Understanding Figurative Meaning through Explainable Visual Entailment

Arkadiy Saakyan [1], Shreyas Kulkarni [1], Tuhin Chakrabarty [1], Smaranda Muresan [1,2]

[1]Columbia University, [2]Barnard College

Motivation

- To what extend do Vision-Language Models understand figurative meaning?
- V-FLUTE, an expert-verified dataset of 6,027 {image, caption, label, explanation} instances built using a human-LLM collaboration framework covering several phenomena: metaphors, similes, idioms, sarcasm, and humor.
- A suite of evaluations to assess current VLMs' capabilities on this new task of explainable visual figurative entailment.
- A detailed human evaluation with error analysis yielding insights into the types of errors for different classes of models.

Related Work and Our Contribution

Phenomenon	Data Source	Visual Style	Figurative Part	Our Contribution	# instances
Metaphor/ Simile	HAIVMet [1]	Illustration	Image	Image Selection Textual Explanations Expert Verification	857 (450 E, 407 C)
	IRFL [5]	Photographic	Caption	Image Selection Textual Explanations Expert Verification	1,149 (574 E, 575 C)
Idiom	IRFL [5]	Photographic	Caption	Image Selection Textual Explanations Expert Verification	370 (186 E, 184 C)
Sarcasm	MuSE [2]	Meme	Caption	Caption Generation Textual Explanations Expert Verification	1,042 (521 E, 521 C)
Humor	MemeCap [4]	Meme	Image	Caption Generation Textual Explanations Expert Verification	1,958 (979 E, 979 C)
	NYCartoons [3]	Illustration	Image+Caption	Taken As Is	651 (651 E)

Table 1. V-FLUTE dataset composition: 5 figurative phenomena, source datasets, visual styles, and our contributions. E denotes number of entailment instances, C - contradiction. Diversity of the dataset ensures coverage of various figurative phenomena, figurative meaning location, and visual styles.

Example: Creating Metaphor Subset for V-FLUTE

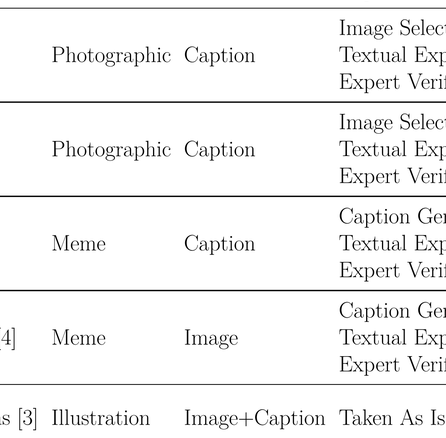


Figure 1. Creation of V-FLUTE instances for metaphors and similes from HAIVMet [1].

Examples

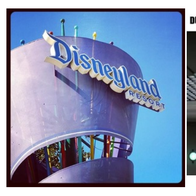
HAIVMet	IRFL	MuSE	MemeCap	NYCartoons
				
The faculty meeting was peaceful.	Their relationship is a house on fire.	Oh I just #love having to stare at this while I #work.	Even death won't exempt you from going to work.	Easy for you to say, you're cured!
Contradiction	Entailment	Contradiction	Entailment	Entailment
The image shows a faculty meeting transformed into a dramatic battlefield ... The visual metaphor suggests the faculty meeting was like a war, and not peaceful.	The photo suggests a conflict or an intense emotional situation ... which aligns with the symbolism of a house on fire representing a relationship filled with turmoil or heated arguments.	The image shows Disneyland Resort sign ... the person would like to experience it in person rather than just looking at the sign during work hours.	The image shows RoboCop ... it humorously illustrates a character who has been reanimated as a cyborg to continue working despite having died.	A play on the word "cured". People seek therapy to have their mental problems remedied or cured. But "cured" can also refer to a meat prep technique ...

Table 2. Sample dataset instances form V-FLUTE corresponding to the source datasets displaying images (premise), captions (hypothesis), labels, and explanations [Row 1-5].

Models and Metrics

2.

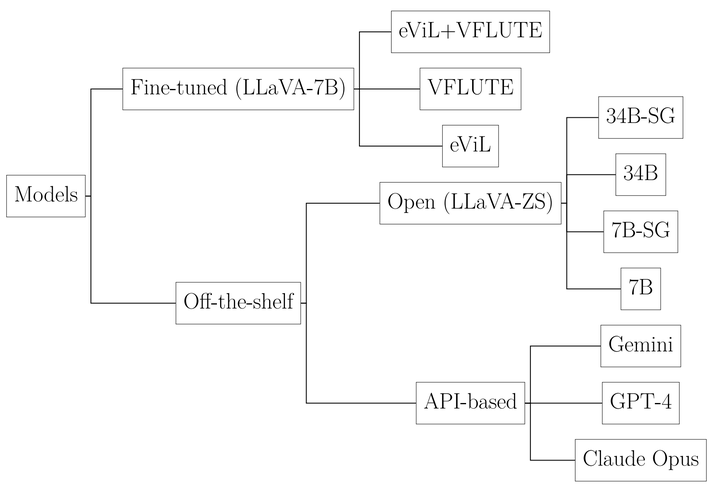


Figure 2. Taxonomy of models used for the study.

F1, F1@ExplanationScore

Main Results

- Result 1: General visual entailment does not solve figurative visual entailment

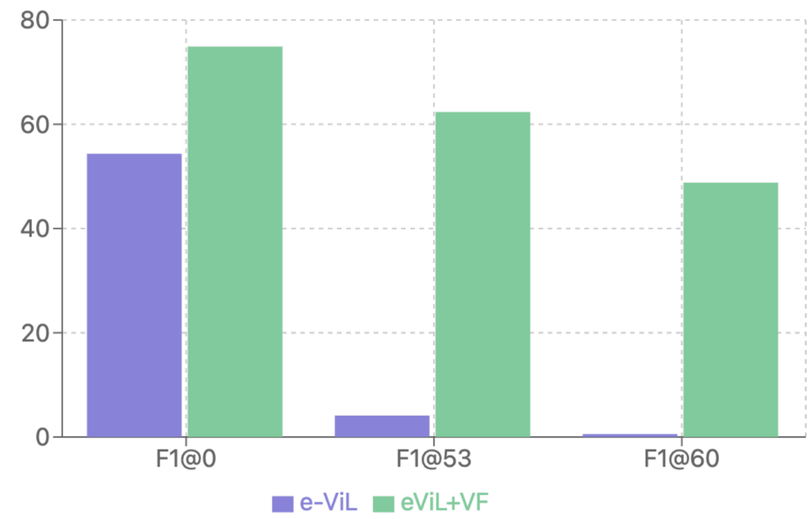


Figure 3. Performance difference when training just on literal entailment vs. figurative + literal.

Result 2: Figurative meaning in image is harder to explain compared to figurative meaning in text

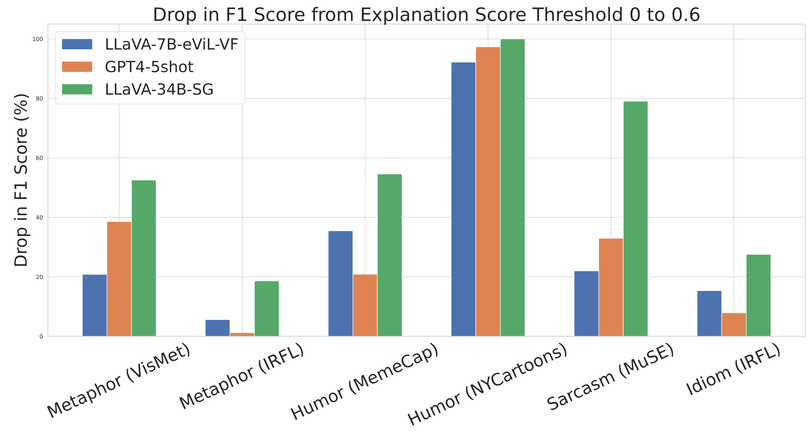


Figure 4. % Drop in F1 score for various models by source dataset between 0 to 0.6. Higher drop indicates higher proportion of wrongly generated explanations.

Result 3: VLMs benefit from visual information when dealing with figurative meaning

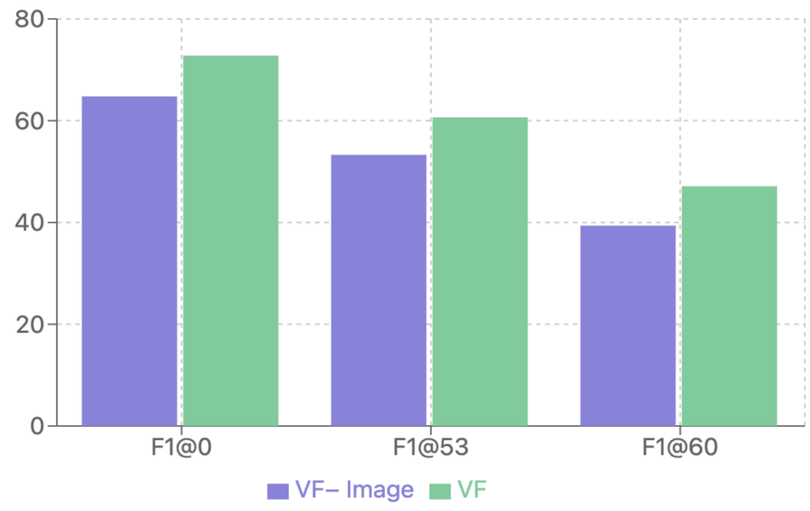


Figure 5. Ablation: performance when training with image vs. not including the image.

Human Evaluation and Error Analysis

	LLaVA-7B eViL+VF	LLaVA-34B SG	GPT-4 (5 shot)
Adequate %	33.78	29.85	50.67
Preference %	23.08	7.69	44.23

Table 3. Adequacy and Preference rates for generated explanations.

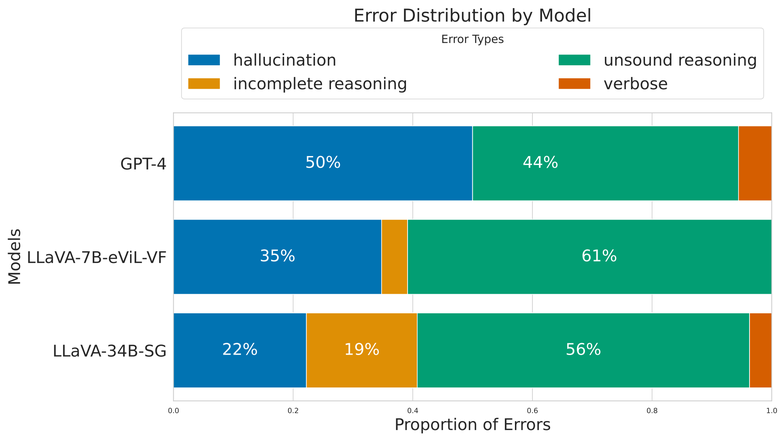


Figure 6. Normalized frequency of main error types in the explanation by model.

References

[1] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan.
I spy a metaphor: Large language models and diffusion models co-create visual metaphors.
In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 7370–7388, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar.
Nice perfume. how long did you marinate in it? multimodal sarcasm explanation.
Proceedings of the AAAI Conference on Artificial Intelligence, 36(10):10563–10571, Jun. 2022.

[3] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi.
Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest.
In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics.

[4] EunJeong Hwang and Vered Shwartz.
MemeCap: A dataset for captioning and interpreting memes.
In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 1433–1445, Singapore, December 2023. Association for Computational Linguistics.

[5] Ron Yosef, Yonatan Bitton, and Dafna Shahaf.
IRFL: Image recognition of figurative language.
In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1044–1058, Singapore, December 2023. Association for Computational Linguistics.