Benchmarking Video-Language Models to understand their grounding capabilities

Joint work with:

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna,

Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto,

Anette Frank, Aykut Erdem, Erkut Erdem

Work under review for ICLR 2024

Counterfactual probing using foils

Introduced in V&L research by Shekhar et al (2017) with the FOIL-It task.

Given an image and **caption** pair, change a word in the caption to create a **foil**.

Is the model able to distinguish the two?

People riding bicycles down a road approaching a ***bird.***

People riding bicycles down a road approaching a ***dog.***

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., & Bernardi, R. (2017). FOIL it! Find One mismatch between Image and Language caption. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)* (pp. 255–265).

Madhyastha, P., Wang, J., & Specia, L. (2018). Defoiling Foiled Image Captions. *Proceedings Of the Conference of the North American Chapter Of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18)*, 433–438. http://arxiv.org/abs/1805.06549

VALSE: Vision And Language Structured Evaluation

A foil-based benchmark focusing on linguistic phenomena

VALSE:

- Targets pretrained V&L models.
- Intended as a zero-shot evaluation benchmark.
- (V&L models have pretrained image-caption alignment heads.)
- Targets specific linguistic features.
- Designed to ensure validity & reliability, especially by mitigating **distributional** and **plausibility** bias.

L Parcalabescu, M Cafagna, L Muradjan, A Frank, I Calixto, and A Gatt (2022). VALSE: A Task-independent benchmark for Vision and Language models centered on linguistic phenomena. (2021). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL'22),*

Meet ViLMA...
(Video-Language Model Assessment)

A successor to VALSE. This time we're waltzing with video.

If we model the interaction between language and vision using videos, then we also have a better grasp of the **temporal dimension**.

Temporal information features strongly in our interpretation of certain linguistic constructions and the inferences we draw from them.

E.g. *changes of state:* X <u>opened</u> the door

E.g. *iterativity*: I knocked <u>several/two</u> times

E.g. *spatio-temporal relations:* He kicked the ball <u>towards</u> the net

We use a counterfactual setup

- Start from a video clip + caption (from an existing dataset)
  - Sometimes captions are generated from templates
- Replace something in the caption which:
  - Makes the caption no longer true wrt the video
  - Is related to a phenomenon of interest (e.g. change of state)
- Every single item in this benchmark is validated with human judges
- We target pretrained Video-Language models and use them in a zero-shot setting (usually, by exploiting their pretrained video/image-text alignment head)

Tests (and subtests) in ViLMA

Test (#exs.)	**Video Caption (blue) / Foil (orange)**	**Foil Generation**	**Sample Frames**
Action Counting (1432)	Someone lifts weights exactly two / five times.	Number replacement	
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	
Change of State (998)	Someone folds the paper / laundry.	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	
Rare Actions (1443)	Drilling into / Calling on a phone.	Action replacement	
	Drilling into a phone / wall.	Object replacement	
Spatial Relations (393)	Moving steel glass towards / from the camera.	Relation replacement	

Proficiency tests

Key idea:

For a model to "solve" a ViLMA test and demonstrate real grounding abilities, it must first have certain "basic" abilities, which are simpler.

Each ViLMA test item has a corresponding "proficiency test" item, which tests for the basic ability. (NB: they are one-to-one)

If the model fails the basic PT and succeeds on the main T, this is likely due to spurious features.

Example:

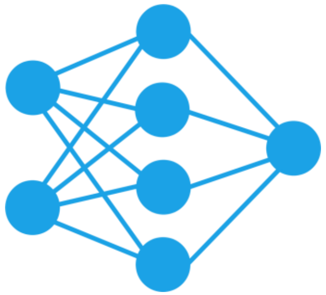
To successfully count the occurrences of an action, the model needs to recognise the event itself. (= Event recognition)

To successfully determine that a door was opened and is therefore no longer closed, the model needs to be able to recognise the door. (= Object recognition)

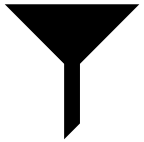
Our testing rationale
(example from situation awareness test)

Proficiency test
C_P: A shirtless man opens the **window** hurriedly
F_P: A shirtless man opens the **door** hurriedly

Main test
C: A shirtless man **opens** the window hurriedly
F: A shirtless man **smashes** the window hurriedly



P-score:
Proportion of cases where:
P(match|v, C_P) > P(match|v, F_P)



Filter for the final, combined score: P+T

T(est)-score:
Proportion of cases where:
P(match|v, C) > P(match|v, F)

We tested a gazillion models

- Random baseline (50%)
- Unimodal models (GPT-2, OPT)
    - *To what extent is the caption/foil distinction solvable based on textual features only?*
- Image-text models (CLIP, BLIP-2)
    - *Offer a baseline against which to compare grounding with and without an explicit temporal dimension*.
- Video-language models
    - ClipBERT, UniVL, VideoCLIP, FiT, CLIP4CLIP, VIOLET, X-CLIP, MCQ, Singularity, UniPerceiver, Merlot Reserve, VindLU
    - *Several architectures and a variety of pretraining objectives (out of scope for today)*

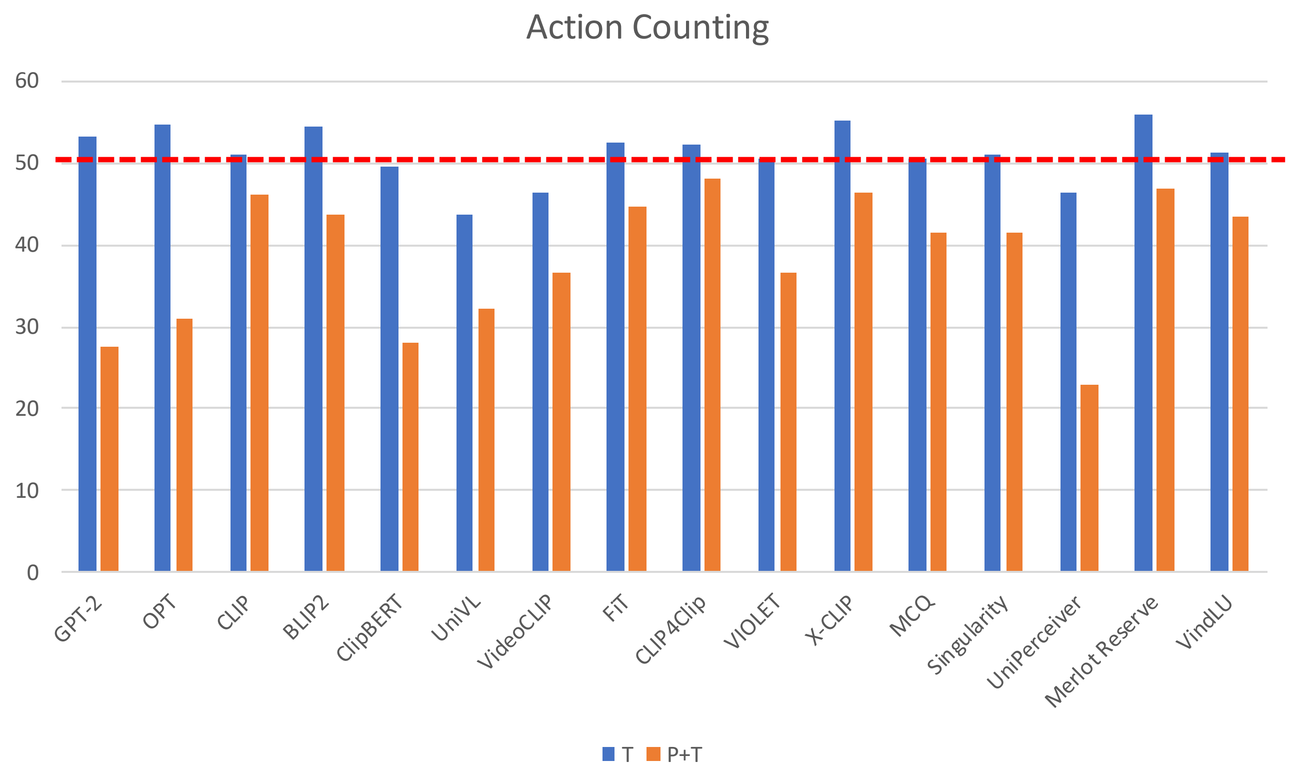
Results (selection)

Performance poor even without taking proficiency into account.

Video-Language models are not notably better than models using static images!

Confirms previous results with image-text models.

Counting is hard!

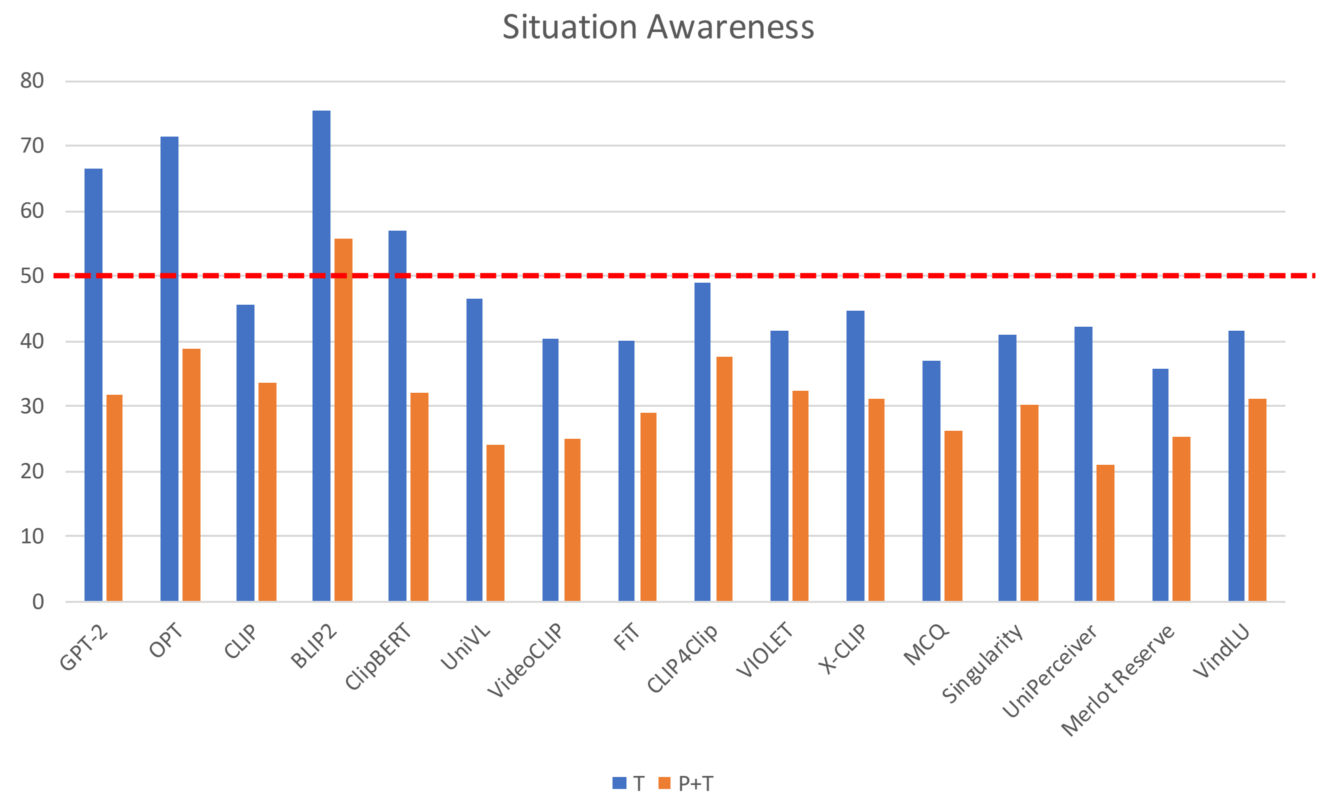


Results (selection)

Unimodal models do better overall. Could be because swapping actors or verbs results in “surprising” foils.

Again, BLIP-2 (static) outperforms a lot of the Video-Language models.

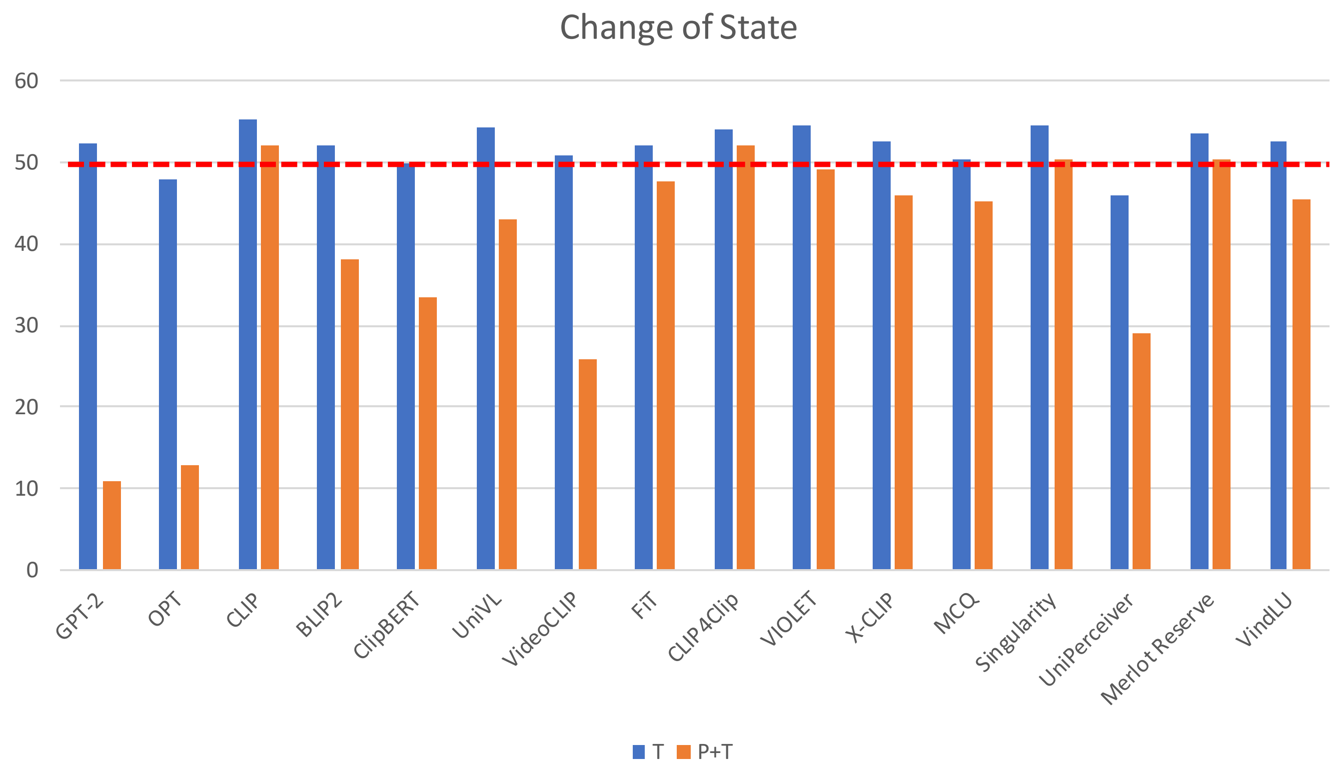
The temporal info in videos isn’t helping.



Results (selection)

Models are barely above chance.

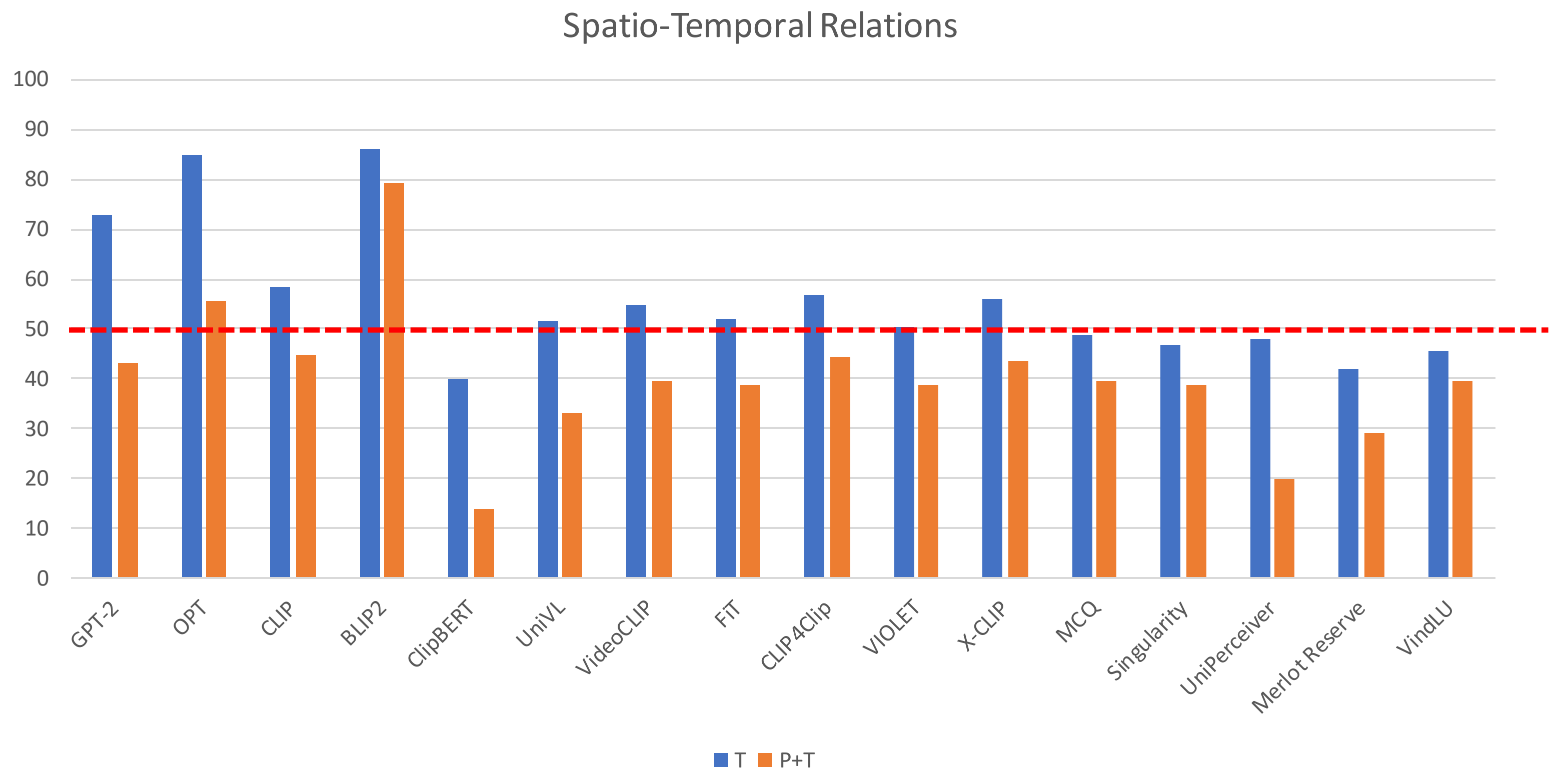
This is one test where temporal info should matter a lot.



Results (selection)

Only static models stay above chance once P is taken into account.

Video-language models do not seem to gain an advantage from the temporal info.



Main observation

Proficiency / main task

- Big drops suggest a reliance on spurious features in the training data.

Video vs. Image

- No clear differences in many tasks – the temporal information in videos doesn't necessarily help.
- (But could also be an impact of pretraining data size etc).

Main takeaway

- The training objectives used do not guarantee that models ground language in video data, making full use of the temporal dimension in that data.