Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

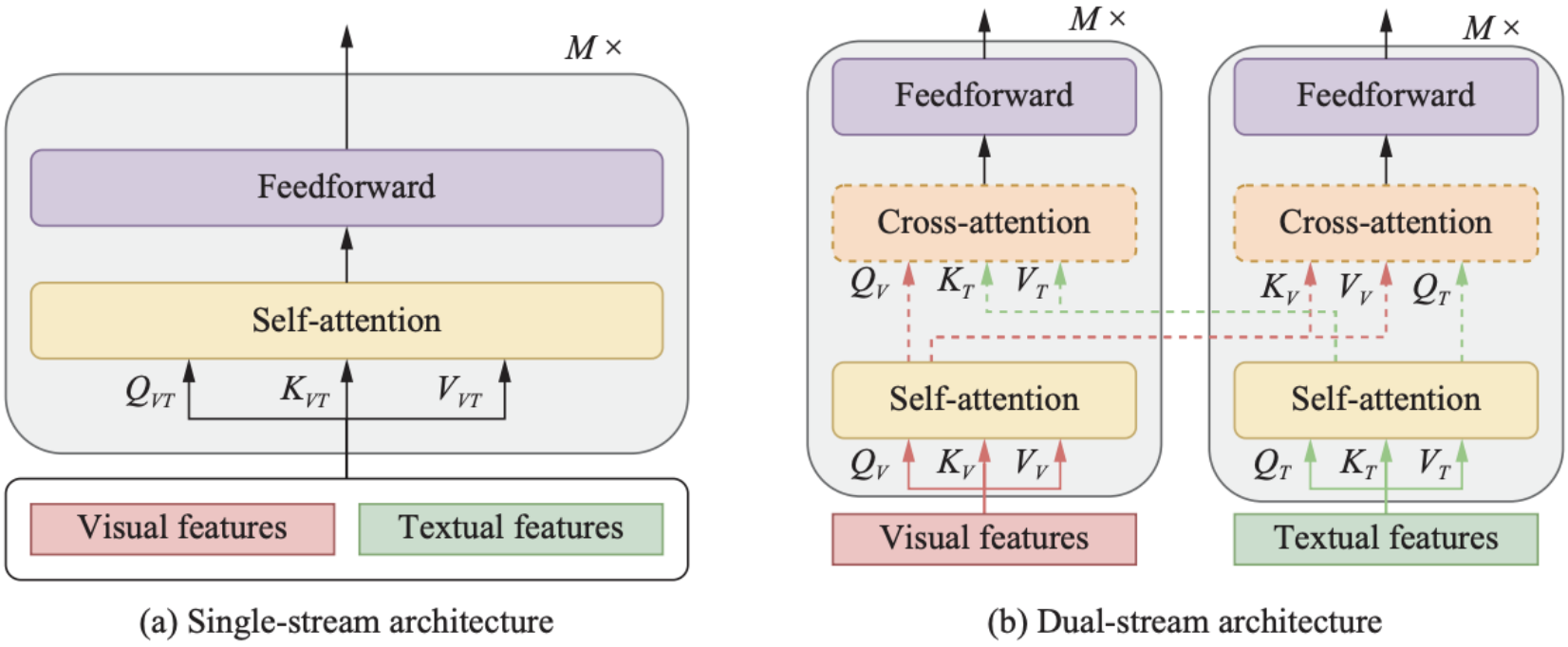
Focus

Multimodal fusion degree (integration);

- Single-stream vs. dual-stream models
- Modality importance

Recap: Vision–Language Models

BERT-like

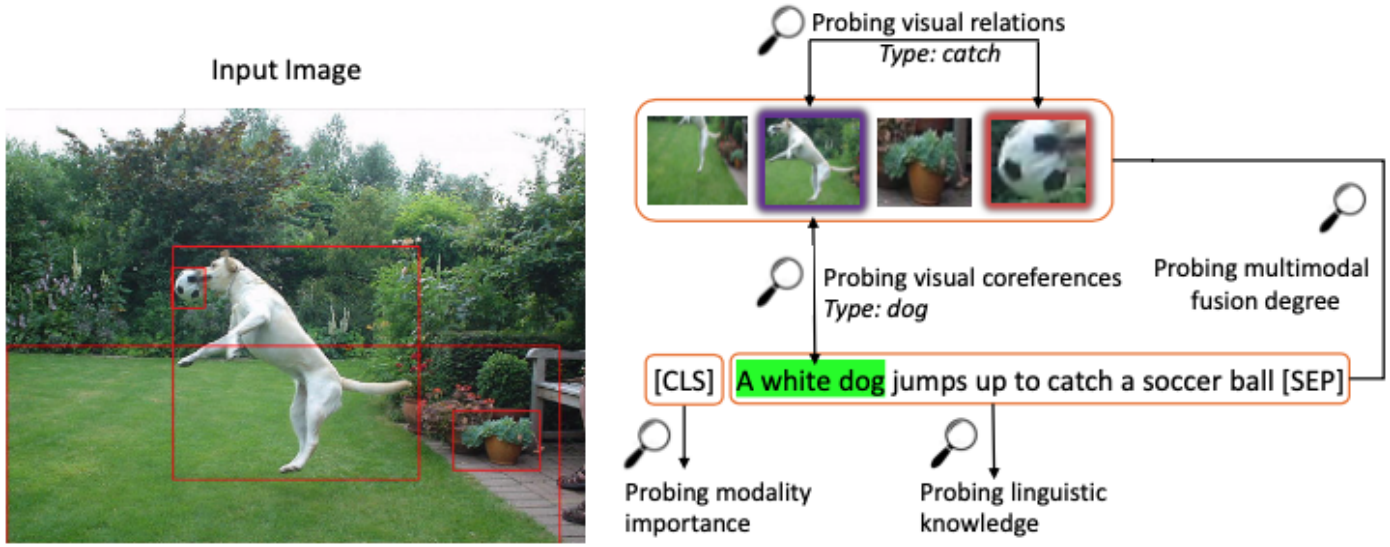


(a) Single-stream architecture

(b) Dual-stream architecture

source: [Chen et al., 2022]

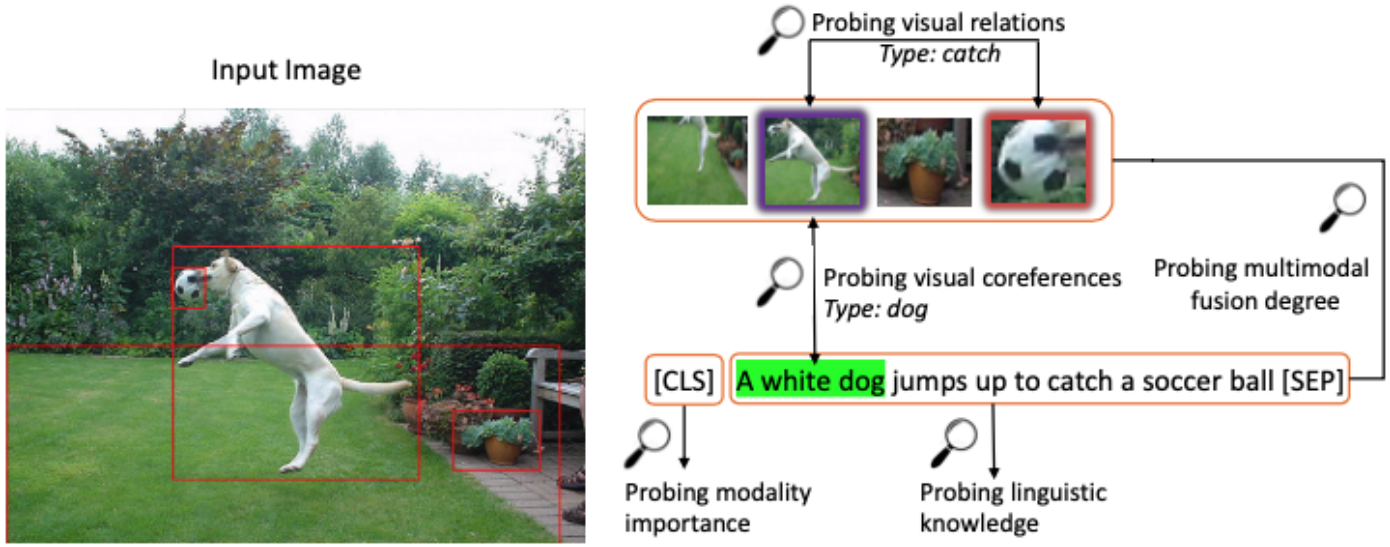
Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]



Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

- Multimodal Fusion Degree: Single-stream models learn more intertwined fusion of modalities, increasing with network layer depth (and decreasing in two-stream models)
- Modality Importance: Text more dominant than Vision
- Single-stream models better at capturing cross-modal interaction (visual coreference resolution), image-to-image interaction (visual relationship detection), and text-to-text interaction (linguistic knowledge - SentEval task)

Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]



Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

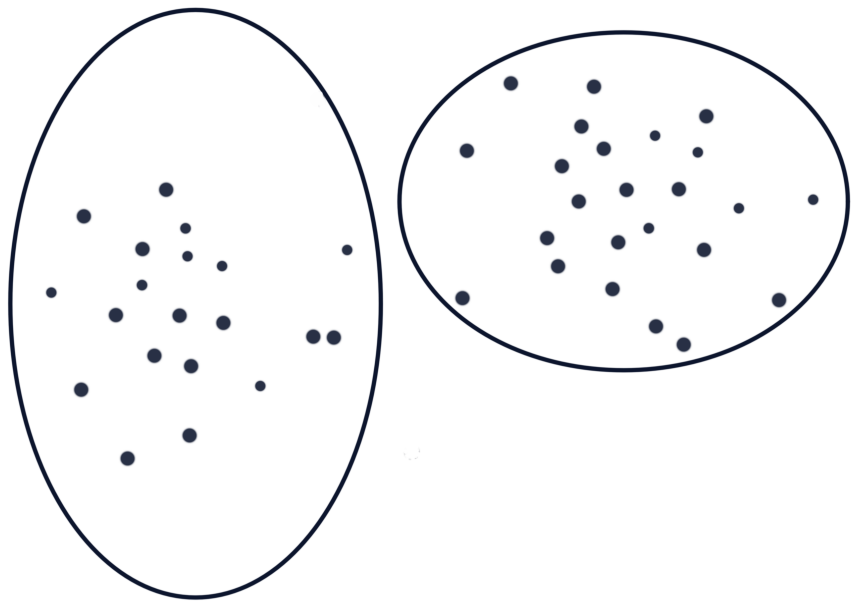
- Multimodal Fusion Degree
  - Extract multimodal embeddings of all tokens and image regions

visual and textual embeddings

Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

- Multimodal Fusion Degree
  - Extract multimodal embeddings of all tokens and image regions
  - Cluster into two sets

visual and textual embeddings



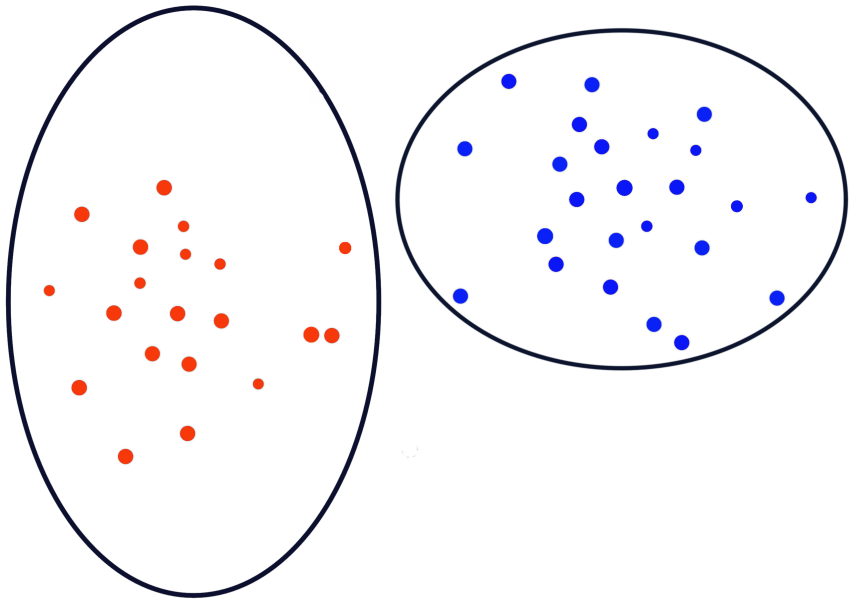
Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

- Multimodal Fusion Degree
  - Extract multimodal embeddings of all tokens and image regions
  - Cluster into two sets
  - Measure difference of clusters and ground-truth visual / textual clusters (Normalised Mutual Information)

$$MI(x, y) = P(x, y) \log \frac{P(x,y)}{P(x)P(x)}$$

  - High difference $\Rightarrow$ each cluster $\approx$ single modality $\Rightarrow$ low degree of multimodal fusion

visual and textual embeddings



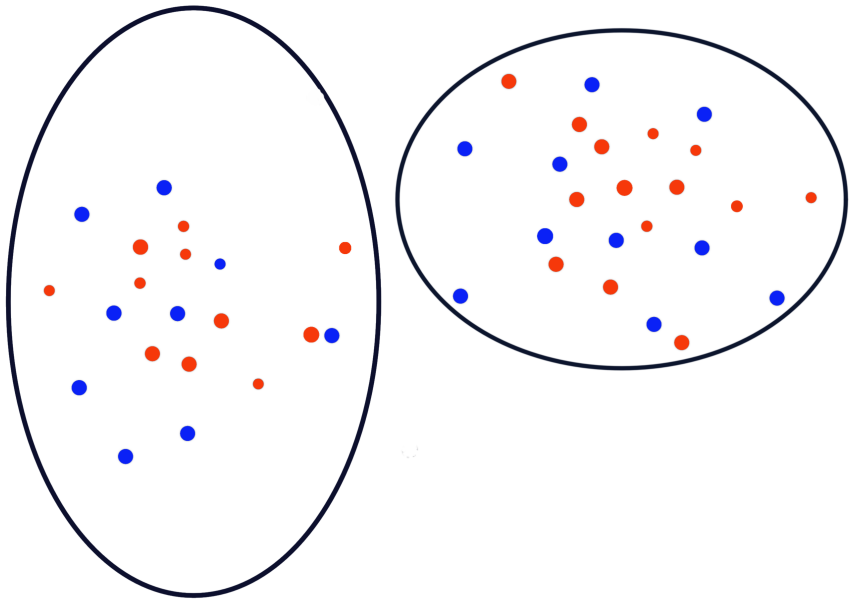
Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

- Multimodal Fusion Degree
  - Extract multimodal embeddings of all tokens and image regions
  - Cluster into two sets
  - Measure difference of clusters and ground-truth visual / textual clusters (Normalised Mutual Information)

$$MI(x, y) = P(x, y) \log \frac{P(x,y)}{P(x)P(x)}$$

  - Low difference $\Rightarrow$ each cluster captures both modalities $\Rightarrow$ high degree of multimodal fusion

visual and textual embeddings



Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

- Multimodal Fusion Degree
  - Extract multimodal embeddings of all tokens and image regions
  - Cluster into two sets
  - Measure difference of clusters and ground-truth visual / textual clusters (Normalised Mutual Information)

$$MI(x, y) = P(x, y) \log \frac{P(x,y)}{P(x)P(x)}$$

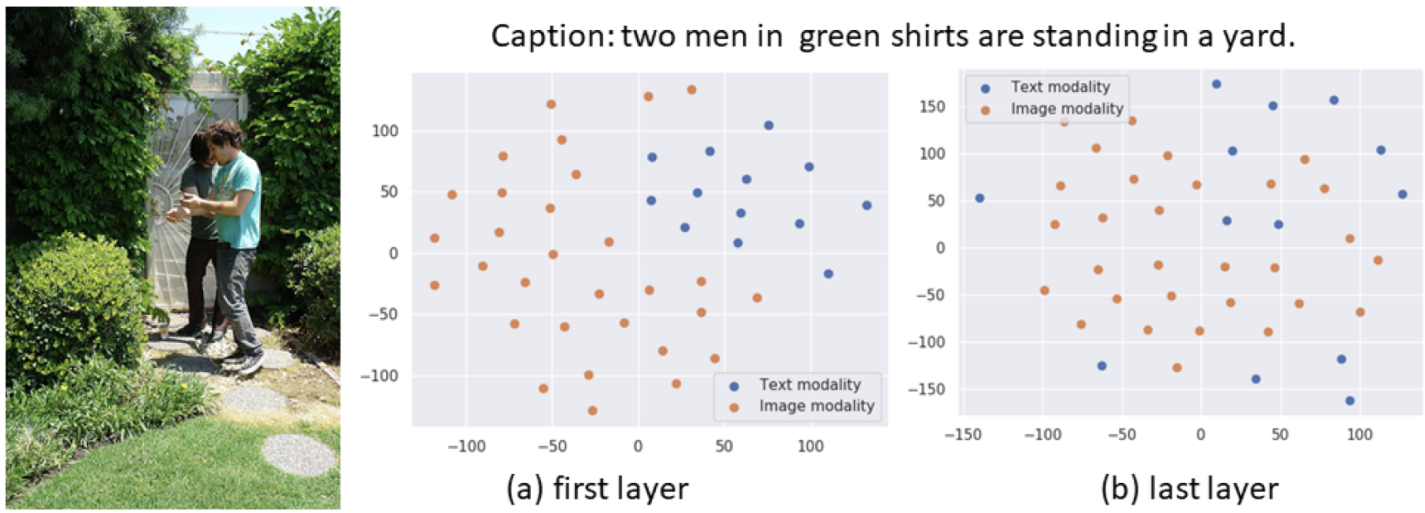


Fig. 6: An t-SNE visualization of multimodal fusion degree of the first and last layer of UNITER over one image-text pair. Each yellow and blue dot corresponds to a visual and textual token, respectively.

Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models [Cao et al., 2020]

- Multimodal Fusion Degree
  - Extract multimodal embeddings of all tokens and image regions
  - Cluster into two sets
  - Measure difference of clusters and ground-truth visual / textual clusters (Normalised Mutual Information)

$$MI(x, y) = P(x, y) \log \frac{P(x,y)}{P(x)P(x)}$$

Layer	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
single-stream												
Flickr30k	0.36	0.38	0.39	0.41	0.38	0.38	0.38	0.38	0.32	**0.20**	0.26	**0.20**
Visual Genome	0.25	0.25	0.24	0.24	0.22	0.22	0.21	0.21	0.20	0.17	**0.16**	**0.16**
two-stream (cross)												
Flickr30k	**0.42**	0.48	0.67	0.75	0.43	–	–	–	–	–	–	–
Visual Genome	**0.43**	0.56	0.68	0.78	0.57	–	–	–	–	–	–	–

Table 1: NMI scores on multimodal fusion probing. A smaller NMI value indicates a higher fusion degree. Note that the two-stream model (LXMERT) only has 5 layers in its cross-modality encoder. A larger layer number corresponds to an upper layer.

Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

Overview

- VL decoders
  - Can generate text given an image and text
  - $\Rightarrow$ Can produce predictions *and* explanations
- *How much is each modality used when generating predictions vs. explanations?*
- *How self-consistent are the models in their explanations?*

Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

Overview

- VL decoders
  - $\Rightarrow$ Can produce predictions *and* explanations
- ***How much is each modality used when generating predictions vs. explanations?***

MM-SHAP for multimodal degree [Parcalabescu and Frank, 2023]

- Determines the contribution of each modality in VL encoders:
- Computes Shapley values for each input token, i.e., the contribution of each input token to each generated token
- Normalises and aggregates them modality-wise
- Textual degree - T-SHAP, visual degree - V-SHAP (proportion of modality contributions)

Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

Overview

- VL decoders
  - $\Rightarrow$ Can produce predictions *and* explanations
- *How much is each modality used when generating predictions vs. explanations?*
- ***How self-consistent are the models in their explanations?***

CC-SHAP for self-consistency [Parcalabescu and Frank, 2024b]

- Determines the contribution of each input token to the *prediction*:
  Compares input contributions when generating answers to input contributions when generating explanation

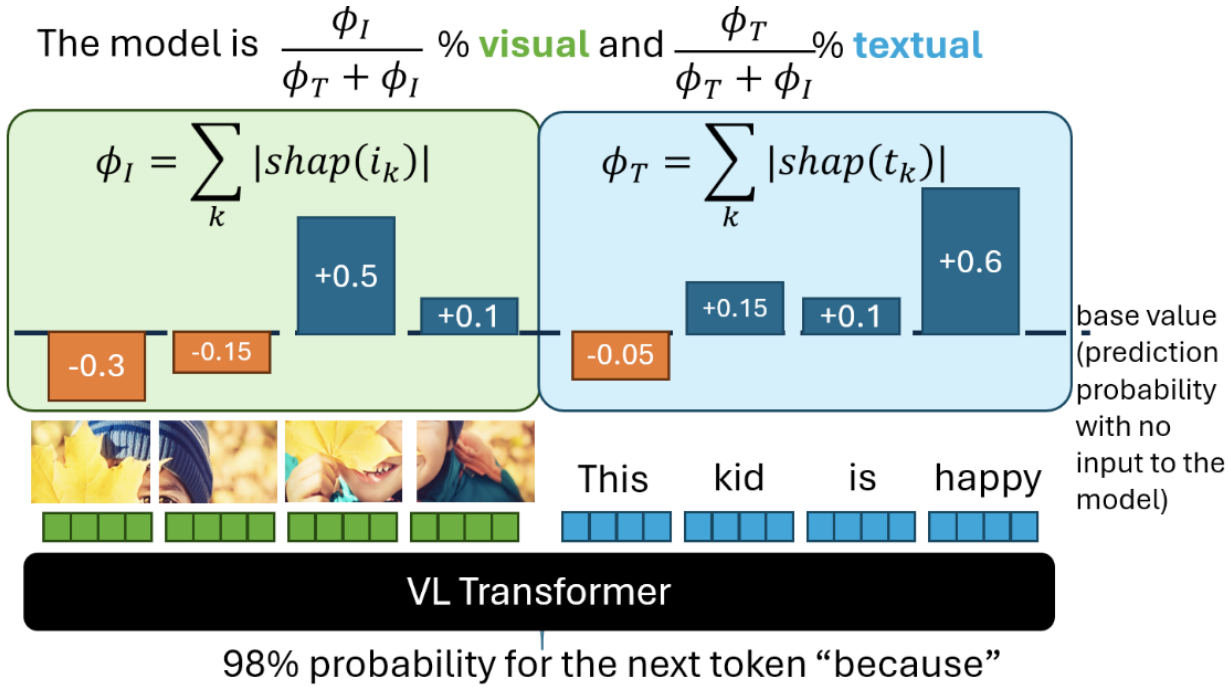
Explanation Self-Consistency Tests

- Compares model outputs before and after input edits

Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

MM-SHAP



Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

Experimental Setup

- Three LLaVA models with strong LLMs
- VALSE benchmark and VQA benchmarks

Results:

- **Prediction**: Text modalities predominating during *prediction*: T-SHAP $> 89\%$
- **Explanation generation:** Visual modalities more influential compared to prediction

$\Rightarrow$ Misalignment between contributions during prediction and explanation — model inconsistency

- Linguistic abilities: decoders still struggle, especially on counting, but less on nouns and existence

Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

Results on VALSE

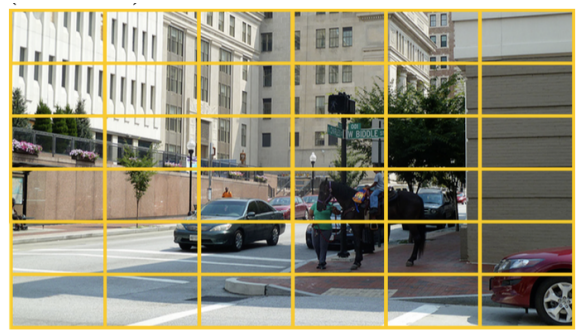
Metric	Model	Existence quantifiers	Plurality number	Counting bal.†	Counting sns.†	Counting adv.†	Sp.rel.‡ relations	Action repl.†	Action swap†	Coreference std.†	Coreference clean	Foil-it nouns
acc_r	BakLLaVA	92	78	77	80	**73**	**84**	**89**	82	78	78	**98**
	LV-Mistral	**96**	**81**	**78**	**82**	67	79	**89**	**90**	**84**	**84**	**98**
	LV-Vicuna	88	71	73	76	59	73	86	88	83	78	96
acc	BakLLaVA	50	50	50	50	**50**	50	50	50	50	50	50
	LV-Mistral	62	**55**	51	50	**50**	**58**	54	54	**57**	**59**	66
	LV-Vicuna	**78**	52	**59**	**60**	48	52	**64**	**62**	55	52	**68**
p_c	BakLLaVA	0	0	0	0	0	0	0	0	0	0	0
	LV-Mistral	24	17	1	1	0	30	9	9	29	39	34
	LV-Vicuna	**64**	**98**	**26**	**35**	10	**99**	**82**	**83**	**97**	**95**	**96**
p_f	BakLLaVA	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
	LV-Mistral	**100**	94	**100**	**100**	**100**	86	98	99	86	78	98
	LV-Vicuna	92	6	91	86	85	4	46	42	13	9	40

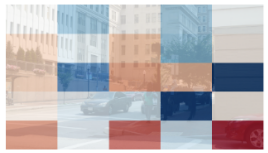
Table 1: Performance (cf. §4.2) of three **VL decoders** on *all samples* of VALSE. We bold-face the b result per metric, and highlight with red results at or below random baseline (50%). The scores in the la are compared in Fig. 1 to VLM encoders and unimodal models. **LV-***: LLaVA-NeXT-*. †**bal.** Counting †**sns.** Counting small numbers. **adv.** Counting adversarial. **repl.** Action replacement. **swap.** Actant swap Spatial relations. †**std.** Coreference standard. **Avg.** ± **SD**: Average over rows and standard deviation.

Analysis: VL decoders and their multimodal degree

[Parcalabescu and Frank, 2024a]

Prediction vs. Explanation (post-hoc)



	Model Prediction		Model Explanation	
	V-SHAP: 3%	T-SHAP: 97%	V-SHAP 40%	T-SHAP: 60% CC-SHAP: 0.22
LLaVA-NeXT-Mistral		[INST]: <image> Where is the horse? [/INST] The best short answer is: The horse is on a street		[INST]: <image> Where is the horse? [/INST] The best short answer is: The horse is on a street [INST] Why? Please explain how you arrived at your answer. [/INST] Explanation: In the image provided, there is a person leading a horse on a city street, most likely navigating through the busy urban environment for a specific purpose, perhaps a special event or a cultural activity that incorporates horse riding. The horse is being guided by its handler across the crosswalk with a person on its back, which could indicate it's being ridden, possibly for ceremonial reasons, tourist attractions, or local customs.
		[INST]: <image> Where is the horse? [/INST] The best short answer is: The horse is on a street		[INST]: <image> Where is the horse? [/INST] The best short answer is: The horse is on a street [INST] Why? Please explain how you arrived at your answer. [/INST] Explanation: In the image provided, there is a person leading a horse on a city street, most likely navigating through the busy urban environment for a specific purpose, perhaps a special event or a cultural activity that incorporates horse riding. The horse is being guided by its handler across the crosswalk with a person on its back, which could indicate it's being ridden, possibly for ceremonial reasons, tourist attractions, or local customs.

Outline

ViLMA: Video–Language Model Assessment

A zero-shot benchmark for linguistic and temporal grounding in video–language models [Kesen et al., 2023]

Overview

- Videos can account for the temporal dimension, for dynamically evolving phenomena (events, actions, physical processes)
- Task-agnostic probing benchmark focusing on temporal understanding and reasoning
- 5177 **Counterfactuals** (foils), each paired with a **proficiency test** (considered as prerequisite to solve the main tests)
- Finding: Grounding abilities of video–language models (VidLMs) not better than vision–language models with still images (VLMs)

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Foils (Counterfactuals)

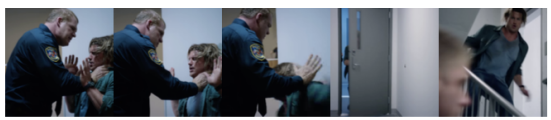
- Created from existing datasets and (V)LLMs using, e.g., Masked Language Modelling
- Validated using NLP tools and manual validation via Amazon Mechanical Turk

Proficiency tests

- Less challenging than the main tests, they do not require temporal modelling
- Bias indicator: Does the model succeed in main test but fail in proficiency test?
- Manually validated

ViLMA: Video–Language Model Assessment

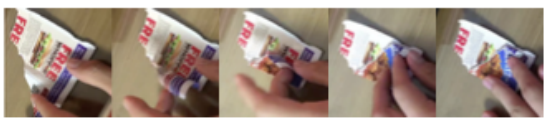
[Kesen et al., 2023]

Test (#exs.)	Video Caption (blue) / Foil (orange)	Foil Generation	Sample Frames
Situation Awareness (911)	A policeman / blond man holds a blond man / policeman against a wall.	Actor swapping	
	A man in blue holds / chops up a man in green.	Action replacement	
Spatial Relations (393)	Moving steel glass towards / from the camera.	Relation replacement	

Proficiency test: Object identification

ViLMA: Video–Language Model Assessment

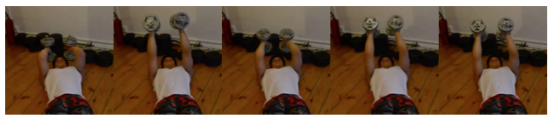
[Kesen et al., 2023]

Test (#exs.)	Video Caption (blue) / Foil (orange)	Foil Generation	Sample Frames
Change of State (998)	Someone folds / unfolds the paper.	Action replacement	
	Initially, the paper is unfolded / folded.	Pre-state replacement	
	At the end, the paper is folded / unfolded.	Post-state replacement	
	Initially, the paper is unfolded / folded. Then, someone folds / unfolds the paper. At the end, the paper is folded / unfolded.	Swap-and-replacement	

Proficiency test: Object identification

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Test (#exs.)	Video Caption (blue) / Foil (orange)	Foil Generation	Sample Frames
Action Counting (1432)	Someone lifts weights exactly two / five times.	Number replacement	
Rare Actions (1443)	Drilling into / Calling on a phone.	Action replacement	
	Drilling into a phone / wall.	Object replacement	

Do VidLMs identify novel compositions or recognise unusual human–object interactions?

Proficiency test: Action recognition & Object existence

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Action Counting

Proficiency Test: a man skips / climbs a rope.
Main Test: a man skips rope exactly three / nine times.

Proficiency Test: someone peels a melon / lemon.
Main Test: someone peels a melon in exactly two / five moves.

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Situation Awareness: Action replacement / Actor swapping

Proficiency Test: A shirtless man opens the window / door hurriedly.
Main Test: A shirtless man opens / smashes the window hurriedly.

Proficiency Test: The woman in red looks upward at the man in a hat / wheelchair.
Main Test: The woman in red / man in a hat looks upward at the man in a hat / woman in red.

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Change of State: Post state / Reverse

Proficiency Test: Someone connects the chair and the base / dots.
Main Test: At the end, the chair and the base is attached / detached.

Proficiency Test: Someone chops an onion / apple.
Main Test: Initially, an onion is whole / in pieces. Then, someone chops / connects an onion. At the end, an onion is in pieces / whole.

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Rare Actions (Action replacement)

Proficiency Test: there is at least one banana / blender
Main Test: weighing / eating a banana

ViLMA: Video–Language Model Assessment

[Kesen et al., 2023]

Comparison: Unimodal, VL and VidL models

Model	Action Counting			Situ. Awareness			Change of State			Rare Actions			Spatial Relations			Avg.
	P	T	P+T	P	T	P+T	P	T	P+T	P	T	P+T	P	T	P+T	P+T
Random	50.0	50.0	25.0	50.0	37.9	18.9	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	23.8
GPT-2†	50.3	53.3	27.6	44.5	66.6	31.7	18.0	52.4	10.8	58.4	25.9	17.7	49.1	72.8	43.0	26.2
OPT†	56.2	54.6	31.0	51.7	71.3	38.7	23.1	48.0	12.9	59.0	23.9	14.9	59.0	84.7	55.7	30.6
CLIP‡	90.5	50.9	46.2	71.0	45.5	33.6	93.0	**55.2**	**52.2**	92.7	93.9	87.8	78.6	58.3	44.8	52.9
BLIP2‡	80.9	54.5	43.7	73.4	**75.4**	**55.7**	74.5	52.1	38.1	93.8	74.5	70.5	**91.1**	**86.0**	**79.4**	**57.5**
ClipBERT	56.4	49.6	28.0	54.1	56.9	31.9	63.7	50.0	33.5	43.5	40.7	17.7	39.7	39.8	14.1	25.0
UniVL	73.4	43.6	32.2	51.6	46.6	24.1	81.3	54.3	43.0	77.5	78.0	59.9	62.5	51.7	33.2	38.5
VideoCLIP	79.1	46.4	36.5	61.6	40.3	24.9	49.8	50.8	25.9	84.0	77.8	67.5	67.9	54.7	39.7	38.9
FiT	83.9	52.4	44.6	69.8	40.0	29.1	93.0	52.1	47.8	89.7	89.4	80.7	70.5	51.9	38.7	48.2
CLIP4Clip	**91.2**	52.3	**48.0**	**73.8**	49.0	37.6	**94.8**	54.1	52.1	83.0	**94.1**	78.7	79.8	56.7	44.2	52.1
VIOLET	79.6	50.6	36.5	70.2	41.6	32.4	88.2	54.6	49.1	87.1	86.6	74.6	73.3	50.4	38.7	46.3
X-CLIP	84.1	55.1	46.4	63.5	44.8	31.0	85.7	52.7	46.0	83.9	85.7	72.3	74.8	56.2	43.5	47.8
MCQ	81.4	50.4	41.5	67.0	37.1	26.3	90.3	50.3	45.3	91.3	88.7	82.3	79.4	48.9	39.4	47.0
Singularity	79.6	51.1	41.5	68.8	40.9	30.1	92.8	54.6	50.3	92.7	88.4	83.1	80.7	46.8	38.9	48.8
UniPerceiver	50.6	46.4	23.0	51.4	42.1	21.1	67.5	46.1	29.1	58.2	58.8	34.7	45.5	48.0	20.1	25.6
Merlot Reserve	84.2	**56.0**	46.9	70.5	35.6	25.3	93.4	53.6	50.4	83.8	90.6	77.6	63.1	41.9	29.2	45.9
VindLU	84.5	51.2	43.5	70.5	41.6	31.2	85.4	52.6	45.6	**94.2**	93.1	**88.0**	83.2	45.6	39.4	49.5

- Temporal reasoning capabilities of VidLMs limited – VLMs and VidLMs mostly comparable

Outline

Take-away Messages

- Innovations in models and training losses more beneficial than scaling up training data
- Rich data sources are relevant
- Two key ingredients for models: contrastive learning and cross-modal attention

References I

- CV-Probes: Studying the interplay of lexical and world knowledge in visually grounded verb understanding [Beňová et al., 2024a]
- Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models [Srinivasan and Bisk, 2022]
- Measuring Progress in Fine-grained Vision-and-Language Understanding [Bugliarello et al., 2023]
- ViLMA: A zero-shot benchmark for linguistic and temporal grounding in video–language models [Kesen et al., 2023]
- MAEA: Multimodal Attribution for Embodied AI [Jain et al.,]
- Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality [Thrush et al., 2022]
- VALSE [Parcalabescu et al., 2022]
- [Dönmez et al., 2023]

References II

- Do Vision & Language Decoders use Images and Text equally? How Self-consistent are their Explanations? [Parcalabescu and Frank, 2024a]
- How and where does CLIP process negation? [Quantmeyer et al., 2024]
- Controlling for Stereotypes in Multimodal Language Model Evaluation [Malik and Johansson, 2022]
- Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation [Wan et al., 2024]
- VerbCLIP: Improving Verb Understanding in Vision-Language Models with Compositional Structures [Wazni et al., 2024]
- Cross-Modal Coherence Relations Inform Eye-gaze Patterns During Comprehension & Production [Inan and Alikhani, 2024]
- Visual Spatial Reasoning [Liu et al., 2023]

References III

- VL-CheckList: An Explainable Toolbox for Evaluating Pre-trained Vision-Language Models [Zhao et al., 2022]
- MM-Shap [Parcalabescu and Frank, 2023]
- COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images [Bogin et al., 2021]
- Understanding Cross-modal Interactions in V&L Models that Generate Scene Descriptions [Cafagna et al., 2022]
- Beyond Image-Text Matching: Verb Understanding in Multimodal Transformers Using Guided Masking [Beňová et al., 2024b]
  `https://github.com/ivana-13/guided_masking`
- VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations
- CREPE [Ma et al., 2023]

References IV

- A closer look at referring expressions for video object segmentation [Bellver et al., 2023]
- Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers [Frank et al., 2021]
- What Does BERT with Vision Look At? [Li et al., 2020]
- Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective [Salin et al., 2022]
- VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers [Lal et al., 2021]
  `https://github.com/IntelLabs/VL-InterpreT`

References I

Bellver, M., Ventura, C., Silberer, C., Kazakos, I., Torres, J., and Giro-i Nieto, X. (2023).

A closer look at referring expressions for video object segmentation.
Multimedia Tools and Applications, 82(3):4419–4438.

Beňová, I., Gregor, M., and Gatt, A. (2024a).

Cv-probes: Studying the interplay of lexical and world knowledge in visually grounded verb understanding.
arXiv e-prints, pages arXiv–2409.

Beňová, I., Košecká, J., Gregor, M., Tamajka, M., Veselỳ, M., and Šimko, M. (2024b).

Beyond image-text matching: Verb understanding in multimodal transformers using guided masking.
arXiv preprint arXiv:2401.16575.

Bogin, B., Gupta, S., Gardner, M., and Berant, J. (2021).

COVR: A test-bed for visually grounded compositional generalization with real images.
In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9824–9846, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bugliarello, E., Sartran, L., Agrawal, A., Hendricks, L. A., and Nematzadeh, A. (2023).

Measuring progress in fine-grained vision-and-language understanding.
In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582.

Cafagna, M., van Deemter, K., and Gatt, A. (2022).

Understanding cross-modal interactions in V&L models that generate scene descriptions.
In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 56–72.

References II

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., and Liu, J. (2020).

Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models.

In *Proceedings of the 2020 European Conference on Computer Vision (ECCV 2020)*.

Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., and Xu, B. (2022).

Vlp: A survey on vision-language pre-training.

arXiv preprint arXiv:2202.09061.

Dönmez, E., Tilli, P., Yang, H.-Y., Vu, N. T., and Silberer, C. (2023).

Hnc: Leveraging hard negative captions towards models with fine-grained visual-linguistic comprehension capabilities.

In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 364–388.

Frank, S., Bugliarello, E., and Elliott, D. (2021).

Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers.

In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Inan, M. and Alikhani, M. (2024).

Seeing eye-to-eye: Cross-modal coherence relations inform eye-gaze patterns during comprehension & production.

In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14494–14512.

Jain, V., Yerramilli, S., Tamarapalli, J. S., and Bisk, Y.

Maea: Multimodal attribution for embodied ai.

In *Progress and Challenges in Building Trustworthy Embodied AI*.

References III

Kesen, I., Pedrotti, A., Dogan, M., Cafagna, M., Acikgoz, E. C., Parcalabescu, L., Calixto, I., Frank, A., Gatt, A., Erdem, A., and Erdem, E. (2023).
Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models.

Lal, V., Ma, A., Aflalo, E., Howard, P., Simoes, A., Korat, D., Pereg, O., Singer, G., and Wasserblat, M. (2021).
InterpreT: An interactive visualization tool for interpreting transformers.
In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 135–142, Online. Association for Computational Linguistics.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2020).
What does BERT with vision look at?
In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Liu, F., Emerson, G., and Collier, N. (2023).
Visual spatial reasoning.
Transactions of the Association for Computational Linguistics, 11:635–651.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. (2023).
Crepe: Can vision-language foundation models reason compositionally?
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921.

Malik, M. and Johansson, R. (2022).
Controlling for stereotypes in multimodal language model evaluation.
In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 263–271.

References IV

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., and Gatt, A. (2022).

VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena.

In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Parcalabescu, L. and Frank, A. (2023).

MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks.

In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Parcalabescu, L. and Frank, A. (2024a).

Do Vision & Language Decoders use Images and Text equally? How Self-consistent are their Explanations?

arXiv preprint arXiv:2404.18624.

Parcalabescu, L. and Frank, A. (2024b).

On measuring faithfulness or self-consistency of natural language explanations.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.

Quantmeyer, V., Mosteiro, P., and Gatt, A. (2024).

How and where does CLIP process negation?

In Gu, J., Fu, T.-J. R., Hudson, D., Celikyilmaz, A., and Wang, W., editors, *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 59–72, Bangkok, Thailand. Association for Computational Linguistics.

References V

Salin, E., Farah, B., Ayache, S., and Favre, B. (2022).

Are vision-language transformers learning multimodal representations? a probing perspective.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.

Srinivasan, T. and Bisk, Y. (2022).

Worst of both worlds: Biases compound in pre-trained vision-and-language models.
In Hardmeier, C., Basta, C., Costa-jussà, M. R., Stanovsky, G., and Gonen, H., editors, *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022).

Winoground: Probing vision and language models for visio-linguistic compositionality.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R., and Chang, K.-W. (2024).

Survey of bias in text-to-image generation: Definition, evaluation, and mitigation.
arXiv preprint arXiv:2404.01030.

Wazni, H., Lo, K. I., and Sadrzadeh, M. (2024).

VerbCLIP: Improving verb understanding in vision-language models with compositional structures.

In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 195–201.

Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., and Yin, J. (2022).

An explainable toolbox for evaluating pre-trained vision-language models.
In Che, W. and Shutova, E., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE. Association for Computational Linguistics.