Good Views Hidden in Plain Sight

A Closer Look at the Next-Token Baselines in "Proving and Improving the Platonic Representation Hypothesis with JEPAs"

Thomas Bernwald

4 November 2026 · Interactive version: thomasaas.github.io/good-views.html

TL;DR. Balestriero and LeCun's *Proving and Improving the Platonic Representation Hypothesis with JEPAs* [1] trains text models on pairs of captions from the same image and finds that a JEPA aligns well with a vision encoder, whereas next-token prediction (NTP) given the same caption pairs barely improves over single-caption training. They conclude that the gain comes from the JEPA objective rather than the paired data. However, that baseline is evaluated by mean-pooling the caption's hidden states. Using the hidden state at `[SEP]` (**h([SEP])**, the boundary token between the two captions in training) instead of mean-pooling, the same checkpoint matches the good-view JEPA on Flickr R@1 through one rotation (24.5% vs 24.6%). Training it to predict only the other caption reaches 27.8%, while the JEPA still leads on mutual 10-NN. Frozen Qwen2.5-0.5B and 1.5B baselines also jump from 24% to 35–43% R@1 when read out after a short prompt. At the paper's smallest web-scale point (2.24M images), reading **h([SEP])** closes most of the WIT gap (.075 → .093 vs .107) but leaves a clear Flickr retrieval gap (25.3% vs 33.9%). Comparisons between NTP and JEPA representations are thus sensitive to how the representation is extracted.

In *Proving and Improving the Platonic Representation Hypothesis with JEPAs* [1], Randall Balestriero and Yann LeCun prove under idealised assumptions that JEPA encoders learn the same representation up to one rotation when trained on *good views* (distinct observations of the same scene, such as two captions of one image). *Bad views* corrupt a single observation, such as masking one caption twice; because those views share surface wording, the encoder retains wording alongside scene content. Empirically, a small BERT trained as a good-view JEPA on Flickr captions aligns with frozen DINOv2 much better than a bad-view JEPA or next-token prediction (NTP). The authors argue that NTP acts like a bad-view JEPA whose positives are a prefix and its continuation, and therefore aligns with vision mainly through scale.

I ran a few quick experiments with LLM coding agents to test that comparison. The paper's paired-caption NTP baseline appears to underestimate what NTP learns from good views because its mean-pooled readout dilutes the cross-caption signal.

The paired baseline

The paper's main NTP baseline trains on single captions, which matches the data given to the bad-view JEPA (both score ~.12–.13 against DINOv2 on Flickr) but not the good-view JEPA, which sees which captions share an image. Section 3.1 addresses this with a paired NTP control:

> "Third, the gain comes from the objective and not from the information that two captions describe the same image: next-token prediction trained on two captions of the same image, `[CLS] a [SEP] b [SEP]`, barely moves (mutual 10-NN .042 to .043, R@1 .021 to .023), i.e., the same positives help next-token prediction far less than they help as JEPA views, as per Proposition 4"

(Table 1 measures alignment against an image JEPA trained from scratch on Flickr. Against frozen DINOv2 in Table 2, R@1 moves from .118 to .137. All results below use frozen DINOv2.)

At evaluation time, the paper feeds a single caption `[CLS] a [SEP]` and mean-pools the final hidden states across all tokens. In earlier layers, tokens in *b* can attend back to all of *a* as well as `[SEP]`, so cross-caption gradients reach the whole prefix. At the final layer, however, each state feeds only the output head at its own position: tokens inside *a* only see a prefix of *a* and predict the next token of *a* (or `[SEP]`), whereas `[SEP]` has seen all of *a* and is the only position trained to predict into the second caption *b* (specifically its first token). That makes **h([SEP])** an interesting readout candidate.

Appendix B of the paper tests a last-token readout only on single-caption NTP, where it trails mean pooling (R@1 .061 vs .119). That is expected: there `[SEP]` is followed by ignored padding and receives no loss. In the paired model, `[SEP]` predicts the first token of *b*, yet wasn't evaluated separately.

Figure 1 · Flickr30k BERT-S (33M): Paper Readout (lighter) vs Alternative Readout (solid)

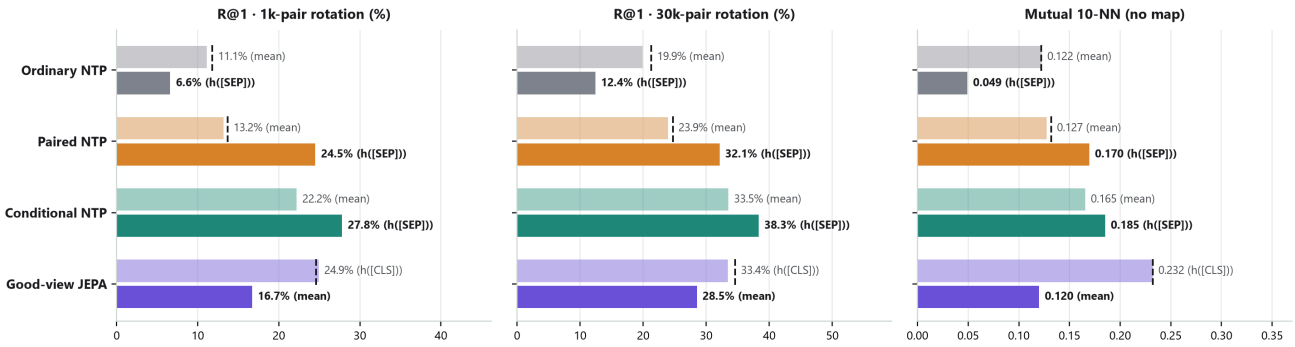


Figure 1. Flickr30k, BERT-S (33M) text encoders trained from scratch for 200 epochs (one seed each), 1,000 test images against frozen DINOv2 ViT-B/14. R@1 is image→caption retrieval through one orthogonal map; chance is 0.1%. The paired-NTP checkpoint was trained with the authors' code, unchanged.

Using the authors' training code unchanged (BERT-S, Flickr30k, 200 epochs, seed 0) and changing only the readout:

- **R@1, 1k-pair rotation: 13.2% → 24.5%** (good-view JEPA: 24.6%).
- **R@1, 30k-pair rotation:** 23.9% → 32.1% (good-view JEPA: 34.6%).
- **Mutual 10-NN:** .127 → .170 (good-view JEPA: .232).

On retrieval through an orthogonal map (the paper's "direct test of equality up to rotation"), paired NTP read at h([SEP]) matches the good-view JEPA on the 1k-pair fit, while the JEPA retains a clear lead on mutual 10-NN. An independent GPT-written run stopped at epoch 139 shows the same jump (14.0% → 22.2% on 1k, 26.1% → 31.5% on 30k).

Reading h([SEP]) hurts both single-caption NTP (11.1% → 6.6%) and the JEPA (11.1% vs 24.9% for h([CLS])). It only helps when `[SEP]` precedes a second caption of the *same* image during training: in a 25-epoch pilot where caption *a* was paired with a caption from a different image, h([SEP]) dropped to chance (Appendix A1).

Alignment at h([SEP]) also keeps rising while the validation loss more than doubles (Figure 2). The paper's validation metric (used for its early-stopped row, though main tables use final checkpoints) is next-token loss on *single* captions, which rises after ~20–30 epochs as the model memorises training phrasing. Because single captions have only ignored padding after `[SEP]`, that validation loss never evaluates `[SEP]` or cross-caption prediction; early-stopping on it would pick epoch 20 (17.5% R@1 instead of 24.5%).

Predicting only the other caption

Paired NTP still spends roughly half its loss predicting caption *a*, whereas the JEPA only penalises disagreement between views. To isolate cross-caption prediction, I trained a **conditional** NTP variant using

the same `[CLS] a [SEP] b [SEP]` input, initialization, and schedule, but masking the loss on *a* so only *b* is predicted (with gradients still flowing into *a* and both caption orders sampled).

Read at h([SEP]), conditional NTP reaches **27.8% / 38.3% R@1** (1k / 30k fit), ahead of the good-view JEPA's 24.6% / 34.6%, and even mean pooling reaches 22.2% / 33.5%. Its mutual 10-NN (.185) remains below the JEPA's (.232). Part of the gain over standard paired NTP may simply come from *b*'s tokens receiving roughly twice the weight in the loss.

Readout choice affects the JEPA as well: its native h([CLS]) readout (the projector input) beats mean pooling on Flickr R@1 (24.9% vs 16.7%), while the projector output itself scores 26.7% on the 1k fit (2.2 points above paired NTP's h([SEP])) and 32.4% on the 30k fit (Appendix A1). On Flickr, the paper thus compares the JEPA at its trained summary token against paired NTP mean-pooled over tokens trained for within-caption prediction. (On WIT-1024 and COCO-5k, the paper instead mean-pools all models across layers following the PRH protocol [2].)

Figure 2 · Alignment and Single-Caption Validation Loss Across Training Epochs

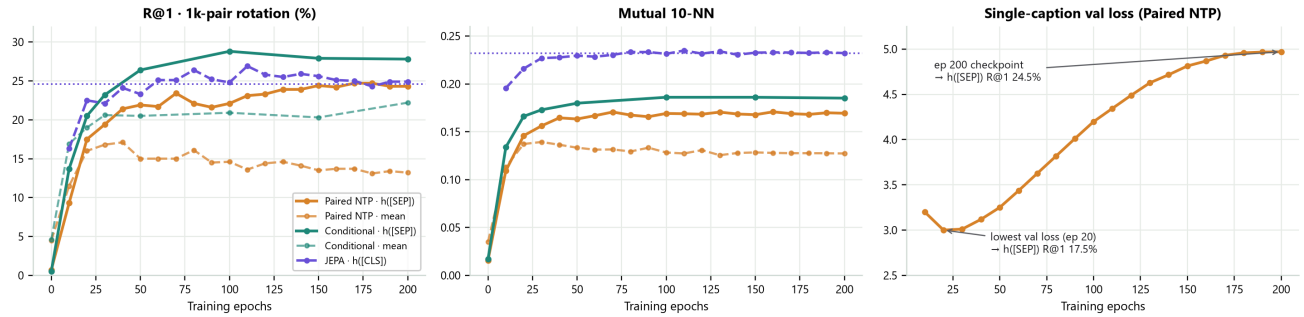


Figure 2. Alignment during training across epochs. Solid lines are the alternative readout, dashed lines the readout the paper uses. Paired NTP: every 10 epochs from the run's own validation dumps (bf16). Other runs: saved checkpoints re-embedded in fp32. Validation loss is single-caption next-token loss on val+test captions; because single captions have only ignored padding after `[SEP]`, it never scores predicting *b* from *a*.

All numbers (Flickr30k, BERT-S 33M vs frozen DINOv2 ViT-B/14)

Model · readout	R@1 (1k)	R@1 (30k)	Mutual 10-NN
Ordinary NTP · mean *(paper's readout)*	**11.1%**	**19.9%**	**.123**
Ordinary NTP · h([SEP])	6.6%	12.4%	.049
paper, Ordinary NTP · mean (3 seeds)	*11.8±0.2%*	*21.3±0.4%*	*.122±.001*
Paired NTP · mean *(paper's readout)*	13.2%	23.9%	.127
Paired NTP · h([SEP])	**24.5%**	**32.1%**	**.170**
paper, Paired NTP · mean (3 seeds)	*13.7±0.4%*	*24.7±1.6%*	*.132±.002*
Conditional NTP · mean *(paper's readout)*	22.2%	33.5%	.165
Conditional NTP · h([SEP])	**27.8%**	**38.3%**	**.185**
Good-view JEPA · mean	16.7%	28.5%	.120
Good-view JEPA · h([CLS]) *(paper's readout)*	**24.9%**	**33.4%**	**.232**
paper, Good-view JEPA · h([CLS]) (3 seeds)	*24.6±1.5%*	*34.6±0.7%*	*.232±.006*
paper, bad-view JEPA · h([CLS])	*13.3±0.7%*	*20.7±0.8%*	*.132±.001*
paper, ordinary NTP early-stopped	*13.4±0.8%*	–	*.124±.002*

Bold marks the better readout per model. Paper values are three-seed means ± half-range; ours are single seeds.

Without a fitted map

Because fitting an orthogonal map could flatter a weaker representation, I also checked three of the paper's map-free text evaluations: caption→caption retrieval, STS-B sentence similarity, and the Table 6 "scene share" estimator (variance explained by the shared image rather than caption wording). At h([SEP]), paired NTP closes most of the gap to the JEPA on all three (caption retrieval 36.7% → 53.8% vs 60.6%), and conditional NTP slightly exceeds the JEPA on each. These three metrics are text-only; on mutual 10-NN against DINOv2, the JEPA still leads.

Model · readout	Caption→caption R@1	STS-B	Scene share (k = 32)	Mutual 10-NN
Paired · mean	36.7%	.447	.633	.127
Paired · h([SEP])	53.8%	**.582**	.791	.170
Conditional · h([SEP])	**62.3%**	.572	**.816**	.185
JEPA · h([CLS])	60.6%	.561	.803	**.232**

Single seeds, 200 epochs. Paired NTP: Claude-run with the authors' code; conditional NTP and the JEPA: GPT-run. More tasks and readouts in Appendix A1.

At web scale: the 2M Recap point

To test beyond Flickr, I trained paired BERT-S NTP on the smallest web-scale point in the paper's scaling plot: 2.24M Recap-DataComp images (one alt-text and two LLaVA recaptions per image), following the exact schedule of the Table 7 good-view JEPA control (20 × 1M samples, learning rate 10^{-4}, gradient clip 1.0, one seed).

Mean pooling again shows little gain from pairing: .075 WIT mutual 10-NN (vs .070–.071 for single-caption NTP) and 23.2% Flickr R@1 (vs 22.3–24.6%). Reading h([SEP]) raises WIT mutual 10-NN to .093, close to the JEPA's range (.095–.115, mean .107), but only lifts Flickr R@1 to 25.3% against the JEPA's 33.9%, a gap substantially larger than the seed variation the paper reports.

Reading one token further by appending "the" (which starts nearly every Recap recaption) reaches .104 on WIT and 30.1% on Flickr, though this was the best of ten readouts scored on the same test sets, so treat it as a lead rather than a result (Appendix A2). Testing the analogous pre-declared heuristic on Flickr by appending "a" (the most common first word in Flickr) gave smaller, inconsistent changes: +2.1 points on 30k-pair retrieval and +0.016 on mutual 10-NN, but −0.2 on 1k-pair retrieval, with +0.9 on 30k-pair retrieval for the JEPA control as well (Appendix A1). One possible reason paired NTP lags further behind on Recap is that two of the three captions per image come from the same LLaVA captioner, so predicting b from a also rewards matching generator style. That is untested.

Model · readout	WIT 10-NN	Flickr R@1 (30k)	Flickr 10-NN
paper, ordinary NTP (3 seeds)	*.070–.071*	*22.3–24.6%*	–
Paired NTP · mean	.075	23.2%	.103
Paired NTP · h([SEP])	.093	25.3%	.140
Paired NTP · "the" *(exploratory)*	.104	30.1%	.182
paper, good-view JEPA (3 seeds)	*.107 (.095–.115)*	*33.9%*	–

One seed. WIT: the paper's PRH protocol (per-layer features, best layer pair; the JEPA's .107 is mean-pooled). Flickr: final layer, rotation fitted on 30k pairs, frozen DINOv2. Full readout search, including generated continuations, in Appendix A2.

Readouts for frozen language models

Pretrained LLM baselines in the paper are also mean-pooled over the raw caption. Appending a short prompt and reading the final token's hidden state (without generating tokens) substantially changes their scores. Prompted last-token embeddings are well known [3–6]; here they just show that the LLM baselines are also sensitive to the readout:

```
Caption: {A}
Alternative caption of the same scene:
```

Under the paper's Flickr protocol (final layer, 30k-pair rotation), Qwen2.5-0.5B R@1 jumps from **24.1% to 35.4%** (14.0% to 22.2% with a 1k-pair fit), passing the Flickr-trained good-view JEPA (34.6%). On Qwen2.5-1.5B, a continuation prompt reaches **42.9%** (raw caption 24.5%, alternative prompt 35.3%). On WIT-1024, mutual 10-NN rises from **.102 to .138** (0.5B), **.108 to .155** (1.5B), and **.153 to .170** (72B). The relative gain shrinks with scale, the continuation prompt hurts at 72B, and 72B remains below the paper's 109M web-trained JEPA (.184).

Prompt wording itself may matter less than providing a post-caption summary position: "Continuation of this caption:" matches "Alternative caption of the same scene:" on 0.5B and beats it on 1.5B Flickr, whereas bare markers like literal "[SEP]" or Qwen's end-of-text token do not help (Appendix A3).

Closest to this, Wang et al. raise frozen Qwen3's alignment with DINOv2 on WIT with sensory prompts ("imagine what it would be like to see…") [5]. They let the model generate a continuation and average its hidden states over the generated tokens and over all layers; in a single forward pass averaged over the input tokens, the prompt did not help. The prompted last-token state here needs no generation. Wang et al. also find that mean-pooling over generated tokens beats single-token readouts [7]. Here, greedily generating an alternative caption and pooling its states did not beat the single prompted last-token state (33.9% vs 35.4% Flickr R@1 on 0.5B; Appendix A3), but that comparison reads single layers only. Averaging over all layers, which Wang et al. find often beats any single layer [5], I haven't tested yet.

Figure 3 · Frozen Qwen2.5: Raw Caption Mean (Paper) vs Prompted Last-Token Readout

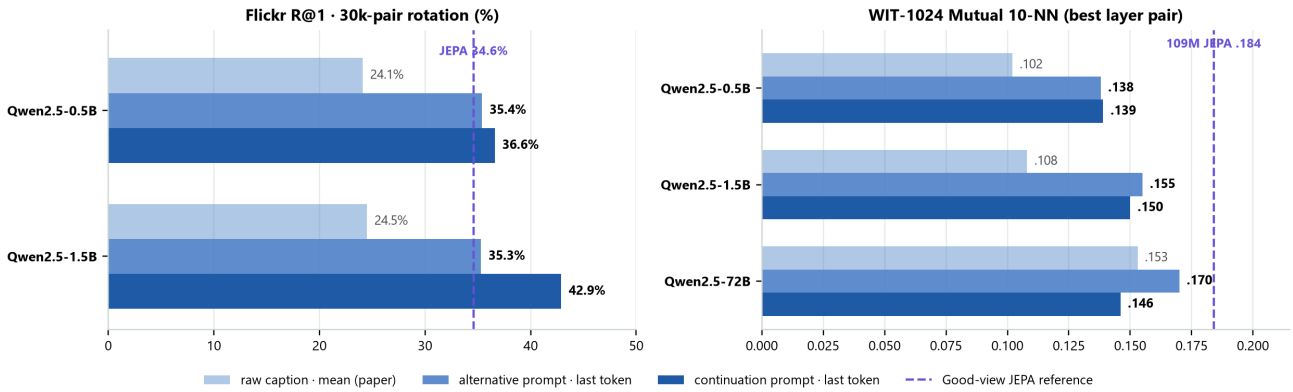


Figure 3. Frozen Qwen2.5 alignment across readout conditions on Flickr30k (final layer, orthogonal map fitted on 30k pairs) and WIT-1024 (mutual 10-NN, best layer pair).

Conclusion

On Flickr with 33M models, the paper's paired control does not show that alignment gains come from the JEPA objective rather than the pairing information. Evaluated at `[SEP]`, the paper's own paired NTP checkpoint gains 11 R@1 points and matches the good-view JEPA on retrieval through a rotation, while

conditional NTP surpasses it. Mean-pooling therefore understates what paired NTP learns, even though the JEPA still leads on mutual 10-NN (.232 vs .170–.185) and on Flickr retrieval at the 2.24M-image Recap point (33.9% vs 25.3%).

Several confounds also remain untested. In the paper, hyperparameters were tuned only for the good-view JEPA, while NTP "use[s] the same recipe and [was] not tuned"; each JEPA sample draws eight views from an image's five captions whereas paired NTP sees two; and the JEPA encoder is bidirectional while NTP is causal. All runs here use a single seed, prompted LLM readouts were chosen after inspecting Flickr results, conditional NTP was not run at web scale, and the paper's largest regime (109M models on 940M images) is not tested here. I also haven't gone through the theory (whose extension to NTP the authors note is heuristic).

To reproduce the main Flickr result with the authors' repository, train with `train.py --modality text --objective ntp --caption-views` and extract features with `evaluate.py embed <checkpoint> <out> --last-token` (where the last token of `[CLS] a [SEP]` is `[SEP]`).

Disclosure. The hypotheses, experimental design, and analysis are the author's; LLM coding agents from OpenAI, Anthropic, and Google were used to implement and run the evaluation pipelines and to assist with editing this post. Two independently written pipelines reproduce the paper's baselines (ordinary NTP 11.1% vs reported 11.8%, paired NTP 13.2% vs 13.7%, good-view JEPA 24.9% vs 24.6%, Qwen2.5-0.5B WIT .102 vs .102) and agree on the h([SEP]) effect.

Appendix

A1 · More Flickr results

Other tasks

The same frozen 200-epoch checkpoints and readouts as in Figure 1, on tasks without a fitted image map (except COCO). Ordinary and conditional NTP and the JEPA were scored by the GPT-run pipeline. Paired NTP was scored with a reimplementation of the same protocols, which reproduces the GPT-run numbers exactly on its 139-epoch checkpoint.

Model · readout	Caption→caption R@1	STS-B	Scene share	COCO R@1
Ordinary · mean	33.1%	.424	.611	3.3%
Ordinary · h([SEP])	12.6%	.428	.432	1.5%
Paired · mean	36.7%	.447	.633	3.4%
Paired · h([SEP])	53.8%	**.582**	.791	5.5%
Conditional · mean	50.6%	.550	.734	6.7%
Conditional · h([SEP])	**62.3%**	.572	**.816**	**7.9%**
JEPA · h([CLS])	60.6%	.561	.803	6.8%
JEPA · mean	51.3%	.541	.704	5.0%

JEPA readouts

JEPA readout	R@1 (1k)	R@1 (30k)	Mutual 10-NN
h([CLS]) *(paper)*	24.9%	**33.4%**	.232
h([CLS]) → projector	**26.7%**	32.4%	**.233**
mean	16.7%	28.5%	.120
h([SEP])	11.1%	21.7%	.071

Control: caption pairs from different images

To check whether h([SEP]) benefits simply from having any second caption after `[SEP]`, the 25-epoch GPT-run pilot included a shuffled control. It matches conditional NTP (loss only on b), except caption a comes from a different image in the batch. With unrelated context, the model ignores a: single-caption embeddings collapse to effective rank ~1, and alignment drops to chance under both readouts. These are 25-epoch pilot checkpoints from an earlier training loop, not the 200-epoch runs; in those, h([SEP]) only overtakes mean pooling after about 20–30 epochs (Figure 2).

Pilot arm (25 epochs) · readout	R@1 (1k)	R@1 (30k)	Mutual 10-NN	Eff. rank
Paired · mean	18.1%	26.5%	.145	40
Paired · h([SEP])	18.9%	27.1%	.160	18
Conditional · mean	20.2%	30.0%	.162	32
Conditional · h([SEP])	**22.9%**	**31.1%**	.160	18
Shuffled · mean	1.3%	2.3%	.027	1.1
Shuffled · h([SEP])	1.2%	3.4%	.020	1.5

Pre-declared check: appending “a”

To test whether reading one token into caption b also helps on Flickr, I appended “a” (the most common first word in Flickr training captions, chosen before scoring) to give `[CLS] a-caption [SEP] a`. Paired NTP is read

at the appended “a”, while the JEPA is still read at h([CLS]) as a control for adding the token (reading the JEPA at “a” performed much worse). Evaluated on the 200-epoch Figure 1 checkpoints; GPT-run (Codex), whose h([SEP]) and h([CLS]) numbers reproduce ours: paired NTP improves on 30k-pair retrieval (+2.1), mutual 10-NN, and caption retrieval, is flat on 1k-pair retrieval, and dips on STS-B; the JEPA control gains +0.9 on 30k-pair retrieval and a little on STS-B. One seed; not significant on its own.

Metric	NTP · h([SEP])	NTP · “a”	JEPA · h([CLS])	JEPA · h([CLS]) + “a”
Flickr R@1, 1k fit	24.5%	24.3%	24.9%	24.9%
Flickr R@1, 30k fit	32.2%	34.3%	33.4%	34.3%
Mutual 10-NN	.169	.185	.232	.232
Caption→caption R@1	53.8%	55.3%	60.6%	61.1%
STS-B Spearman	.582	.574	.561	.572
Scene share (k = 32)	.791	.795	.803	.804
COCO R@1 (Flickr means)	5.4%	5.7%	6.8%	6.9%

A2 · Web-scale readout search (exploratory)

All readouts of the Recap paired-NTP checkpoint (final, seed 0) that were tried, reported equally. Prefixes are appended after `[SEP]` as the beginning of a caption b; the state of the last prefix token is read. Caption a is unchanged and nothing is generated. These were scored on the same Flickr and WIT test sets, so the best entry is optimistically selected. GPT-run (Codex); its mean and h([SEP]) numbers match the Claude-run evaluation exactly.

One untested reading of why “the” beats h([SEP]) here: in Recap, caption b is either an alt-text (short, varied) or a LLaVA recaption (“The image shows…”). So part of what h([SEP]) is trained to predict is which kind of caption comes next, which is about format rather than the scene. Appending “the” settles that, and the state can encode content instead. On Flickr all captions are of one kind, which would fit appending “a” helping only a little there (Appendix A1). The final-layer state at “the” itself mostly predicts “image”, but like h([SEP]) it gets its scene content from the lower layers, where all of b’s later tokens attend to it.

Readout	Flickr R@1 (1k)	Flickr R@1 (30k)	Flickr 10-NN	WIT 10-NN
full mean (paper)	16.4%	23.2%	.103	.075
h([SEP])	16.5%	25.3%	.140	.093
last caption token	10.9%	19.7%	.056	.049
caption-only mean	15.4%	22.4%	.096	.072
mean without [SEP]	15.5%	22.5%	.096	.072
“this image shows”	14.8%	23.0%	.091	.102
“the image shows”	15.7%	22.5%	.087	.102
“a photo of”	15.6%	23.9%	.111	.093
“this text says”	15.7%	24.1%	.153	.062
“the”	**21.2%**	**30.1%**	**.182**	**.104**

Generating a continuation and pooling its hidden states (8 or 32 greedy tokens, with or without a prefix) reaches at most 31.4% Flickr R@1 (30k) and .094 on WIT: no clear gain over appending “the”, and worse on WIT. Wang et al. report that pooling generated tokens beats single tokens for pretrained LLMs [7]; this small trained model doesn’t show that.

A3 · Frozen Qwen2.5: all 15 readout conditions

WIT columns use the PRH protocol (best layer pair). The Flickr columns use the paper's Flickr protocol: final layer, rotation fitted on 30k pairs, R@1 among 1,000 test captions.

Condition	Description	0.5B · WIT	1.5B · WIT	72B · WIT	0.5B · Flickr R@1	1.5B · Flickr R@1
`raw_mean`	Caption alone; mean of all token states *(paper's readout)*	.102	.108	.153	24.1%	24.5%
`raw_last`	Caption alone; state of the last caption token	.073	.077	.091	20.2%	–
`alternative`	`Caption: {A}\nAlternative caption of the same scene:`, final `:` state	.138	**.155**	**.170**	35.4%	35.3%
`continuation`	`Caption: {A}\nContinuation of this caption:`, final `:` state	**.139**	.150	.146	36.6%	**42.9%**
`echo`	`Caption: {A}\nRepeated caption: {A}`, mean over second copy [4]	.113	.128	.155	**36.7%**	–
`cls_sep_mean`	Literal `[CLS] {A} [SEP]`, mean of all tokens	.098	.107	.108	24.3%	–
`cls_sep_last`	Literal `[CLS] {A} [SEP]`, state of final token	.108	.118	.121	23.1%	–
`cls_sep_echo`	Literal `[CLS] {A} [SEP] {A} [SEP]`, mean over second `{A}`	.118	.130	.126	32.3%	–
`native_eos_mean`	Caption + native EOS token, mean of all tokens	.098	.105	.149	23.7%	–
`native_eos_last`	Caption + native EOS token, state of EOS token	.051	.072	.067	1.6%	–
`generated_alternative`	Greedily generate up to 32 tokens *B* from alternative prompt; mean over *B* with prompt in context	.103	.126	.133	33.9%	–
`isolated_alternative`	Same generated *B*, encoded alone without prompt or *A*	.104	.119	.136	27.8%	–
`generated_continuation`	Generated continuation *B*, mean over *B* with prompt in context	.113	.119	.138	27.5%	–
`isolated_continuation`	Same generated continuation, encoded alone	.107	.110	.131	22.2%	–
`cls_sep_generated_mean`	Literal `[CLS] {A} [SEP] {B_alt} [SEP]`, mean of all tokens	.114	.130	.131	33.0%	–

A4 · Methods, reproduction and caveats

- **Metrics.** Mutual 10-NN and R@1 follow the authors' `prh.metrics.alignment`. The 1,014 validation pairs (or 30,014 train+validation pairs) fit the means and one orthogonal map. Scores are on the 1,000 Flickr test pairs, using one caption per image.
- **Readouts.** All are from the final layer. "Mean" averages all non-padding positions, including `[CLS]` and `[SEP]`. h([SEP]) is the state at the final `[SEP]` of `[CLS] a [SEP]`. Evaluation inputs never contain a second caption.
- **Uncertainty.** Every training arm has one seed. The paper's seed spread for paired NTP is ±0.4 (1k) and ±1.6 (30k) points. With 1,000 test queries, the standard error of R@1 near 25% is about 1.4 points. The h([SEP]) gain for paired NTP (+11 points) is far outside both, but the smaller differences between conditional NTP and JEPA are not.
- **Selection.** The Flickr test set was looked at repeatedly while exploring readouts. h([SEP]) was proposed beforehand, for the mechanistic reason above, but the prompt variants for frozen LLMs were not.
- The JEPA was trained with the authors' code and default recipe. The seed-0 run collapsed around epoch 98, when its h([CLS]) became nearly constant and R@1 fell to 0.2%. The seed-1 run shown here stayed healthy and matches the paper (24.9% vs 24.6%, mutual 10-NN .232 vs .232).

- The paired NTP run was also trained with the authors' unchanged code (`prh-proof` `a2b9949`, `stable-pretraining` `ab836bf`, PyTorch 2.11) on one H100. The conditional run used the same pinned code, with only the loss mask changed.
- Flickr-trained models of every kind barely transfer to WIT-1024 (mutual 10-NN ≈ .03), as the paper also notes, so I don't use WIT for them.

References

1. R. Balestriero, Y. LeCun. Proving and Improving the Platonic Representation Hypothesis with JEPAs. arXiv:2610.39302, 2026. Code: galilai-group/prh-proof.
2. M. Huh, B. Cheung, T. Wang, P. Isola. The Platonic Representation Hypothesis. ICML 2024.
3. T. Jiang et al. Scaling Sentence Embeddings with Large Language Models (PromptEOL). arXiv:2307.16645, 2023.
4. J. Springer et al. Repetition Improves Language Model Embeddings (echo embeddings). arXiv:2402.15449, 2024.
5. S. L. Wang, P. Isola, B. Cheung. Words That Make Language Models Perceive. arXiv:2510.02425, 2025.
6. S. Zhuang, X. Ma, B. Koopman, J. Lin. PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval. arXiv:2404.18424, 2024.
7. S. L. Wang, P. Isola, B. Cheung. The Truth Lies Somewhere in the Middle (of the Generated Tokens). arXiv:2605.09969, 2026.

Citation

```
@article{bernwald2026goodviews,
  title  = {Good Views Hidden in Plain Sight},
  author = {Bernwald, Thomas},
  year   = {2026},
  month  = {November},
  url    = {https://thomasaas.github.io/good-views.html}
}
```