Frege in the (Residual) Stream

Sense and Reference inside Language and Vision–Language Models

Denis Paperno

Utrecht University

ESSLLI 2026 workshop *Referring expression choice in grounded contexts: Linguistic, cognitive, and computational aspects*

Where this talk is going

1. A referent can be presented in many ways. Frege gave us the vocabulary: **Sinn** (*sense*) vs. **Bedeutung** (*reference*).
2. Classical Referring Expression Generation algorithm (Dale & Reiter 1995) runs to completion before any expression even starts being generated: it is **a procedure for choosing a *sense***.
3. Neural models appear to skip that step: text (and image) in, text out. There is no explicit "sense" (or "reference").
4. **The core of this talk:** interpretability suggests that models encode the whole Frege triangle: **forms**, **concepts**, and **referential information**; the last might not even be named specifically in either input or output.
5. A case study where referential and conceptual information are **separately decoded and separately manipulated**.
6. Future direction: What we can now ask empirically.

[I will assume Frege and REG and go slowly on the interpretability. Please interrupt.]

Part I. Sense and reference

Frege's observations

the morning star and _the evening star_ mean the same thing

The morning star circles the Sun and _The evening star circles the Sun_ must be simultaneously true (or simultaneously false)

Frege's observations

the morning star and *the evening star* mean the same thing

The morning star circles the Sun and *The evening star circles the Sun*
must be simultaneously true (or simultaneously false)

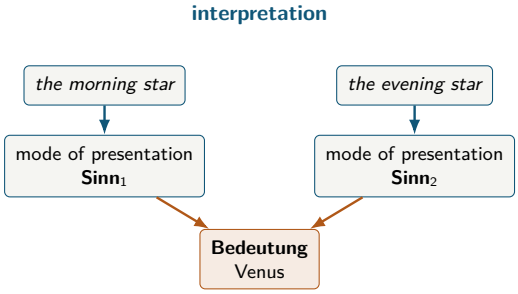
the morning star and *the evening star* DON'T mean the same thing

The morning star is the evening star and
The morning star is the morning star
are not the same statement.

Frege's triangle

Frege's observations

the morning star and *the evening star* mean the same thing
the morning star and *the evening star* don't mean the same thing

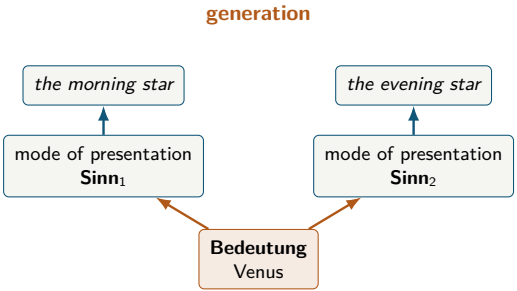


[Frege 1892, *Über Sinn und Bedeutung*]

Frege's triangle

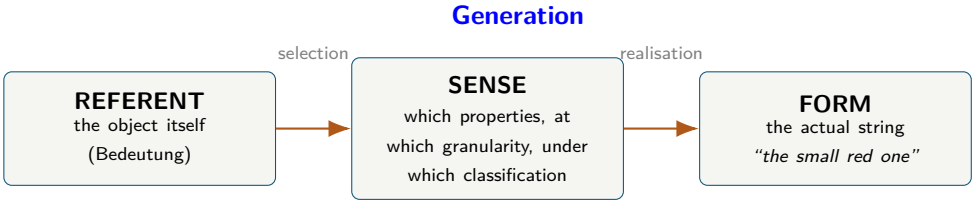
Frege's observations

the morning star and *the evening star* mean the same thing
the morning star and *the evening star* don't mean the same thing



[Frege 1892, *Über Sinn und Bedeutung*]

What we take forward



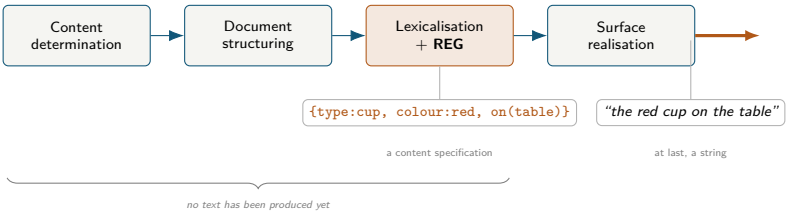
The question for the rest of the talk

Is this three-way split a description of *our theory of language*, or also of *computational models like LLMs*? And if the latter, in what order do the boxes get filled?

Part II. Referring Expression Generation

one substantive claim: this literature is about senses, not expressions

“Referring Expression Generation” generates nothing



Reiter & Dale (2000) define REG as the “selection of words and phrases to identify domain entities”, and then immediately set it apart from lexicalisation as a different kind of task. Lexicalisation picks words for concepts; REG must **discriminate** between descriptions singling one entity out from its distractors. It is a discrimination task, not a wording task.

The properly generative step (producing text that is well formed syntactically, morphologically and orthographically) is **surface realisation**, further downstream.

So the classical pipeline already has a dedicated module for *choosing* a sense. It just calls it “generation”.

Dale & Reiter (1995): three algorithms, and why the third won

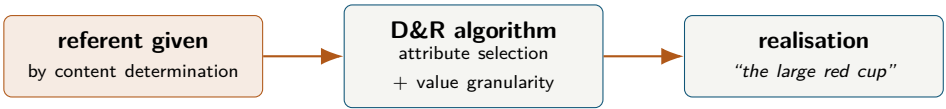
["Computational interpretations of the Gricean maxims in the generation of referring expressions", *Cognitive Science* 19(2)]

- **Full Brevity.** Shortest distinguishing description. Faithful to Quantity; NP-hard; empirically wrong about people.
- **Greedy Heuristic.** Repeatedly add whatever rules out most distractors. Cheaper, still purely discriminative.
- **Incremental Algorithm.** Walk a **fixed preference order** over attributes; include one iff it rules out at least one distractor; stop when $C = \emptyset$. **No backtracking.** Always include the type.

Efficient (linear time) and **predicts human overspecification**, the redundant colour and size that a brevity-optimal speaker would never produce.

Two of the three guiding principles are not about discriminability: the preference order, and granularity policy (basically, prefer basic-level values like *dog* vs *poodle* or *mammal*).

The Fregean reading

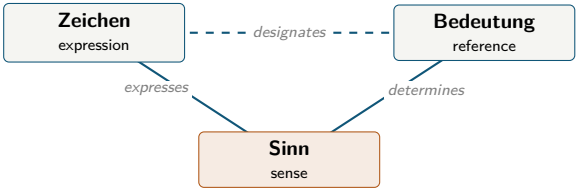


- The algorithm terminates **before any referring expression exists**.
- Its output is a mode of presentation of an already-fixed referent.
- The parameters of REG (**which properties**, **at what taxonomic level**, **under which classification**, **anchored to which relatum**) = where “modes of presentation” (senses compatible with one referent) differ.

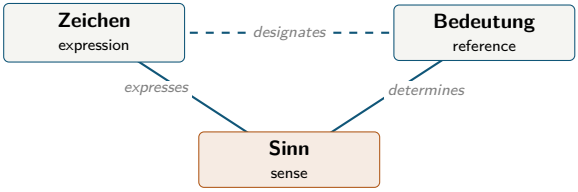
Classical generation pipeline: referent first, sense second, form third.

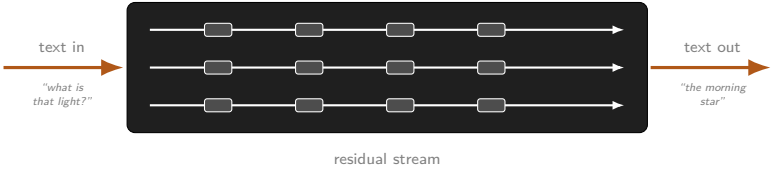
The question: How abour LLMs/VLMs?

The question: How abour LLMs/VLMs?

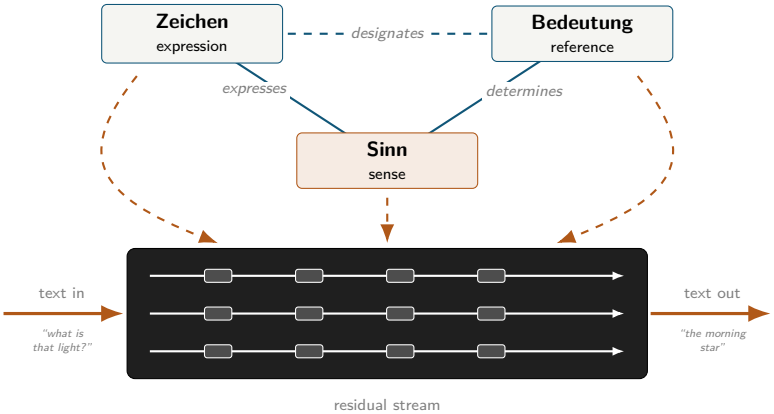


The question: How abour LLMs/VLMs?



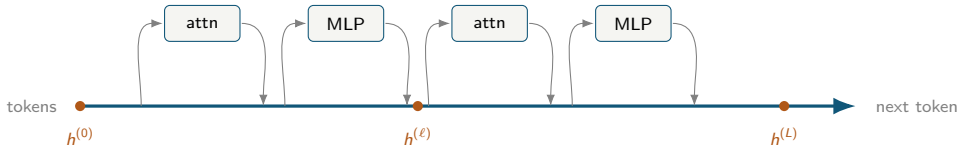


The question: How abour LLMs/VLMs?



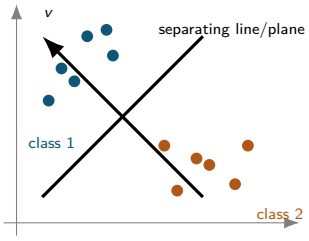
Part III. What is legible inside a model

What a transformer keeps around



- At every token position the model carries a vector $h^{(\ell)} \in \mathbb{R}^d$ (typically $d \approx 4000$).
- Each layer *reads* from it and *adds* to it. Nothing is overwritten: it is a running sum.
- So the stream is a **working memory that is fully observable from outside**. Every intermediate state can be inspected, and, crucially, edited.
- "Interpretability" means:
  - what can we recover from $h^{(\ell)}$
  - what happens if we change it?

What “a direction” means



- A **feature** is claimed to be a *direction*: a unit vector v such that the projection $\langle h, v \rangle$ tracks some property.
- A **probe** is a classifier fitted to find such a v from labelled examples.
- A successful probe shows only that the information is *linearly separable in the state*. It does not show the model uses it.
- Hence the two extra moves:
  - **steer**: set $h \leftarrow h + \alpha v$ and see if behaviour changes as predicted;
  - **control**: check that a random direction of the same length does *not* do it.

[The linear framing is a working assumption, see Park et al. 2024a; Elhage et al. 2022.]

The instruments, in plain terms

Reading: is the information there?

probing classifier	can a simple classifier recover property π from $h^{(\ell)}$?
activation difference	estimate direction as $h^{(\ell)}_{C1} - h^{(\ell)}_{C2}$ for mean activations of examples of classes $C1, C2$
logit lens	if we stop at layer ℓ and decode straight to vocabulary, what word do we get?
sparse autoencoder	decompose h into a large sparse set of reusable, inspectable features

Intervening: is it used as we may hypothesise?

activation patching	run on input A, run on input B, then *transplant* A's state at one layer and position into B's run. Did the answer move toward A's?
steering	add a direction to the state and watch the output change
ablation	remove a component and see what changes

Three claims, in increasing order of interest

Claim A: **Forms** are recoverable

Surface, lexical and orthographic facts about the tokens read and the tokens to come.

Claim B: **Concepts** are recoverable

Category membership, taxonomic structure, abstract features that are not any single token.

Claim C: **Referents' properties** are recoverable

Information about entities that appears neither in the input string nor in the output string.

Claim A: forms are recoverable

- **Trivially true at the endpoints.** Tokens are input and output; the embedding and unembedding matrices are literally form dictionaries.
- **Non-trivially true in between.** The logit lens shows the stream can be decoded into vocabulary at *every* layer; feed-forward layers act by promoting vocabulary items [nostalgebraist 2020; Geva et al. 2022; Belrose et al. 2023].

In a nutshell

Forms are represented **before they are emitted**. A single hidden state predicts tokens several positions ahead [Pal et al. 2023]; models select a rhyme word before generating the line that ends in it [Lindsey et al. 2025; Maar et al. 2026].

So form choice is not a last-moment event.

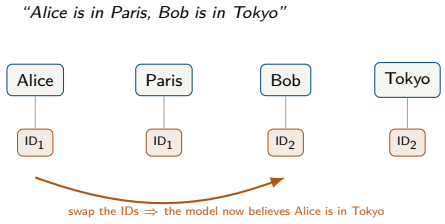
Claim B: sense is recoverable

- Sparse autoencoders decompose the stream into large inventories of interpretable features [Huben et al. 2024; Templeton et al. 2024].
- **Well-known illustration:** the "Golden Gate Bridge" feature. It fires on descriptions of the bridge in English, on the same content in other languages, *and on images of it*. Amplify it and the model begins to describe itself as the bridge.
- In vision–language models: visual tokens become readable in language space in middle-to-late layers [Venhoff et al. 2025]; recent method: LatentLens [Krojer et al. 2026]. CLIP representations decompose into text-labellable directions [Gandelsman et al. 2024]; language models can be used to *name* visual features by steering them [Ferrando et al. 2026].
  Already on visual input senses are activated before any text is produced

Claim C, evidence I: entities have addresses

Converging evidence

To keep track of who is where, models assign entities **abstract identifiers** i.e. vectors that encode *which attribute belongs to which entity*, and that are neither a word nor a position in the string.



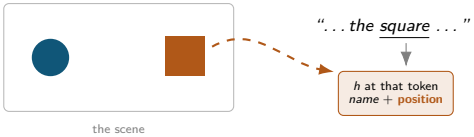
The same picture recurs as binding IDs [Feng & Steinhardt 2024], as a structured and manipulable binding subspace [Dai et al. 2024].

Discourse referents (Heim's file cards) in vector space.

Claim C, evidence II: the name carries the location

The finding

In a vision–language model, at the **text token that spells an object's name**, the residual stream carries **where that object is**: a property expressed by neither the question nor the answer.



Reported as content-independent position indices scaffolding binding [Assouel et al. 2025], as a linear mechanism binding location to object-name activations [Kang et al. 2026], as relations computed language-side from distributed visual signals [Cui et al. 2026], and in work tracing localisation and relational circuits [Schaumlöffel et al. 2026; Ma et al. 2026].

Both the *form* and the spatial position (property of the *referent*) are encoded in the same vector.

Part IV. A case study

Yingjin Song, Denis Paperno, and Albert Gatt.
Who Is Left of Whom? Tracing Spatial Evidence and Role Binding in Relative-Position Reasoning. *Actionable Interpretability* workshop at COLM 2026

a referent property and a description component, pulled apart

Task and design

Image:

=

Text Description:
*In a scene with three objects arranged horizontally from left to right, there are a **red circle**, a **blue square**, and a **green triangle**.*

Query: **Target** **Reference**
Is the **blue square** to the left or right of the **green triangle**?
Answer: Left

Task.: "Is the [target] left or right of the [reference]?"

Task and design

Task. "Is the [target] left or right of the [reference]?"

Three matched conditions

- vision–language model + image
- the same model + a text description of the same scene
- its language backbone + the same text

[LLaVA-1.6, InternVL3.5, Pixtral, 7–12B, with their backbones]

Two paired perturbations, orthogonal by construction

- **location swap**: the world changes, the question does not ⇒ the answer must flip.
- **role reversal**: the world is unchanged, the question swaps target and reference ⇒ the answer must flip.

One perturbation touches the referents; the other touches only the description of them.

[*target* = the object being localised (figure); *reference* = the anchor (ground), after Talmy.]

First: accuracy hides brittle relational computation

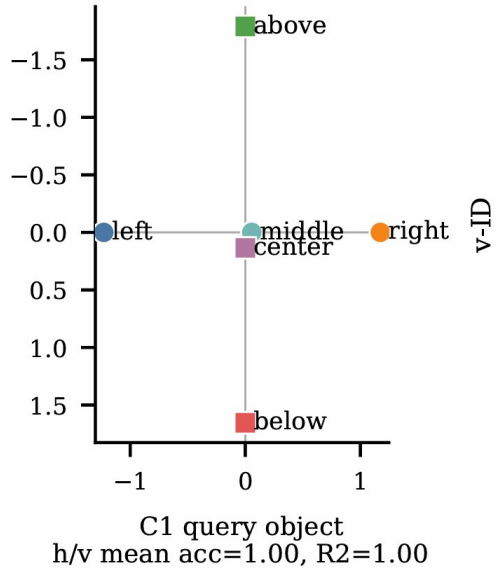
- Models answer many individual questions correctly while failing the *paired* checks. For example, getting "is A left of B" right and "is B left of A" wrong.
- Ordering: image $>$ text-to-the-VLM $>$ text-to-the-backbone, and the gap is far wider on consistency than on accuracy.
- Note the middle term: **the vision–language model beats its own language backbone on identical text-only input.** Return to this at the end.

A property of the referent: object-centred spatial IDs

Construction. For each object mention, subtract the mean state of objects with the *same attribute* ("red circle"), then average within position label:

$$\mathrm{ID}_\pi = \mathbb{E}\left[\bar{h}_{i,o} \mid \pi_{i,o} = \pi\right]$$

Object identity is removed by construction. What remains is position.



A component of the description: a figure/ground direction

Geometry. Take the *same* object in a reversal pair. Once used as target, once as reference:

$$d_{p,o} = h_{o,\text{target}} - h_{o,\text{reference}}$$

These contrasts point in a common direction across models and conditions.

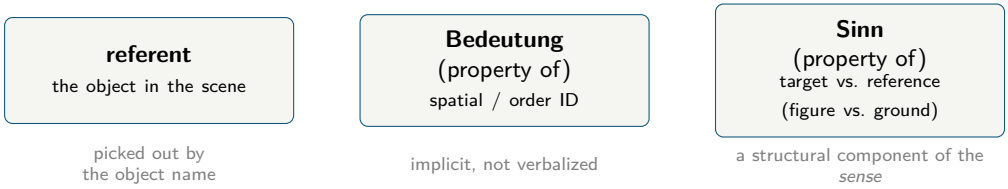
Causality. Push the target state away from d_{role} and the reference toward it:

- the correct-answer margin drops, more than for a norm-matched orthogonal control; [3, 5]
- the damage scales with intervention strength;

If steered positively, directions estimated on synthetic scenes transfer to real photographs, improving paired consistency with frozen weights:

Model	Condition	Accuracy	Target-Reference Reversal Consistency	Location-Swap Consistency
LLaVA	VLM+image	89.2 → 89.8 (+0.7)	81.5 → 81.8 (+0.3)	79.0 → 80.2 (+1.2)
LLaVA	VLM+text	66.1 → 67.0 (+0.9)	47.1 → 47.7 (+0.5)	36.2 → 37.6 (+1.4)
LLaVA	LLM+text	47.3 → 49.1 (+1.8)	19.8 → 23.2 (+3.4)	15.9 → 17.7 (+1.8)

Reading the case study back into Frege



- The position ID behaves like a property of the **Bedeutung**: it survives swapping the roles, and it changes when the world changes.
- The role direction behaves like a component of the **Sinn**: it survives changing the world, and it changes when the description changes.
- They are **separately decodable and separately steerable**. That is Frege's separability commitment, turned into an experiment with a result.

The speculative takehome

Observation

The vision–language model beats its own language backbone **on identical text-only input**. In accuracy, in consistency, in how localisable its spatial representations are, and in how far the answer rides on object representations rather than on the relation words *left* and *right*.

Interpretation, offered as such

Training on **referentially grounded** data appears to sharpen **conceptual-structural** representations (here: figure/ground role binding and object-centred position) and the sharpening persists when the grounding is taken away.

Grounding does not just add a modality; it enhances what the residual stream keeps track of.

[Consistent with Buzeta et al. 2026 (visual data correcting binding shortcuts), Nikankin et al. 2025 (modality-specific circuits), Tong et al. 2026.]

Part V. Sense and reference as an empirical programme

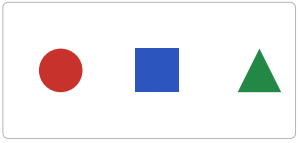
The question I actually want to ask

When a model produces a referring expression,
in what order are the referent and the sense fixed?

Three hypotheses, all consistent with the behaviour:

1. **Fregean pipeline.** Referent, then sense, then form (as in classic REG, Dale & Reiter).
2. **Form-first.** A description is produced by surface statistics and the referent is whatever satisfies it. No separable sense stage.
3. **Joint.** Referent and description settle together in one constraint-satisfaction step; the "stages" are our gloss on a smear.

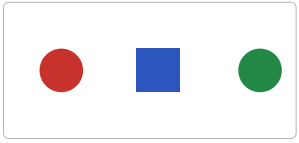
A concrete next step: REG with controlled distractors



shape suffices

"The blue square is to the right of _____"

"the circle"



colour needed

"The blue square is to the right of _____"

"the red circle"

The manipulation

The **target never changes**: same red circle, same position, same question. Only the *third* object changes, and with it **which properties suffice to pick the target out**.

Behavioural precondition: do the distractors actually shift the distribution over generated expressions? If not, there is nothing internal to look for.

How to test this

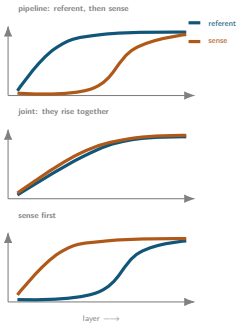
1. Referent: from which layer can we decode *which object's position* is about to be described?

probe, then patch across distractor conditions

2. Sense: does the distribution over *{"circle", "red"}* at the post-*"the"* position evolve differently across layers depending on the distractors?

logit lens; no fitted decoder needed

3. Sense *before* referent? [speculative]
LatentLens the target's *image patches*. Is the *circle* vs. *red circle* contrast already there: before any text token, and so before any referent has been chosen for description?



The referent is fixed by the question in every condition. We should manipulate only context: distractors/rest of the scene.

Takeaways

1. Frege's distinction is at the basis of referroing expression generation.
2. Classic (Dale & Reiter) REG is **sense selection**.
3. Inside language models, all vertices of the Frege triangle are recoverable.
4. The case study (Song et al.) illustrates this; referential grounding appears to strengthen conceptual structure even in the text-only stream.
5. Open questions: are sense and reference ordered and separable?

Thank you!

Follow-up: do givenness and deixis share a space?

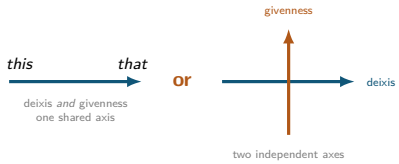
Gundel, Hedberg & Zacharski (1993)

"different determiners and pronominal forms conventionally signal different cognitive statuses"

in focus *(it)* > activated *(this, that)* > familiar *(that N)* > uniquely identifiable *(the N)* > referential *(indef. this N)* > type identifiable *(a N)*

Motivation. The same forms do double duty. *This* and *that* signal **cognitive status** in discourse and **proximal vs. distal** deictically, pointing at things in a scene.

Two functions, one form. Inside a model: **one direction, or two?**



Two things this could settle

- **Interference.** Part IV gives a validated, steerable spatial direction. Steer an object's spatial ID in a context with *no* spatial content: does the choice between *this* and *that* move? Then the converse.
- **Scalar or categorical?** One continuous axis with thresholds (Ariel), or discrete statuses (Gundel et al.)? The geometry can adjudicate what the behavioural data has not.

Backup: glossary

residual stream	the running vector at each token position that every layer reads from and adds to
activation	the value of that vector (or part of it) at a given layer and position
probe	a classifier fitted to recover some property from an activation
logit lens	decoding an intermediate activation directly into vocabulary
activation patching	transplanting an activation from one run into another and observing the change
steering	adding a chosen direction to an activation to change behaviour
sparse autoencoder	a method for decomposing activations into many sparse, inspectable features
superposition	more features than dimensions, stored non-orthogonally

Backup: the two perturbations

Object-location swap

Scene: `[red circle] [blue square]`
"Is the red circle left of the blue square?" → yes.

Swap positions: `[blue square] [red circle]`
Same question → the answer must flip.

Tests whether the position representation updates.

Target/reference reversal

Scene unchanged: `[red circle] [blue square]`
Q1: red circle relative to blue square → left.
Q2: blue square relative to red circle → right.

Tests whether the role binding is correct.

One perturbation changes the scene; the other changes only how it is described. That orthogonality is what makes the referent/sense separation testable rather than merely nameable.

Backup: staged pipeline from activation patching

- Under **location swap**: three stages: source-side evidence recovers the answer in early-to-middle layers, query-side target/reference tokens in middle layers, the final pre-answer position in late layers.
- Under **role reversal**: source-side patches do nothing. Indeed, the scene did not change. Recovery is carried by query-side role tokens and the last position.
- Largest effects come from patches covering the **object pair jointly**, not single mentions and not the relation words.
- Description-side recovery: ≈ 0.95 for the VLM on text vs ≈ 0.73 for its backbone; query-side ≈ 1.0 vs ≈ 0.87.

Backup: numbers

Source → query chain patching (corrupt-answer flip rate)

Setting	LLaVA	InternVL	Pixtral
VLM + image, all visual tokens	62.9%	69.1%	51.2%
VLM + image, object bbox + strip	34.0%	11.6%	5.1%
VLM + text, all object mentions	17.7%	24.3%	13.7%
LLM + text, all object mentions	20.0%	15.6%	10.3%

Role-direction steering transferred to real images (What'sUp)

Model	Condition	Accuracy	Reversal consistency
InternVL	VLM + image	93.4 → 94.1	89.8 → 91.1
InternVL	LLM + text	64.6 → 70.7	31.2 → 44.3
Pixtral	VLM + text	81.8 → 84.1	66.1 → 71.4

[Directions estimated on synthetic scenes; model weights frozen.]

Backup: models and data

VLM	Architecture	LLM backbone
LLaVA-v1.6-Mistral-7B	projector-concat	Mistral-7B-Instruct
InternVL3.5-8B-Instruct	ViT–MLP–LLM	Qwen3-8B
Pixtral-12B	projector-concat	Mistral-Nemo-Instruct

- Benchmarking additionally covered LLaVA-1.5 (7B/13B), LLaVA-1.6-Vicuna (7B/13B), Qwen3-VL (4B/8B), InternVL3.5-14B.
- Synthetic scenes: 2 and 3 objects on a 3×3 grid, 36 colour–shape types, balanced across axes and layouts; corrupt counterparts generated by swapping positions.
- Natural images: What'sUp [Kamath et al. 2023], with systematically swapped object positions.

Backup: limitations of the case study

- Four discrete relations (*left*, *right*, *above*, *below*). No depth, distance, containment, occlusion, or viewpoint relations.
- The text conditions *serialise* a layout into language. They test whether linguistically supplied spatial structure is used like visual spatial structure.
- Three open models at 7–12B; smaller or bigger models might have different properties.

References — philosophy and linguistics I

- Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. Routledge.
- Donnellan, K. (1966). Reference and definite descriptions. *Philosophical Review* 75(3).
- Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100.
- Grosz, B., Joshi, A. & Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2).
- Gundel, J., Hedberg, N. & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69(2).
- Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, UMass Amherst.
- Kamp, H. (1981). A theory of truth and semantic representation.
- Karttunen, L. (1976). Discourse referents. In J. McCawley (Ed.), *Syntax and Semantics 3: Notes from the Linguistic Underground*, pp. 363–385. Academic Press.
- Kripke, S. (1980). *Naming and Necessity*. Harvard UP.
- Langacker, R. (1987). *Foundations of Cognitive Grammar*, vol. 1.
- Putnam, H. (1975). The meaning of "meaning".
- Rosch, E. et al. (1976). Basic objects in natural categories. *Cognitive Psychology* 8(3).
- Russell, B. (1905). On denoting. *Mind* 14(56).
- Strawson, P.F. (1950). On referring. *Mind* 59(235).
- Talmy, L. (2000). *Toward a Cognitive Semantics*. MIT Press.

References — REG and NLG I

- Andreas, J. & Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. EMNLP.
- Dale, R. (1989). Cooking up referring expressions. ACL.
- Dale, R. & Haddock, N. (1991). Generating referring expressions involving relations. EACL.
- Dale, R. & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2), 233–263.
- Frank, M. & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science* 336.
- Gatt, A. & Belz, A. (2009). Introducing shared tasks to NLG: The TUNA shared task evaluation challenges.
- Gatt, A., van der Sluis, I. & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. ENLG, 49–56.
- Kazemzadeh, S. et al. (2014). ReferItGame: referring to objects in photographs of natural scenes. EMNLP.
- Krahmer, E., van Erk, S. & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics* 29(1).
- Krahmer, E. & van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Computational Linguistics* 38(1).
- Mao, J. et al. (2016). Generation and comprehension of unambiguous object descriptions. CVPR.
- Monroe, W. et al. (2017). Colors in context. *TACL* 5.
- Reiter, E. & Dale, R. (2000). *Building Natural Language Generation Systems*. CUP.
- van Deemter, K. (2002). Generating referring expressions: boolean extensions of the incremental algorithm. *Computational Linguistics* 28(1).
- Yu, L. et al. (2016). Modeling context in referring expressions. ECCV.

References — interpretability I

- Assouel, R., Campbell, D., Bengio, Y. & Webb, T. (2025). Visual symbolic mechanisms: emergent symbol processing in vision language models.
- Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Computational Linguistics*.
- Belrose, N. et al. (2023). Eliciting latent predictions from transformers with the tuned lens.
- Buzeta, N. et al. (2026). Seeing to generalize: how visual data corrects binding shortcuts.
- Cui, K., Prakash, N., Messica, S., Raina, A., Bau, D., Torralba, A. & Rott Shaham, T. (2026). The dual mechanisms of spatial variable binding in vision–language models. arXiv:2603.22278.
- Dai, Q., Heinzerling, B. & Inui, K. (2024). Representational analysis of binding in language models. EMNLP.
- Elhage, N. et al. (2022). Toy models of superposition.
- Feng, J. & Steinhardt, J. (2024). How do language models bind entities in context? ICLR.
- Ferrando, J. et al. (2026). Language models can explain visual features via steering.
- Gandelsman, Y., Efros, A. & Steinhardt, J. (2024). Interpreting CLIP's image representation via text-based decomposition. ICLR.
- Geva, M. et al. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. EMNLP.
- Gurnee, W. & Tegmark, M. (2024). Language models represent space and time. ICLR.
- Hewitt, J. & Liang, P. (2019). Designing and interpreting probes with control tasks. EMNLP.
- Hewitt, J. & Manning, C. (2019). A structural probe for finding syntax in word representations. NAACL.
- Huben, R. et al. (2024). Sparse autoencoders find highly interpretable features in language models. ICLR.
- Kamath, A. et al. (2023). What's "up" with vision-language models? EMNLP.
- Kang, R., Chen, H., Gkioxari, G. & Perona, P. (2026). Linear mechanisms for spatiotemporal reasoning in vision language models. ICLR.
- Kim, N. & Schuster, S. (2023). Entity tracking in language models. ACL.

References — interpretability II

- Krojer, B. et al. (2026). LatentLens: revealing highly interpretable visual tokens in LLMs.
- Li, B.Z., Nye, M. & Andreas, J. (2021). Implicit representations of meaning in neural language models. ACL.
- Li, K. et al. (2023). Emergent world representations: exploring a sequence model trained on a synthetic task. ICLR.
- Li, Q. et al. (2026). Causal tracing of object representations in large vision language models: mechanistic interpretability and hallucination mitigation. AAAI.
- Lindsey, J. et al. (2025). On the biology of a large language model. Anthropic.
- Ma, X. et al. (2026). Attention in space: functional roles of VLM heads for spatial reasoning.
- Maar, J. et al. (2026) What's the plan? Metrics for implicit planning in LLMs and their application. ICLR
- Meng, K. et al. (2022). Locating and editing factual associations in GPT. NeurIPS.
- Nanda, N., Lee, A. & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. BlackboxNLP.
- Nikankin, Y., Arad, D., Gandelsman, Y. & Belinkov, Y. (2025). Same task, different circuits: disentangling modality-specific mechanisms in VLMs. NeurIPS.
- nostalgebraist (2020). Interpreting GPT: the logit lens.
- Pal, K. et al. (2023). Future Lens: anticipating subsequent tokens from a single hidden state. CoNLL.
- Park, K., Choe, Y.J. & Veitch, V. (2024a). The linear representation hypothesis and the geometry of large language models. ICML.
- Park, K., Choe, Y.J., Jiang, Y. & Veitch, V. (2024b). The geometry of categorical and hierarchical concepts in large language models. arXiv:2406.01506; published ICLR 2025.
- Saphra, N. & Wiegreffe, S. (2024). Mechanistic? BlackboxNLP.
- Schaumlöffel, T., Vilas, M.G. & Roig, G. (2026). Mechanisms of object localization in vision–language models. CVPR, 31356–31365. arXiv:2605.19792.
- Subramani, N. et al. (2022). Extracting latent steering vectors from pretrained language models. Findings of ACL.

References — interpretability III

- Templeton, A. et al. (2024). Scaling monosemanticity. Transformer Circuits.
- Tenney, I., Das, D. & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. ACL.
- Tong, S. et al. (2026). Beyond language modeling: an exploration of multimodal pretraining.
- Turner, A. et al. (2023). Steering language models with activation engineering. arXiv:2308.10248.
- Venhoff, C. et al. (2025). How visual representations map to language feature space in multimodal LLMs.
- Venhoff, C., Khakzar, A., Joseph, S., Torr, P. & Nanda, N. (2025b). Too late to recall: the two-hop problem in multimodal knowledge retrieval. CVPR Workshop on Mechanistic Interpretability for Vision.
- Vig, J. et al. (2020). Investigating gender bias in language models using causal mediation analysis. NeurIPS.
- Voita, E. & Titov, I. (2020). Information-theoretic probing with minimum description length. EMNLP.