolute terms but also when compared to the reference levels established by the control descriptions: clarity ($\mu = 4.29$ vs control $\mu = 2.80$), accuracy ($\mu = 3.76$ vs control $\mu = 1.43$), detail ($\mu = 4.06$ vs control $\mu = 2.96$) and flow ($\mu = 3.96$ vs control $\mu = 2.07$). These differences are statistically significant (t-test, all $p \ll 10^{-5}$).

5 Method's Behavior

We report analyses on `ImageSet2Text`'s behavior, considering an ablation study, scalability estimate and failure cases analysis (Fig. 2) and discussing potential applications.

Ablation Study. Inspired by ongoing research that combines symbolic and data-driven AI (Marcus 2018; Guo et al. 2024), we conducted an ablation study with four versions of `ImageSet2Text`, each progressively integrating more structured information:

`v1` relies only on LLMs to generate hypothesis h_0, the set of general hypotheses $\mathcal{H}$, and the supporting $\mathcal{H}_i^+$ and contradicting $\mathcal{H}_i^-$ alternatives. The extracted insights are directly used to refine the description iteratively;

`v2` introduces the external lexical graph $\mathcal{G}_l$ to generate the sets $\mathcal{H}$, $\mathcal{H}_i^+$ and $\mathcal{H}_i^-$. No concept graph is kept in memory and the textual description is updated at every iteration;

[5] 18+ native English-speakers w/o visual/reading impairments.

	$\mathcal{G}_l$	$\mathcal{G}_c^\tau$	dependency parsing	Acc@1	Acc@5
v1	-	-	-	0.67	0.87
v2	✓	-	-	0.77	0.87
v3*	✓	✓	-	**0.90**	**1.00**
v4	✓	✓	✓	0.67	0.87

Table 2: Ablation study with incremental structured knowledge representation on a subset of the PIS dataset (* indicates the version corresponding to ImageSet2Text).

v3 introduces the iterative concept graph $\mathcal{G}_c^\tau$, with the final description being generated from $\mathcal{G}_c^T$. This is the version of ImageSet2Text introduced in this paper;

v4 decomposes hypothesis formulation into two steps: (1) the LLM to summarize the VQA answers $\mathcal{A}$ into a sentence, and (2) POS tagging and dependency parsing to extract the object o_0 and its candidate predicates.

We report completeness (as in Section 4) on a random subset of 15 image set pairs (5 easy, 5 medium, and 5 hard). Table 2 shows that the progressive integration of structured information improves the performance up to v3, followed by a decline in v4. This result is confirmed by manual assessment, where v3 produced the highest quality descriptions. The generation of $\mathcal{H}$, $\mathcal{H}_i^+$, and $\mathcal{H}_i^-$ in v2 is more effective compared to v1, which is subject to hallucinations due to relying only on the LLM. However, both v2 and v1 occasionally produce descriptions with a broken flow. This limitation is addressed in v3 by directly generating the final description from the concept graph. POS tagging and dependency parsing in v4 are error-prone and hard to adapt, making them less reliable than the LLM alone for hypothesis formulation. In conclusion, the best-performing version, v3, is the version that best leverages the advantages of both symbolic and data-centric AI.

Scalability. ImageSet2Text scales efficiently to large image sets of size N. Hypotheses are generated from a small random sample $M \ll N$, keeping computational cost independent of N (with M depending on set heterogeneity rather than size). For verification, 1280-dimensional embeddings (*i.e.*, $d = 1280$) for all N images are precomputed *once* on the CVL model at ≈ 12 images/s on an NVIDIA RTX 3090, requiring $T_{embed}(N) = N/12$ seconds (TACIXAT 2023). The kNN step, repeated over T iterations, compares N images with $S+C$ example points, taking $T_{iter}(N, |\mathcal{H}_i^+|, |\mathcal{H}_i^-|) = N \cdot (|\mathcal{H}_i^+| + |\mathcal{H}_i^-|) \cdot d \cdot 2/\text{FLOPs}$ seconds, where FLOPs denotes the GPU's floating-point throughput. The factor of 2 accounts for one multiplication and one addition per dimension when computing cosine distance on L2-normalized embeddings. For example, using FP32 precision on an NVIDIA RTX 3090 (35.58 TFLOPs), with 1 million images and $|\mathcal{H}_i^+| + |\mathcal{H}_i^-| = 1000$, each kNN iteration takes < 0.1 s, which is ImageSet2Text's only N-dependent step. Across our experiments, T ranges from 10 to 30, independently of N. In addition, avg graphs #nodes (depth) are: 10.84 (3.28) in GroupConceptualCaptions, 9.83 (3.13) in GroupWikiArt; in PIS: 10.84 (3.31) easy, 10.48 (3.48) medium, 10.37 (3.57) hard, indicating fewer nodes but similar/higher depth due to higher visual diversity.

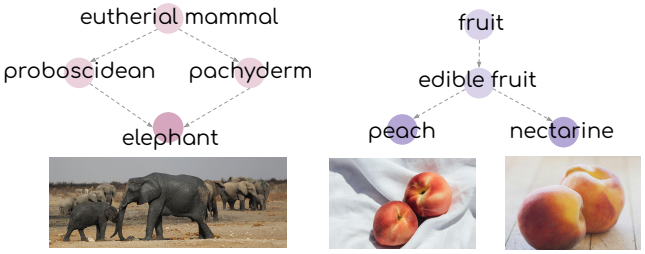

Figure 6: Illustrative examples of failure cases: (1) sibling nodes that are not mutually exclusive (*pachyderm* vs. *proboscidean*), causing incorrect formulation of $\mathcal{H}_i^-$; and (2) sibling nodes that cannot be visually distinguished (*peach* vs. *nectarine*), causing kNN misclassifications.

Failure Cases. We identify two main failure types (see Fig. 6) due to the integration of WordNet and CLIP: (1) sibling nodes that are not mutually exclusive cause $\mathcal{H}_i^-$ elements to be supportive (in 1.49% of cases); (2) sibling nodes that are hard to distinguish visually cause kNN misclassifications (in around 2% of cases). While these issues affect the verification process, they are sufficiently rare to have minimal overall impact. Details on how we estimate their incidence are provided in App. E, along with other limitations inherited from the modules constituting ImageSet2Text.

Potential Applications. The high readability of the descriptions generated by ImageSet2Text suggests broader applications beyond the already demonstrated possibilities of the SDC task (Dunlap et al. 2024) (Section 4.2), such as dataset exploration under user-guidance, explainable AI, cultural analytics, and the identification of potential biases in image sets. Additionally, we initiated a collaboration with Fundación ONCE[6] to collect feedback on ImageSet2Text's value for visually-impaired individuals, incorporating community insights early in the design process (Costanza-Chock 2020). Details in App. F.

6 Conclusion

In this paper, we have proposed ImageSet2Text, a novel method to generate natural language descriptions of image sets. We have shown its competitive performance in: (1) a large-scale group captioning experiment with two newly proposed benchmarks (GroupConceptualCaptions and GroupWikiArt); (2) set difference captioning to assess description completeness; and (3) a user study. Through an ablation study, a scalability analysis, and failure case examination, we have illustrated how ImageSet2Text successfully integrates symbolic and data-centric approaches. Overall, ImageSet2Text effectively describes large image collections, proving valuable for diverse applications.

[6]Spanish national association for universal accessibility.

Acknowledgements

We are grateful to Fundación ONCE for their willingness to collaborate on our research. **PR** and **NO** have been partially supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Resolución de la Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación). **PR** was also supported by a grant by Fundación Banc Sabadell. **NG** is partly supported by JSPS KAKENHI No. JP22K12091. **KS** is part of the ELLIS Unit Linz, the LIT AI Lab and the Institute for Machine Learning at Johannes Kepler University Linz, which are supported by the Federal State Upper Austria. We thank the projects FWF AIRI FG 9-N (10.55776/FG9), AI4GreenHeatingGrids (FFG- 899943), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01), FWF Bilateral Artificial Intelligence (10.55776/COE12). We thank NXAI GmbH, Audi AG, Silicon Austria Labs (SAL), Merck Healthcare KGaA, GLS (Univ. Waterloo), TÜV Holding GmbH, Software Competence Center Hagenberg GmbH, dSPACE GmbH, TRUMPF SE + Co. KG. We thank **Federico Brunero**, **Julien Colin**, **Erik Derner**, **Aditya Gulati**, **Benedikt Höltgen**, **Fabian Paischer** and **Korbinian Pöppel** for helpful discussions.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv*, 2502.13923.

Bai, Z.; Nakashima, Y.; and Garcia, N. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *ICCV*, 5422–5432.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342.

Boiman, O.; and Irani, M. 2007. Detecting irregularities in images and in video. *IJCV*, 74(1): 17–31.

Cambridge Dictionary. 2025. Definition of "caption". Accessed: February 12, 2025.

Cetinic, E. 2021. Towards Generating and Evaluating Iconographic Image Captions of Artworks. *Journal of Imaging*, 7: 123.

Chang, S.; and Ghamisi, P. 2023. Changes to captions: An attentive network for remote sensing change captioning. *IEEE Transactions on Image Processing*.

Chattopadhyay, A.; Chan, K. H. R.; and Vidal, R. 2024. Bootstrapping variational information pursuit with large language and vision models for interpretable image classification. In *ICLR*.

Chen, F.; Ji, R.; Sun, X.; Wu, Y.; and Su, J. 2018. GroupCap: Group-Based Image Captioning With Structured Relevance and Diversity Constraints. In *CVPR*.

Chen, J.; Zhu, D.; Haydarov, K.; Li, X.; and Elhoseiny, M. 2023. Video ChatCaptioner: Towards Enriched Spatiotemporal Descriptions. *arXiv:2304.04227*.

Chung, Y.; Kraska, T.; Polyzotis, N.; Tae, K. H.; and Whang, S. E. 2019. Automated data slicing for model validation: A big data-ai integration approach. *IEEE Transactions on Knowledge and Data Engineering*, 32(12): 2284–2296.

Costanza-Chock, S. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.

Deng, B.; Peng, S.; Genova, K.; Wetzstein, G.; Snavely, N.; Guibas, L.; and Funkhouser, T. 2025. Visual Chronicles: Using Multimodal LLMs to Analyze Massive Collections of Images. *arXiv:2504.08727*.

d'Eon, G.; d'Eon, J.; Wright, J. R.; and Leyton-Brown, K. 2022. The spotlight: A general method for discovering systematic errors in deep learning models. In *ACM FAccT*, 1962–1981.

D'Incà, M.; Peruzzo, E.; Mancini, M.; Xu, D.; Goel, V.; Xu, X.; Wang, Z.; Shi, H.; and Sebe, N. 2024. OpenBias: Open-set Bias Detection in Text-to-Image Generative Models. In *CVPR*, 12225–12235.

Doersch, C.; Singh, S.; Gupta, A.; Sivic, J.; and Efros, A. A. 2015. What makes paris look like paris? *Communications of the ACM*, 58(12): 103–110.

Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; and Hidalgo, C. A. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *ECCV*, 196–212. Springer.

Dunlap, L.; Zhang, Y.; Wang, X.; Zhong, R.; Darrell, T.; Steinhardt, J.; Gonzalez, J. E.; and Yeung-Levy, S. 2024. Describing differences in image sets with natural language. In *CVPR*, 24199–24208.

Eyuboglu, S.; Varma, M.; Saab, K.; Delbrouck, J.-B.; Lee-Messer, C.; Dunnmon, J.; Zou, J.; and Ré, C. 2022. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv:2203.14960*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grice, H. P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Guo, Y.; Bo, D.; Yang, C.; Lu, Z.; Zhang, Z.; Liu, J.; Peng, Y.; and Shi, C. 2024. Data-centric graph learning: A survey. *IEEE Transactions on Big Data*.

Gurari, D.; Zhao, Y.; Zhang, M.; and Bhattacharya, N. 2020. Captioning images taken by people who are blind. In *ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 417–434. Springer.

Hossain, M. Z.; Sohel, F.; Shiratuddin, M. F.; and Laga, H. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6): 1–36.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 20406–20417.

Hupert, O.; Schwartz, I.; and Wolf, L. 2022. Describing Sets of Images with Textual-PCA. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the ACL: EMNLP 2022*, 3811–3821. Abu Dhabi, United Arab Emirates: ACL.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP.

Jean, N.; Burke, M.; Xie, M.; Alampay Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301): 790–794.

Kahneman, D. 2011. *Thinking, fast and slow*. macmillan.

Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the damage of dataset bias. In *ECCV*, 158–171. Springer.

Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic change captioning with cycle consistency. In *ICCV*, 2095–2104.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *ICML*, 5338–5348. PMLR.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Yang, J.; Li, C.; and Liu, Z. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv:2306.05425*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.

Li, Z.; Tran, Q.; Mai, L.; Lin, Z.; and Yuille, A. L. 2020. Context-aware group captioning via self-attention and contrastive features. In *CVPR*, 3440–3450.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved Baselines with Visual Instruction Tuning.

Liu, Z.; Chu, T.; Zang, Y.; Wei, X.; Dong, X.; Zhang, P.; Liang, Z.; Xiong, Y.; Qiao, Y.; Lin, D.; et al. 2024. MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs. *arXiv:2406.11833*.

Lu, Y.; Guo, C.; Dai, X.; and Wang, F.-Y. 2024. ArtCap: A Dataset for Image Captioning of Fine Art Paintings. *IEEE Transactions on Computational Social Systems*, 11(1): 576–587.

Manovich, L. 2012. How to compare one million images? In *Understanding digital humanities*, 249–278. Springer.

Manovich, L. 2020. *Cultural analytics*. Mit Press.

Mao, S.; Zhang, C.; Su, H.; Song, H.; Shalyminov, I.; and Cai, W. 2024. Controllable Contextualized Image Captioning: Directing the Visual Narrative through User-Defined Highlights. *arXiv:2407.11449*.

Marcus, G. 2018. Deep Learning: A Critical Appraisal. *arXiv*, 1801.00631.

Meng, F.; Wang, J.; Li, C.; Lu, Q.; Tian, H.; Liao, J.; Zhu, X.; Dai, J.; Qiao, Y.; Luo, P.; et al. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv:2408.02718*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *ICCV*, 4624–4633.

Park, S. M.; Georgiev, K.; Ilyas, A.; Leclerc, G.; and Madry, A. 2023. Trak: Attributing model behavior at scale. *arXiv:2303.14186*.

Parliament, E.; and the Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts. https://eur-lex.europa.eu/eli/reg/2024/1689. Official Journal of the European Union, L 168, 12 July 2024.

Phueaksri, I.; Kastner, M. A.; Kawanishi, Y.; Komamizu, T.; and Ide, I. 2023. Towards captioning an image collection from a combined scene graph representation approach. In *MMM*, 178–190. Springer.

Phueaksri, I.; Kastner, M. A.; Kawanishi, Y.; Komamizu, T.; and Ide, I. 2024. Image-Collection Summarization using Scene-Graph Generation with External Knowledge. *IEEE Access*.

Pi, R.; Zhang, J.; Zhang, J.; Pan, R.; Chen, Z.; and Zhang, T. 2024. Image Textualization: An Automatic Framework for Generating Rich and Detailed Image Descriptions. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *NeurIPS*, volume 37, 108116–108139. Curran Associates, Inc.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Rogers, Y. 2023. Interaction Design: Beyond Human-Computer Interaction.

Shah, H.; Park, S. M.; Ilyas, A.; and Madry, A. 2023. Modeldiff: A framework for comparing learning algorithms. In *ICML*, 30646–30688. PMLR.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of ACL*.

Shen, X.; Efros, A. A.; and Aubry, M. 2019. Discovering visual patterns in art collections with spatially-consistent feature learning. In *CVPR*, 9278–9287.

TACIXAT. 2023. OpenCLIP Throughput Benchmark (image embeddings per second). https://gist.github.com/TACIXAT/ecd4f636bf6af28cb69d641e29d7b362. Accessed: 2025-07-28.

Tan, A.; Zhou, F.; and Chen, H. 2024. Explain via any concept: Concept bottleneck model with open vocabulary concepts. *arXiv:2408.02265*.

Tan, W. R.; Chan, C. S.; Aguirre, H.; and Tanaka, K. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Van Noord, N. 2023. Prototype-based dataset comparison. In *ICCV*, 1944–1954.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164.

Wang, A.; Liu, A.; Zhang, R.; Kleiman, A.; Kim, L.; Zhao, D.; Shirai, I.; Narayanan, A.; and Russakovsky, O. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *IJCV*, 130(7): 1790–1810.

Wang, B.; Ma, L.; Zhang, W.; Jiang, W.; and Zhang, F. 2019. Hierarchical photo-scene encoder for album storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8909–8916.

Xu, K. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044*.

Yao, L.; Wang, W.; and Jin, Q. 2022. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3108–3116.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; and Elhoseiny, M. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv:2303.06594*.