Mining Contextualized Visual Associations from Images for Creativity Understanding

Anonymous ACL submission

Abstract

Understanding another person's creative output requires a shared language of association. However, when training vision-language models such as CLIP, we rely on web-scraped datasets containing short, predominantly literal, alt-text. In this work, we introduce a method for mining contextualized associations for salient visual elements in an image that can scale to any unlabeled dataset. Given an image, we can use these mined associations to generate high quality creative captions at increasing degrees of abstraction. With our method, we produce a new dataset of visual associations and 1.7m creative captions for the images in MSCOCO. Human evaluation confirms that these captions remain visually grounded while exhibiting recognizably increasing abstraction. Moreover, fine-tuning a visual encoder on this dataset yields meaningful improvements in zero-shot image-text retrieval in two creative domains: poetry and metaphor visualization. We release our dataset, our generation code and our models for use by the broader community.

1 Introduction

We make sense of visual art through shared associations (Gombrich, 2023). Studies from the cognitive sciences have shown that these associations come from our collective biological, social, cultural and environmental contexts (Ward and Kolomyts, 2010). For example, skulls evoke death in many Western viewers. Consequently, creative vision-language tasks like art interpretation or image-to-poetry generation require models that can leverage these same associations (Huang et al., 2016; Hu et al., 2020; Liu et al., 2018; Lu et al., 2022).

However, while training on image-text pairs scraped from the web has yielded powerful models such as CLIP that are able to adapt to many downstream tasks, research has found that they often fail to achieve similar zero-shot performance in tasks where the domain is largely different from their pre-training data (Menon et al., 2024). This is especially true in creative domains. In poetry and metaphor visualization, CLIP's capabilities are limited (Guljajeva et al., 2023). We hypothesize that this is because the text seen during its pre-training is predominantly short alt-text which does not explicitly include any associations for its accompanying imagery (see Figure 1).

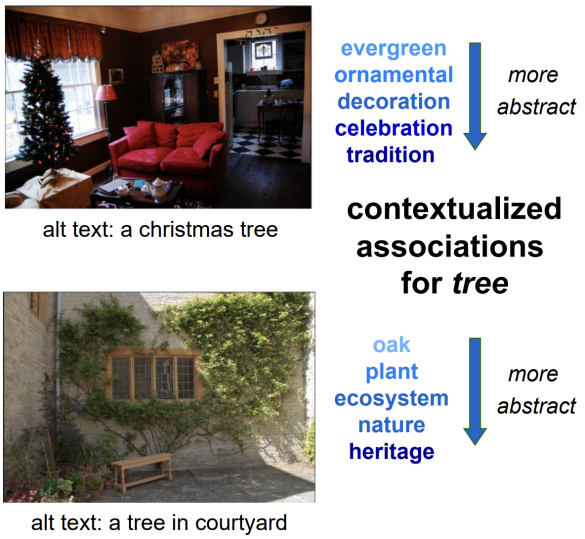


Figure 1: Two images depicting *trees* in different settings. Their alt-text makes no mention of the diverse concepts that each tree evokes. Using our method, we are able to mine contextualized associations at degrees of abstraction that extend beyond literal description.

Prior work has improved vision-language models (VLMs) by training on synthetic captions with fine-grained detail, resulting in more nuanced image understanding (Chen et al., 2024; Fan et al.; Lai et al., 2024). This has produced meaningful performance gains in classification and cross-modal retrieval tasks in non-creative domains.

In our work, we extend this effort to creative domains. We develop a method for mining contextualized visual associations for the salient elements in an unlabeled image. Here, we define contextu-

alized associations as concepts related to a particular visual element based on broader scene context (e.g., “celebration" for the Christmas tree in Figure 1). Then, we use these mined visual associations to synthetically produce creative captions for each image at increasing degrees of abstraction, informed by Hayakawa’s “ladder of abstraction" from linguistics (Hayakawa, 1967). This results in captions that remain grounded to an image while making explicit the associations that the image evokes.

Our data generation process is general purpose and can be arbitrarily scaled to any unlabeled corpus of images. We validate the quality of the resulting creative captions through 1) human evaluation and 2) testing the ability of a visual encoder fine-tuned on our synthetic dataset to adapt to two creative vision-language tasks: image-to-poetry retrieval and linguistic metaphor-to-visual metaphor retrieval (Liu et al., 2018; Chakrabarty et al., 2023a). We find that our synthetic captions reflect increasingly creative abstraction that aligns well with human judgment without introducing hallucination. Moreover, fine-tuning on these captions improves zero-shot multi-modal retrieval in both of our creative vision-language tasks.

In summary, the contributions of our work are:

- A novel approach for mining contextualized associations for visual elements in unlabeled images at increasing degrees of abstraction
- A new dataset, extending MSCOCO with increasingly abstract visual associations and accompanying high quality creative captions
- A human evaluation of our dataset, validating both the increasing abstraction and visual grounding of our synthetic captions
- An evaluation of CLIP, fine-tuned on our dataset, showing improved performance for multiple creative cross-modal retrieval tasks

Additionally, we release our dataset, generation code and models for use by the broader community: https://anonymous.4open.science/r/mining_visual_associations-1F0B.

2 Related Work

2.1 Conceptual Associations and Creativity

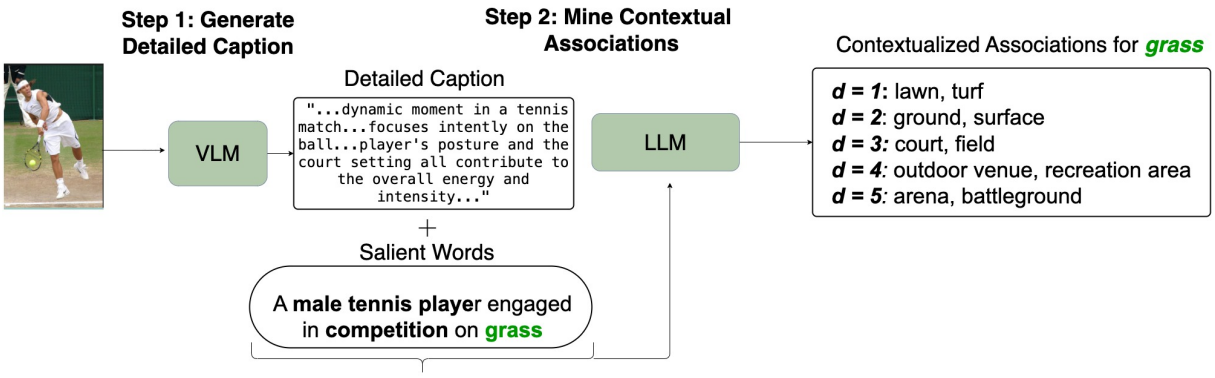
Cognitive science has shown that creativity involves associative thinking (Ward and Kolomyts, 2010). Often, this entails linking together related concepts through abstraction (Beaty and Kenett, 2023). In NLP, attempts to understand poetic language, including metaphor, simile and emotion, have required external associative knowledge to help make sense of implicit meaning (Chakrabarty et al., 2022). A common method for incorporating such knowledge is through the use of association lexicons. Previous studies have collected rich lexicons for the colors and emotions evoked by different words through painstaking human annotation (Mohammad, 2013; Mohammad and Turney, 2013). These were complemented by efforts at automating association mining through word embeddings (Bolukbasi et al., 2016; Hu et al., 2019). In contrast with this prior work, we present a method for automatically mining *contextualized* associations, where the same word’s related concepts vary based on its surroundings. Moreover, while previous lexicons have typically focused on text, our associations are *visually contextualized*, extending association mining to a new modality.

2.2 Synthetic Image-Text Data

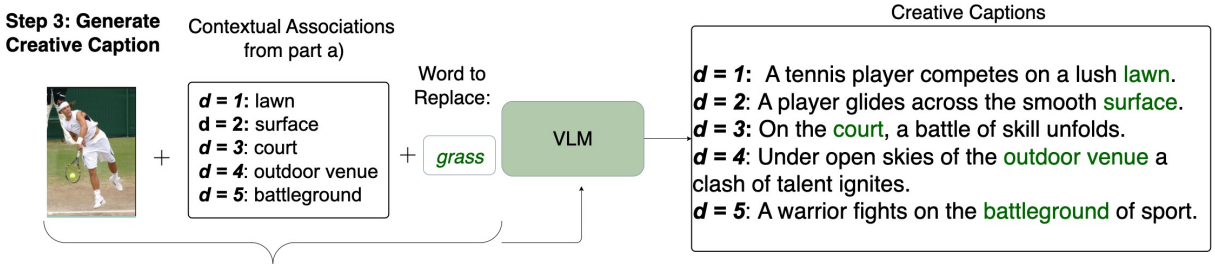
Due to the strength of current VLMs, recent work has exploited synthetic data to improve the downstream performance of image encoders on vision-language tasks (Zheng et al., 2024; Kong et al., 2024; Xiao et al., 2024; Liu et al., 2024; Yang et al., 2023). Studies have found that generated captions can be longer and more descriptive of images than their naturally occurring references (Chen et al., 2024; Sharifzadeh et al., 2024). Some have even shown that training on such captions can yield higher performance than training on those from human annotators (Santurkar et al., 2022). While exciting, the focus of much of this work has been on improving the performance of VLMs on standard image understanding tasks. In our work, we expand this line of inquiry to include creative domains. Building on our method for mining contextualized associations, we generate a corpus of creative captions and show that training on these captions yields significant improvements on zero-shot image-poetry and image-metaphor retrieval.

3 Generating Abstracted Captions

Given an image I featuring visual elements V_I, we mine a set of contextualized associations $A_d(v_j)$ for each $v_j \in V_I$ at increasing degrees of abstraction, $d \in D$. Then, using these associations, we generate a set of captions $C_d(I)$ for each image I



(a) Mining contextualized associations for *grass*.



(b) Generating creative captions using contextualized associations for *grass*.

Figure 2: Our method for mining contextualized associations and generating creative captions with increasing abstraction. In **Step 1**, given an image, we prompt a VLM to generate a detailed caption. Then, in **Step 2**, we prompt an LLM to mine associations for each of its salient visual elements at increasing degrees of abstraction. Finally, in **Step 3**, we prompt a VLM to generate synthetic creative captions using our mined associations.

that reflect the specified degree of abstraction, d.

In this work, we define contextualized associations as concepts that are related to the specified visual element v_i based on its broader scene context. For example, in Figure 1, a *tree* outside evokes different associations than an indoor Christmas tree in many Western viewers – these associations are mediated by each tree's surroundings.

We define five degrees of abstraction d, inspired by Hayakawa's "ladder of abstraction" from linguistics (Hayakawa, 1967):

1. **Near Synonyms** ($\mathbf{d = 1}$): Close in meaning or form (e.g., Ball → Sphere).
2. **Slight Abstractions** ($\mathbf{d = 2}$): Slightly broader category (e.g., Ball → Toy).
3. **Broader Context** ($\mathbf{d = 3}$): Indirect, but linked through situational and emotional context (e.g., Ball → Game).
4. **Conceptual Associations** ($\mathbf{d = 4}$): More abstract or thematic (e.g., Ball → Competition).
5. **Full Abstractions** ($\mathbf{d = 5}$): Highly abstract or metaphorical (e.g., Ball → Journey).

3.1 Mining Contextualized Associations

Given an image I with a short caption c_{short}, first, we generate a detailed caption $c_{detailed}$ using an off-the-shelf vision-language model (VLM). Then, we extract salient visual elements $v_1, \ldots, v_n \in V_I$ by identifying nouns, adjectives, and verbs in the short caption c_{short} with high concreteness ratings according to a lexicon (Brysbaert et al., 2014).

As large language models are trained on language that is much richer than the language typically found in image alt-text, they function as high quality repositories of common associations, especially when conditioned with complete scene context (Tsimpoukelli et al., 2021). Thus, we prompt a text-only frontier language model with both our detailed caption $c_{detailed}$ and our extracted visual elements V_I to mine contextualized associations $A_d(v_j)$ for each $v_j \in V$ at every degree of abstraction d. We include the full prompt in A.2.2.

3.2 Generating Captions

Given an image I, a salient visual element v_j and its conceptual associations $A_d(v_j)$ at degree of abstraction d, we prompt a VLM to generate a

Object	Image 1	Img 1 Associations	Image 2	Img 2 Associations
'horse'		D1: mare, stallion, equine D2: animal, creature D3: livestock, farm animal D4: nature's beauty, freedom D5: spirit, gentleness		D1: steed, stallion D2: equine, animal D3: transportation, driving D4: equestrianism, competition D5: freedom, power
'people'		D1: group, hikers D2: team, collective D3: explorers,adventurers D4: community, fellowship D5: existence, connection		D1: customers, vendors D2: crowd, patrons D3: community, society D4: interaction, gathering D5: human experience, societal dynamics
'bear'		D1: teddy, toy, stuffed animal D2: companion, cuddle buddy D3: comfort object, friend D4: comfort, affection D5: innocence, joy		D1: cub, grizzly D2: mammal, wildlife D3: nature, habitat D4: instinct, survival, family D5: wild spirit, biological heritage

Figure 3: Example from our corpus. For each object, we depict its contextualized associations at increasing degrees of abstraction for two of its representative images. These associations change with its visual surroundings.

creative caption $c_{creative}$ for each association in $A_d(v_j)$. We include the full prompt in A.2.3.

4 Experiments

4.1 Corpus Generation

For our corpus of images with short captions (I, c_{short}), we use Microsoft Common Objects in Context (MSCOCO) (Lin et al., 2014) due to its extensive study in vision language modeling. We note, however, that our method can be applied to any corpus of unlabeled images for which we can obtain high quality short captions; given the strength of current VLMs, this includes most image corpora (Bordes et al., 2024). To extract salient visual elements from each short caption c_{short}, we employ `SpaCy`'s part of speech tagger and filter words based on their concreteness ratings using the lexicon from Brysbaert et al. (2014) (requiring a minimum concreteness of 3). We produce detailed descriptions of each image using `Molmo-7B-D-0924` (Deitke et al., 2024). We use text-only `GPT-4o-mini`[1] to mine contextualized associations at different degrees of abstraction for each image's salient visual elements based on its detailed description. Finally, we use `Molmo-7B-D-0924` once again to generate a creative caption $c_{creative}$ for each extracted visual association. In total, we produce $1,671,835$ creative captions for MSCOCO$_{train}$ and $102,552$ creative captions MSCOCO$_{validation}$ respectively.

[1] specifically `gpt-4o-mini-2024-07-18`

4.2 Human Evaluation

In order to validate our method for mining contextualized visual associations and generating creative captions, we conduct a human evaluation of our synthetic dataset. We recruit five native English speakers to annotate a random sample of our corpus, answering two questions of interest: First, how visually grounded (i.e. free of mistakes / errors / hallucinations) are the creative captions? And second, how well do the generated creative captions reflect increasing abstraction?

To evaluate visual grounding, for 100 creative captions, we ask annotators to label whether the caption is completely contradictory to or not relevant to its image (rating of 1), contains many erroneous details but still describes its image (rating of 2), is an almost perfect caption with minor errors (rating of 3) or represents a perfect caption where there are no errors (rating of 4).

To evaluate abstraction, for each of 100 images, we ask annotators to rank six of its captions in order of increasing abstraction: its original caption and one creative caption from each of our five abstraction degrees, presented in randomized order.

We include our task instructions and screenshots of our annotation interfaces in section A.3.2.

4.3 Automatic Evaluation

In addition to a human evaluation, we validate our method for mining contextualized associations and generating creative captions by fine-tuning a pre-trained visual encoder on our corpus of creative captions for MSCOCO. In particular, we ex-

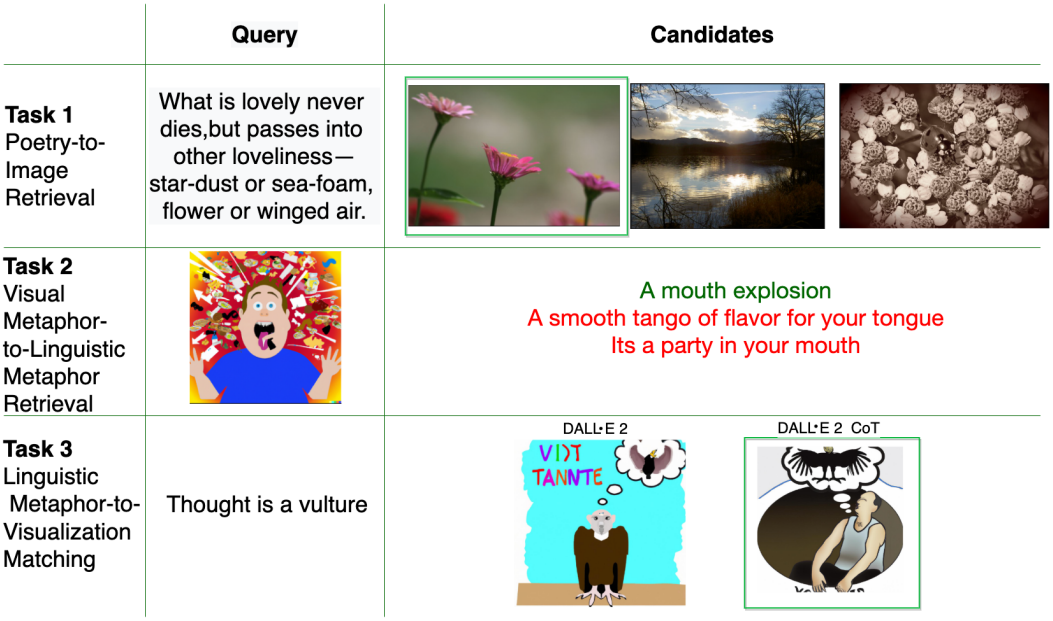


Figure 4: Examples from each of our three evaluation tasks. Correct answers are highlighted in green.

pand `OpenCLIP-ViT-B/32`[2] with a learnable prefix specific to each of our five degrees of abstraction $d \in D$ (Li and Liang, 2021; Menon et al., 2024). Keeping the rest of the model frozen, we update only these prefix embeddings by optimizing CLIP's contrastive image-text matching loss on our corpus. We fine-tune these weights for a single epoch.

We compare our baseline, `OpenCLIP-ViT-B/32` without any fine-tuning, to our fine-tuned model at all five different degrees of abstraction – that is, using each of our five learned abstraction prefixes. We evaluate image-text similarity scores from these models on three zero-shot tasks constructed from datasets in two creative domains:

- **Multi-Modal Poem (MultiM-Poem)** (Liu et al., 2018): Contains $8,292$ images from Flickr paired by English majors with short poems (around 7 lines) from several online poetry sites[3]. We use MultiM-Poem for **Task 1**, poetry-to-image retrieval: given a poem, retrieve its corresponding image.

- **HAIVMet** (Chakrabarty et al., 2023b): Contains $1,540$ linguistic metaphors paired with both incorrect, overly literal, visualizations generated by DALL·E2 and correct, appropriately metaphorical, visualizations generated by DALL·E2 through chain-of-thought. We use HAIVMet for **Task 2**, visual metaphor-to-linguistic metaphor retrieval, and **Task 3**, linguistic metaphor-to-visualization matching.

For our retrieval tasks, we report recall at $k = 1, 5, 10, 20$ as well as the average rank of the correct text or image among all candidate texts or images (where lower is better). For our matching task, we report how often the correct visualization is chosen over the incorrect visualization. We provide examples of each evaluation task in Figure 4.

5 Results and Discussion

Abstraction	% with Grounding ≥ 3
Captions at $d = 1$	0.9
Captions at $d = 2$	0.87
Captions at $d = 3$	0.93
Captions at $d = 4$	0.77
Captions at $d = 5$	0.92

Table 1: The percentage of our creative captions at each degree of abstraction that our annotators judge as exhibiting visual grounding ≥ 3 on our 4-point Likert scale. Our captions demonstrate consistent alignment with their paired images, despite increasing abstraction.

5.1 How good are our creative captions?

In Table 1, we show the results of our first human evaluation task, rating the visual grounding of our creative captions. While annotators rated captions on a four-point Likert scale, we bucket the resulting labels into two groups, $(1, 2)$, indicating poor

[2]pre-trained on the `laion2b_s34b_b79k` dataset
[3]Foundation3, PoetrySoup4, best-poem.net and poets.org

Caption Type	Average Rank
Original Captions	1.47
Captions at $d = 1$	2.69
Captions at $d = 2$	3.39
Captions at $d = 3$	4.03
Captions at $d = 4$	4.50
Captions at $d = 5$	4.98

Table 2: The average abstraction rank (out of 6) for MSCOCO's original and our creative captions. We find that as our specified degree of abstraction increases, annotators rank the resulting creative captions as exhibiting more abstraction, validating our method.

visual grounding, and $(3, 4)$ indicating acceptable visual grounding. First, we note we observe a Fleiss κ of 0.303, indicating fair agreement, for this visual grounding assessment as calculated from three-way annotation on 20% of our tasks (Fleiss, 1971). When considering our overall results, we can see that our creative captions demonstrate consistent visual alignment with their images – in fact, at abstraction degrees 1, 3 and 5, this is true of more than 90% of our creative captions. Our method for mining contextualized associations and generating creative captions generally avoids introducing errors and hallucinations.

In Table 2, we show the results of our second human evaluation task, ranking an image's captions (its original caption and a creative caption sampled for each of our five degrees of abstraction) in order of increasing abstraction. We collect three annotations for 20% of these tasks and observe a Fleiss κ of 0.283, indicating fair agreement, especially given its subjective nature. When considering our overall results, we can see the average rank of captions at each degree of abstraction reflects the intended abstraction, even relative to one another. Captions at smaller degrees have lower rank than captions at larger degrees. Original captions receive the lowest average rank of 1.47 and captions at $d = 5$ receive the highest average rank of 4.98. Our method is capable of consistently generating increasingly abstract creative captions.

5.2 Does OpenCLIP agree?

Our human evaluation makes clear that our synthetic creative captions are both visually grounded and exhibit recognizably increasing abstraction. Does the original OpenCLIP model agree?

To test this, for each image in our corpus, we calculate its similarity with its original caption and its similarity with its creative captions. Additionally, as a baseline, we calculate its similarity with captions containing obvious hallucinations from the FOIL dataset (Shekhar et al., 2017).

On the left side of Figure 5, we plot how often the original captions score higher than 1) our creative captions at increasing degrees of abstraction and 2) the FOIL captions. As the degree of abstraction increases, OpenCLIP favors the original captions more and more. In fact, at our highest degree of abstraction, OpenCLIP prefers the original caption 80% of the time, nearly the same rate at which it prefers the original caption over the hallucinatory captions from FOIL. This suggests a strong preference for literal over creative captions.

On the right side of Figure 5, we plot how often hallucinatory FOIL captions score higher than 1) original captions and 2) our creative captions at increasing degrees of abstraction. As the degree of abstraction increases, it becomes more and more difficult for OpenCLIP to distinguish between obvious hallucinations and abstraction. In fact, at our highest degree of abstraction, OpenCLIP does no better than random guessing. This shows that standard image-text datasets result in models unable to differentiate between hallucination and abstraction.

5.3 Do contextualized associations improve downstream creative understanding?

In Tables 3, 4 and 5, we compare the performance of OpenCLIP against the performance of our model fine-tuned on our synthetic captions at all five degrees of abstraction across our selected creative understanding tasks. In all three tasks, our creative captions yield significant improvements over the baseline despite no task-specific fine-tuning.

On poetry-to-image retrieval (Table 3), our fine-tuned variant improves over the baseline in both recall and average rank when the degree of abstraction is set to either 4 or 5, with 5, our highest degree of abstraction, exhibiting the best performance.

On visual metaphor-to-textual metaphor retrieval, both zero-shot CLIP and our fine-tuned variant struggle to achieve reasonable recall values. However, when we plot the average rank of the correct textual metaphor, we see that increasing the degree of abstraction in our fine-tuned visual encoder yields consistent reductions.

On linguistic metaphor-to-visualization matching (Table 5), our fine-tuned variant improves over the baseline at every degree of abstraction. Interestingly, we observe the largest improvement at

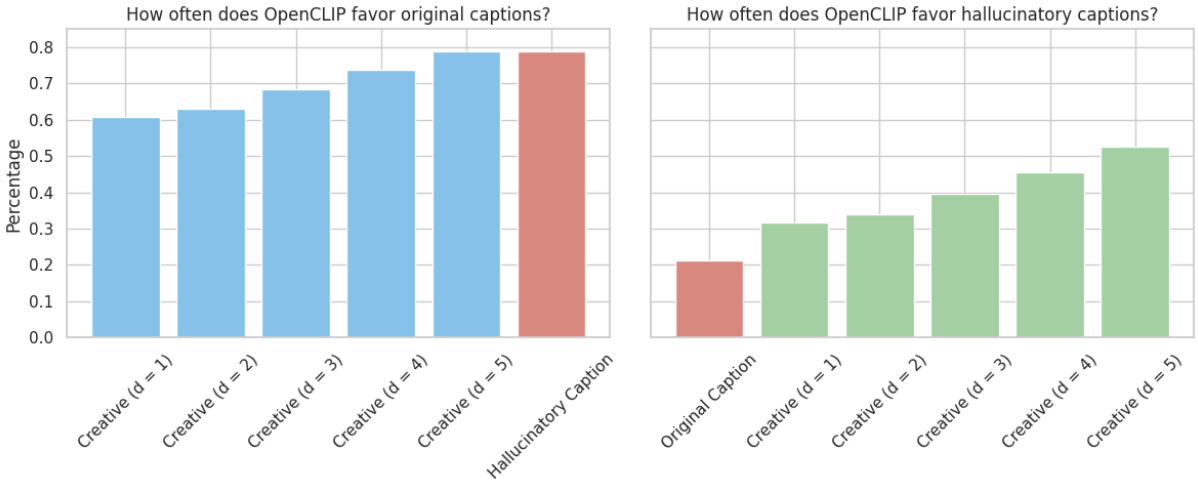


Figure 5: Plots comparing OpenCLIP's scores for original (left) and hallucinatory (right) captions against its scores for our creative captions. OpenCLIP favors literalism and cannot distinguish between hallucination and abstraction.

Model	**k = 1 (↑)**	**k = 5 (↑)**	**k = 10 (↑)**	**k = 20 (↑)**	**Avg Rank (↓)**
OpenCLIP	0.1505	0.3089	0.4033	0.5043	70.46
OpenCLIP-FT ($d = 1$)	0.1454	0.3086	0.3934	0.4977	72.37
OpenCLIP-FT ($d = 2$)	0.1446	0.3048	0.3913	0.4877	72.37
OpenCLIP-FT ($d = 3$)	0.1485	0.3106	0.3935	0.4912	72.41
OpenCLIP-FT ($d = 4$)	0.1591	0.3233	0.4150	0.5147	68.96*
OpenCLIP-FT ($d = 5$)	**0.1624**	**0.3341**	**0.4162**	**0.5222**	**67.60***

Table 3: **Task 1: Poetry-to-Image Retrieval.** Recall@k and average rank of OpenCLIP and a variant fine-tuned on our creative caption corpus at increasing degrees of abstraction. At degrees 4 and 5, our fine-tuned model outperforms the baseline across all metrics. * indicates significance at $\alpha = 0.05$.

Model	**Avg Rank (↓)**
OpenCLIP	3288.9
OpenCLIP-FT ($d = 1$)	3262.6
OpenCLIP-FT ($d = 2$)	3266.4
OpenCLIP-FT ($d = 3$)	3264.2
OpenCLIP-FT ($d = 4$)	3253.4
OpenCLIP-FT ($d = 5$)	**3244.5***

Table 4: **Task 2: Visual Metaphor-to-Linguistic Metaphor Retrieval.** The average rank of OpenCLIP and a variant fine-tuned on our creative caption corpus at increasing degrees of abstraction. As abstraction increases, our model's average rank improves over the baseline. * indicates significance at $\alpha = 0.05$.

a relatively low degree of abstraction ($d = 1$), a break with prior trends.

In making sense of the differences among our model's performances across all three tasks, we hypothesize that one important source of variation could be the composition of our evaluation data. Much like our synthetic corpus, which contains creative captions paired with ordinary images, MultiM-Poem contains figurative language paired with photographs from Flickr. This poses a smaller domain shift than HAIVMet, where creative language is paired with creative imagery. Nevertheless, given the improvements exhibited by our fine-tuned model and the relative ease of applying our corpus generation technique to other image corpora, we view our results as strong evidence for the value of our mined associations in adapting vision-language understanding to creative domains.

6 Conclusion & Future Work

In this work, we introduce a scalable method for mining contextualized associations for visual elements that can be applied to any corpus of unlabeled images. We use these associations to produce a new dataset of increasingly abstract creative captions for MSCOCO. Both human judgment and automatic evaluation across three challenging image-language tasks confirm the value of this method for enabling creativity understanding. In the future, we plan to extend this study beyond English and

Model Name	% preference for DALL·E 2 (CoT) ↑
OpenCLIP	0.43
OpenCLIP-FT ($d = 1$)	$\mathbf{0.59}^*$
OpenCLIP-FT ($d = 2$)	0.47
OpenCLIP-FT ($d = 3$)	0.54^*
OpenCLIP-FT ($d = 4$)	0.49
OpenCLIP-FT ($d = 5$)	0.50

Table 5: **Task 3: Linguistic Metaphor-to-Visualization Matching.** Preference for the correct visualization of OpenCLIP and a variant fine-tuned on our creative caption corpus at increasing degrees of abstraction. All abstraction settings improve over the baseline. * indicates significance at $\alpha = 0.05$.

Western associations – recent work has shown that, in some cases, VLMs exhibit culturally specific regularities when prompted in different languages (Ananthram et al., 2024). It is our hope to leverage this to mine multicultural associations at scale.

7 Limitations

While our method for mining visual associations and generating creative captions is easy to scale, we acknowledge its reliance on `gpt4o-mini`, a paid closed source model. Additionally, we use `Molmo` to generate both the detailed descriptions and the creative descriptions of the images in our corpus. LLMs and VLMs are both prone to hallucinations and biases which could be reflected and reinforced by both our method and our dataset. Moreover, there is room for improvement across all evaluation tasks which can be achieved through using additional datasets including more variation in images and captions as well as other prompting techniques that have not been explored in this work. Finally, our contextualized associations are limited to the English language and likely reflect a Western-centric perspective. However our methods allows for scalability in other languages which can be conducted in future work.

References

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.

Roger E Beaty and Yoed N Kenett. 2023. Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7):671–683.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Tuhin Chakrabarty, Vishakh Padmakumar, He He, and Nanyun Peng. 2023a. Creative natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 34–40.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023b. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *arXiv preprint arXiv:2305.14724*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ernst Hans Gombrich. 2023. Art and illusion: A study in the psychology of pictorial representation-millennium edition.

Varvara Guljajeva, Mar Canet Solà, and Isaac Joseph Clarke. 2023. Explaining clip through co-creative drawings and interaction. *arXiv preprint arXiv:2306.07429*.

Samuel Ichiyé Hayakawa. 1967. Language in thought and action. *The Florida English Journal*, 3(2):1–12.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976.

Zheng Hu, Jiao Luo, Chunhong Zhang, and Wei Li. 2019. A natural language process-based framework for automatic association word extraction. *IEEE Access*, 8:1986–1997.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Fanjie Kong, Yanbei Chen, Jiarui Cai, and Davide Modolo. 2024. Hyperbolic learning with synthetic captions for open-world detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16762–16771.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. 2024. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pages 111–127. Springer.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. *arXiv preprint*.

Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 783–791.

Yanqing Liu, Xianhang Li, Zeyu Wang, Bingchen Zhao, and Cihang Xie. 2024. Clips: An enhanced clip framework for learning with synthetic captions. *arXiv preprint arXiv:2411.16828*.

Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. 2022. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 11(1):576–587.

Sachit Menon, Ishaan Preetam Chandratreya, and Carl Vondrick. 2024. Task bias in contrastive vision-language models. *International Journal of Computer Vision*, 132(6):2026–2040.

Saif Mohammad. 2013. Colourful language: Measuring word-colour associations. *arXiv preprint arXiv:1309.5942*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. 2022. Is a caption worth a thousand images. *A Controlled Study for Representation Learning. CoRR abs/2207.07635*, 5.

Sahand Sharifzadeh, Christos Kaplanis, Shreya Pathak, Dharshan Kumaran, Anastasija Ilic, Jovana Mitrovic, Charles Blundell, and Andrea Banino. 2024. Synth ^{2}: Boosting visual-language models with synthetic captions and image embeddings. *arXiv preprint arXiv:2403.07750*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Thomas B Ward and Yuliya Kolomyts. 2010. Cognition and creativity. *The Cambridge handbook of creativity*, 5:93–112.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. 2024. Flair: Vlm with fine-grained language-informed image representations. *arXiv preprint arXiv:2412.03561*.

Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. 2024. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pages 73–90. Springer.

A Appendix

A.1 Computation and Model Specifics

The base CLIP model has around 86 million parameters. Our CLIP model variant was trained on 1 NVIDIA RTX A6000 GPU for 1 epoch taking roughly 3 hours. We used a learning rate of `1e-4` and `torch.optim.Adam` optimizer. We use early stopping and lowest validation loss with a `patience = 3` on our synthetic corpus validation dataset to determine the best model.

The specific version of Spacy we use is spacy 3.8.4

We use the OpenAI batch API to generate our associations. The specific hyperparemters we use apart from defaults is `max_tokens: 1000`

The hyperparameters we use for Molmo is `temperature=0.7`, `top_p=0.9`, `max_tokens=150`, `n=1`.

A.2 Generating Abstracted Captions

A.2.1 Detailed Caption Prompt

Below is the prompt used for generating the detailed caption of a given image.

""USER: <image> Please generate a detailed caption of this image. ASSISTANT:"

A.2.2 Mining Associations Prompt

We prompt `GPT-4o-mini` using the batch API with the following system prompt where {context_caption} refers to the detailed caption generated for an image and {original_caption} refers to the MSCOCO caption of the image:

"For a given list of words, generate a new list for each word using the same part of speech. The words should follow a semantic abstraction scale where distance increases from near-synonyms to abstract concepts.
Approach:
1. Distance 1 – Near Synonyms: Close in meaning or form (e.g., Ball → Sphere). 2. Distance 2 – Slight Abstraction: Slightly broader category (e.g., Ball → Toy). 3. Distance 3 – Broader Context: Indirectly linked through situational and emotional context (e.g., Ball → Game). 4. Distance 4 – Conceptual Association: More abstract or theme-related (e.g., Ball → Competition). 5. Distance 5 – Full Abstraction: Highly abstract or metaphorical (e.g., Ball → Journey).

Generate three words each for distances 1 to 5. Generated words should fit into the overall emotional and situational context of this context caption: {context_caption}

Generated words, when replaced with the original word in this short caption {original_caption}, should be semantically correct.

Do not generate the original word in the new generations. Use JSON format: the key is the original word, and the value is a dictionary with distances as keys and lists of generated words as values."

A.2.3 Abstracted Caption Prompt

Below is the prompt used to obtain creative captions for an image for each of its salient objects and associations generated. {all_words} contain the salient words for the image with the original word replaced with the association word at distance {level}. {new_word} is the association word at distance {level}. <image> is the input image

"USER: <image> Write a short caption grounded in this image and semantically correct, using fewer than 10 words. Choose some or all of these words: {all_words} to best represent the image.

Steer the caption's style toward the abstraction level _label_ following these rules:

- Distance 1: Near Synonyms – Close in meaning to the original image - Distance 2: Slight Abstraction – Slightly more abstract than the image - Distance 3: Broader Context – Indirectly linked through situational or emotional context - Distance 4: Conceptual Association – More abstract, theme-related to the image - Distance 5: Full Abstraction – Highly abstract or metaphorical

The caption MUST include the word: {new_word}. ASSISTANT:"

A.3 Evaluation

A.3.1 Significance Tests

We use pairwise t-tests to report significance results on the results of task 3 involving a pairwise preference of images. Specifically we use `scipy.stats ttest_rel` implementation

We use wilcoxin tests to report significance results on task 1 and 2 involving average ranks of the correctly retrieved image/text. Specifically we use `scipy.stats wilcoxon` implementation

A.3.2 Annotators and Annotation Interfaces

We do not report demographics of annotators to maintain full anonymity. Collected annotator data are fully anonymized. Annotators were informed of their annotations would be used for research purposes. Below are the instructions and interfaces annotators used to complete the annotations tasks.

1. Each image has 6 captions. *Your task is to rank each caption on a scale of 1 through 6 where 1 refers to the caption that has the most literal description of the image and 6 refers to caption that have the most abstract descriptions of the image.*

For clarity, literal descriptions of images will provide a straightforward, factual account of what is visually present in the image. Meanwhile abstract descriptions take more creative liberty, conveying the essence, emotions, or symbolic meaning rather than capturing the concrete details.

2. **Each rank should be unique**—no duplicate rankings for a single image.

3. **Select a rank for every caption** before submitting.

Thank you!

Figure 6: Instruction given to annotators for task 1 of Human Evaluation

1. Each image has 5 captions. ***Your task is to label each caption on a scale of 1 through 4 where 1 refers to the captions that completely contain language that is contradictory to or not present in the image, 2 refers to captions that contain many erroneous details but still describe the image, 3 refers to almost perfect captions with minor errors and 4 represents a perfect caption where there are no errors.***

For clarity, if the caption is vague or abstract but does not contain mistakes (hallucinated objects or actions) this is a correct caption and should be labeled as 4. However if a caption describes the image well but contains minor details which are factually incorrect based on the objects present in the image this would be considered incorrect and be given a label of 2(more than two details incorrect) or 3 (two or fewer details incorrect) depending on the amount of errors.

2. **Multiple captions for a given image can be given the same label**

3. **Each caption must be given a label** before submitting.

Figure 7: Instruction given to annotators for task 1 of Human Evaluation

Rank Each Caption for the given image with labels 1(most literal) to 6 (most abstract) *

	1	2	3	4	5	6
Expecting to conquer the wave, the surfer rides the crest of a powerful swell.	○	○	○	○	○	○
Pausing at the wave's edge, a surfer rides time.	○	○	○	○	○	○
Surfer balances on board as wave rests beneath.	○	○	○	○	○	○
A person on a surfboard about to hit a wave	○	○	○	○	○	○
Patience allows a surfer to navigate life's waves with grace.	○	○	○	○	○	○
Timeless surfer riding wave on board.	○	○	○	○	○	○

label each caption 1(captions that completely contain language that is contradictory to or not present in the image) through 4 (represents a perfect caption where there are no errors) (note: multiple captions can have same label) *

	1	2	3	4
Sailing vessel glides across expansive waters.	○	○	○	○
Freedom on the water: A sailboat glides through vast expanse.	○	○	○	○
Horizon: A sailboat glides on calm waters, framed by a distant mountainous landscape.	○	○	○	○
A sailboat glides across a vast expanse of water near a expansive field.	○	○	○	○
A tranquil setting where a sailboat gently glides on calm waters.	○	○	○	○

Figure 9: Example interface for one annotation from task 1