MODELS ARE MORTAL, MEASUREMENTS ARE NOT: A PORTABLE, MODEL-AGNOSTIC REPRESENTATION FOR VIDEO MISINFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A representation belongs to the model that produced it: replace the model and everything fitted on it must be rebuilt, and a closed model exposes no representation at all. We ask whether the coordinates of a representation can instead be defined in language, as a fixed bank of yes/no questions that any model can answer, and test this on short-video misinformation across eleven frozen vision-language models, two prompting languages, a closed API and human annotators. A linear readout fitted once on any one open model deploys unchanged on the other ten, losing only .040 AUC on average, and, with same-event corroboration, comes within a point of a fine-tuned detector's accuracy. A matched control separates two properties, compatibility and validity. A bank of questions unrelated to the task also transfers across models, yet loses much of its discrimination across datasets, so fixing the coordinates buys compatibility while their content buys validity. A measurement view splits each reading into what the question measures and how reliably a model reads it; it accounts for why aggregates transfer when single questions disagree, and locates where reading thins, above all in the smallest models and in collapsed answer channels. In simpler terms, the questions hold the representation, and a model is only one way of reading it.

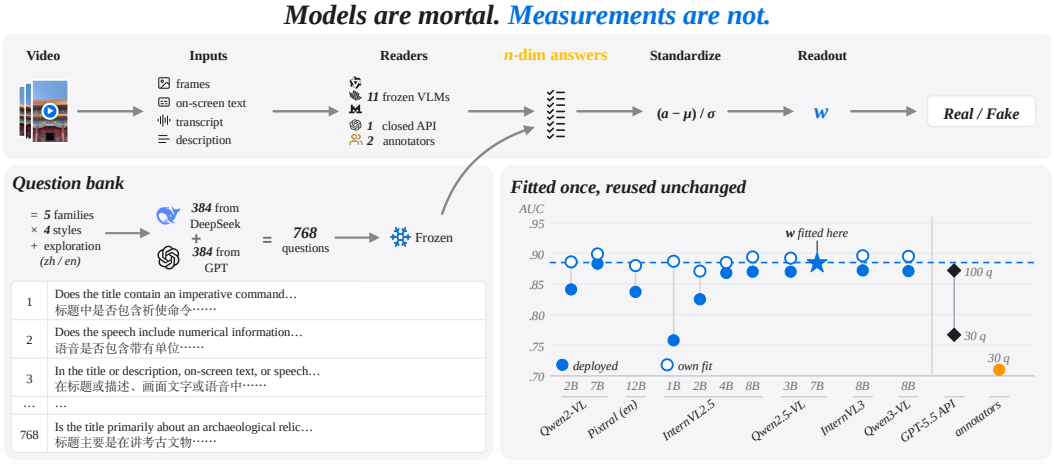


Figure 1: A representation defined by questions. Top: every reader answers the same frozen questions about a video's declared inputs; each answer sheet is standardized per question without labels, and one linear readout w, fitted once, decides real or fake. Bottom left: how the bank is built, with example questions 1, 2, 3 and 768. Bottom right: FakeSV AUC of w fitted on Qwen2.5-VL-7B (star; dashed: its own fit) and deployed unchanged on the other VLMs (filled; hollow: own fit; 1,200 videos). For GPT-5.5 (200 videos) and the annotators (100), w is fitted on Qwen's binarized answers to their 100 or 30 questions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Models keep being replaced, and downstream systems struggle to retain the predictors already built on them. When a predictor reads a model's internal representation, replacing the model can change the dimensions, the correspondence and the numerical distribution of its features, and the classifiers, thresholds and calibrations fitted on the old representation stop working; Qwen alone released three vision-language generations in two years (Wang et al., 2024; Bai et al., 2025; Qwen Team, 2025). Closed models restrict access to the representation further, and even where a representation is available its dimensions carry no stated meaning, so the basis of a decision cannot be audited item by item. These difficulties point to one cause: the coordinates of the representation are set by the model. Existing methods align the representations of different models with anchors, adapters or translators (Sec. 2), but each needs access to the models' internal states, most need another fit, and the aligned coordinates still carry no meaning. We therefore ask: can a coordinate system with stated meaning be defined first, and different models then asked to supply readings in it?

A fixed set of natural-language questions defines a feature space that belongs to no particular model. For each input, every model answers the same questions in the same order, and the answers are encoded as a numerical vector: dimension k always answers question k, and its definition is the same across samples and across models. A fixed bank thus gives answers the role that hidden states play as downstream features, while giving every dimension a stated, readable meaning. We find, further, that the correspondence is not only semantic: a classifier fitted on one model's answer sheets keeps its predictive power on another model's sheets. Replacing the model then replaces only the reader that supplies the features, while the definition of the features and the decision rule already learned are kept (Fig. 1).

We test this on short-video misinformation detection. On the official FakeSV test set, asking Qwen2.5-VL-7B directly whether a video is fake yields an accuracy of 63.5, while a linear probe fitted on the same model's hidden states (Alain & Bengio, 2017; Burns et al., 2023; Li et al., 2023) reaches 86.4. The gap shows that direct answering leaves signal in the internal representation unused, and that a linear probe reads it out but remains tied to that model's hidden states. We therefore build the readout on the coordinates defined by a fixed bank: the bank and the readers stay frozen, a linear readout is fitted once on the source model's sheets, and when the model is replaced only per-question statistics are estimated on the target side from unlabelled material to standardize its readings; the coefficients are reused verbatim.

This paper makes four contributions. First, we define a representation by questions and give a measurement view that splits each reading into what the question measures and how reliably a model reads it (Sec. 3). Second, models are interchangeable readers: coefficients fitted once on one open model read the other ten, and the same model prompted in another language, unchanged, with an average loss of .040 AUC among the open models; fitted on binarized answers, they also read a closed interface and human annotators, within the limits of their answer channels. Single questions are read differently by each model while aggregates measure the same thing (Sec. 4). Third, questions carry the meaning: between two models that were never aligned, the correspondence comes only from the question text, and a bank of questions unrelated to veracity also transfers across models, yet loses much of its discrimination across datasets, so compatibility and validity are distinct properties (Sec. 5). Fourth, task utility: with the same-event corroboration channel the readout comes within a point of a fine-tuned detector's accuracy and holds up better across events, and its within-dataset properties replicate on a second construct (Sec. 5).

2 RELATED WORK

Convergence and transfer. The Platonic Representation Hypothesis proposes that representations converge as models scale (Huh et al., 2024), and later work examines shared representational geometry across models, modalities and brains (Zhu et al., 2026; Marcos-Manchón et al., 2026; Marcos-Manchón & Fuentemilla, 2026). Kernel alignment, however, is correlational and yields no executable shared coordinates; once scale confounds are calibrated away, most of the global convergence disappears and only local-neighbourhood agreement remains (Gröger et al., 2026); mutual-nearest-neighbour evidence degrades as the evaluation set grows (Koepke et al., 2026); and linear translators between embedding models fail on incompatible pairs (Hamideche et al., 2026). En-

gineering routes move what was learned through the models' internal states or weights: anchors (Moschella et al., 2023; Svendsen et al., 2026), adapters (Lenc & Vedaldi, 2015; Bansal et al., 2021), translators (Jha et al., 2025), anchor activations (Kim & Han, 2026), CLIP pivots (Moayeri et al., 2023), ported steering directions (Oozeer et al., 2025; Poppi et al., 2026; Agarwal, 2026) and task vectors (Ilharco et al., 2023; Rinaldi et al., 2026); others constrain how the new model is trained (Shen et al., 2020; Ramanujan et al., 2022). We measure a different object: how much of an already-trained readout survives on a target model that was never jointly aligned.

Questions as representations. Answers to a fixed set of questions have served as a representation before: QA-Emb (Benara et al., 2024), TELLER (Liu et al., 2024), Pastel (Leite et al., 2025), CheckEval and BINEVAL (Lee et al., 2025; Cho et al., 2026), MemeScouts (Bueno et al., 2026), and the concept-bottleneck paradigm (Koh et al., 2020; Yuksekgonul et al., 2023; Oikarinen et al., 2023; Yang et al., 2023; Feng et al., 2025). Question-answer embeddings are therefore not new, and TELLER and CheckEval already pose the same questions to several models; what has not been measured is whether coefficients fitted on one model's answers transfer verbatim to another's, the role of label-free standardization, and the relation between compatibility and task validity.

Models as instruments. Computational social science is building validity norms for LLM annotators (Ziems et al., 2024; Lin & Zhang, 2026; Wang et al., 2026), including regression-based backups for when an annotating model is retired (Cheng et al., 2024), and work on prompt sensitivity and judge-specific biases documents how fragile a single judgment is (Sclar et al., 2024; Zheng et al., 2023; Liu, 2026). When a single reading is treated as a measurement with instrument error, a fixed bank, logit readout, per-question standardization and linear aggregation make frozen models interchangeable readers; the psychometric tools are classical (Spearman, 1904; Brown, 1910; Wang et al., 1991), with recent refinements (Ellis & Sijtsma, 2024; Huang, 2026). Task-side work is reviewed in App. A; the closest neighbour, SAFE, generates free-form question-answer pairs per video and fuses them with video features in a trained network (Liu et al., 2026), with no fixed coordinate system.

3 A Representation Defined by Questions

An instrument consists of three objects: a question bank, a table of per-question statistics and a vector of readout coefficients w, all plain text and numbers, none of them model weights (Fig. 1). The bank holds K fixed natural-language yes/no questions, each asked about declared material, which can be frames, the title, on-screen text, the speech transcript or text excerpts of same-event videos published earlier. A reader is a frozen model paired with a prompting language. For each question it returns the logit difference between *yes* and *no* at the first answer token, following the binary-VQA formulation of Lin et al. (2024), with no generation and no chain of thought. The K differences form the sheet $a_m(x)$, and coordinate k is question k for every reader. Per-question means μ_{mk} and standard deviations σ_{mk} of reader m are estimated from its own unlabelled sheets, and the readout is one L2-regularized logistic regression on the standardized sheet,

$$z_{mk}(x) = \frac{a_{mk}(x) - \mu_{mk}}{\sigma_{mk}}, \qquad u_m(x) = w^\top z_m(x) = \sum_{k=1}^{K} w_k \, z_{mk}(x), \tag{1}$$

with w and an intercept w_0 fitted once on the labelled sheets of a source reader. Reuse takes two forms. When the reader changes, w is kept and only the new reader's μ and σ are estimated from its unlabelled sheets. When the dataset changes and the reader does not, w and the source statistics are both kept and only the intercept is reset. Every score is a sum of named terms $w_k z_{mk}(x)$, so a decision decomposes exactly into per-question contributions.

A model of how a reading arises shows why subtracting a mean and dividing by a standard deviation should suffice. Let $s_k(x)$ be the answer tendency of content x on question k, fixed by the bank and the content and never observed directly. We model a reader's reading as

$$a_{mk}(x) = g_{mk} \, s_k(x) + b_{mk} + \varepsilon_{mk}(x), \qquad g_{mk} > 0, \tag{2}$$

where $\varepsilon_{mk}(x)$ is the part of the reading that the tendency does not explain. A frozen reader is deterministic, so this deviation is a fixed function of the content, specific to the reader; over videos we take it to have zero mean and to be uncorrelated with the tendency and across readers. This is the congeneric model of psychometrics (Jöreskog, 1971; Lord & Novick, 1968), with videos rather

than repeated trials as the source of variation. After per-question standardization,

$$z_{mk}(x) = \lambda_{mk}\, \tilde{s}_k(x) + \tilde{\varepsilon}_{mk}(x), \qquad \lambda_{mk} \in (0, 1], \tag{3}$$

where $\tilde{s}_k$ is the standardized tendency and λ_{mk} the reliability index of reader m on question k, whose square is the share of the reading's variance that the tendency explains (derivation in App. B). The tendency $\tilde{s}_k$ is fixed by the question and the content and shared by all readers; it is what is measured. The reliability index λ_{mk} belongs to the reader; it is how well the question is read. Standardization removes each model's own gain and offset without labels, so readers differ only in reliability and in their own deviations, the subject of Sec. 4; when the dataset changes, what may change is the relation between $\tilde{s}$ and the label, the subject of Sec. 5.

The design of the bank serves one requirement, that the coordinates are fixed before any reader answers them. Two things are necessary. The wording and order of each question are fixed and registered before any reading of it exists, so that coordinate k names the same question for every reader and a budget is simply the first k questions. Each question declares the material it reads and the reader sees only that material, so that what a coordinate measures is fixed by the question and not by whichever modality a model happens to lean on, as VQA models answer from language priors without looking at the image (Goyal et al., 2017); this also gives every reader, annotators and closed interfaces included, the same evidence. The remaining choices serve coverage and the absence of bias. Coverage comes from the construct's literature rather than from any dataset, so that content validity does not depend on the dataset. Each design family is drafted in several elicitation styles, to reduce each coordinate's dependence on one wording (Sclar et al., 2024; Mizrahi et al., 2024). The generators that draft the questions belong to families different from every logit reader, to reduce any bias toward a reader (Panickssery et al., 2024). A free exploration slot scans for cues the literature's priors may miss. Every drafted question enters the bank, and the readout is linear and uses every question, because selecting on finite-sample scores would freeze one draw of sampling noise into the instrument, whereas regularization shrinks weak questions continuously. The closed-API and human readers answer a subset of 30 or 100 questions, selected with labels on development videos (App. E).

Our instantiation has five design families (forgery and production traces; tone and affect; checkability; cross-modal consistency; same-event corroboration) (Bu et al., 2024; Hassan et al., 2017; Luo et al., 2021; Aneja et al., 2023; Abdelnabi et al., 2022; Qi et al., 2023b). They are crossed with four elicitation styles (direct, adversarial, contrastive, authenticating) into 20 cells of 32 questions, and an exploration slot of 128 completes 768 bilingual questions, half from each of two generators of different families (App. A). The slot began with 40 questions and was extended by 88 after the base bank had been scored, each question frozen before it was scored. Corroboration questions read the title and on-screen text of up to six same-event videos published earlier, with no labels and no later debunking; same-event membership comes from the benchmarks' event annotation, so a deployed system would need its own event grouping. Their readings do not flip under negation, so we read them as the availability of corroborating material, and always report the single-video result without this channel alongside.

The readers are eleven frozen vision-language models (VLMs), Qwen2-VL 2B/7B, Qwen2.5-VL 3B/7B, Qwen3-VL 8B (Wang et al., 2024; Bai et al., 2025; Qwen Team, 2025), InternVL2.5 1B/2B/4B/8B, InternVL3 8B (Chen et al., 2024; Zhu et al., 2025) (the 1B, 4B and InternVL3 models build on Qwen2.5 language models), and Pixtral 12B from another laboratory answering English questions (Agrawal et al., 2024); GPT-5.5, which answers on a five-level scale through its API and receives the same declared inputs, including eight keyframes (OpenAI, 2026); and two human annotators. All readers receive the same underlying material, each through its own ingestion format. Measurements use the FakeSV development set (Qi et al., 2023a): the eleven-reader matrix on a common panel of 1,200 development videos scored by every reader, and the six fully covered readers on all 3,082, with event-grouped five-fold cross-validation (CV); English FakeTT (Bu et al., 2024) is the cross-dataset target. In cross-validation a target reader's μ and σ come from the training side of each fold and from its own readings. Across datasets the intercept is set at the score rank that gives the unlabelled target pool the source's label prior as its predicted positive rate, the same access to unlabelled target data that RADAR (Lang et al., 2026) has. Regularization strengths, prompt templates, ingestion formats, splits and protocol details are in App. C.

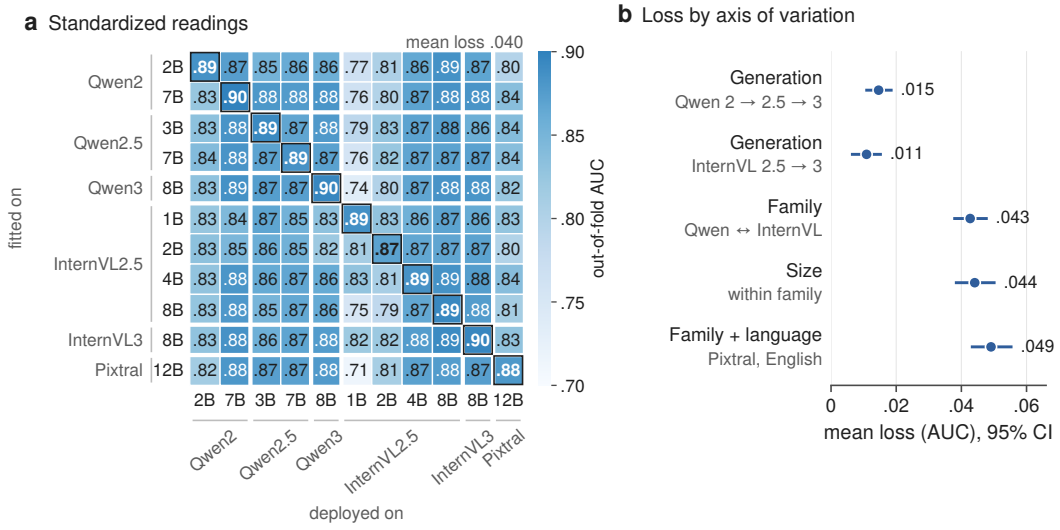


Figure 2: The eleven-reader transfer matrix. (a) Each cell is the out-of-fold AUC of the row reader's coefficients on the column reader's standardized sheets (1,200-video panel, 768 questions, event-grouped five-fold CV); the diagonal is the self-fit, and the annotation gives the mean off-diagonal loss. (b) Mean loss per axis of variation with 95% CIs (videos resampled, out-of-fold predictions fixed). After standardization only six of the 110 off-diagonal cells lose more than .10, all with InternVL2.5-1B as target; on all 3,082 videos the six fully covered readers lose .022 on average. The raw-reading matrix is in App. D.

4 Models Are Interchangeable Readers

Transfer. Among the eight readers with at least 3B parameters, coefficients fitted once on one model and applied unchanged to the others lose only .022 AUC against each model's own fit. Figure 2 shows the whole picture: each cell is a readout fitted on the row reader and applied unchanged to the column reader's sheets. Standardization is the decisive step. On raw readings the mean loss is .087; standardized with each reader's own label-free statistics it falls to .040. By axis of variation, generation is flattest, and family and size cost about the same, most of it through the 1B and 2B models: among readers of at least 3B parameters, family, size and generation lose .021, .017 and .014. Pixtral, from another laboratory, with its own language model, English questions and a mosaic input, loses .056 on average as a target, as much as the 2B models: it still sits in the same coordinate system as the others, although it shares neither their language model nor their training recipe.

Recognition. Without labels or alignment, the standardized sheet of one model picks out the same video among the 1,200 sheets of another: top-1 recognition is the share of videos whose sheet correlates most, across questions, with the same video's sheet from the other model. Among the eight readers of at least 3B parameters it averages .974, against chance below .001. Gröger et al. (2026) find that once scale is calibrated away, activation spaces agree only in their local neighbourhoods; in question coordinates that agreement appears directly, with no alignment step. Prompting language is a variation of the same kind. Switching 38 questions from Chinese to English on the same model costs a Chinese-fitted readout .022 AUC on English sheets, the scale of a change of model, and the Chinese and English sheets of the same video recognize each other in 599 of 600 cases.

Questions and aggregates. Question by question, two models' readings correlate only moderately, .65 in the median, whereas one model's readings before and after paraphrase correlate at .97. Under Eq. (3) the first estimates $\lambda_{Ak}\lambda_{Bk}$, so about two thirds of a single reading's variance is shared across models, while the paraphrase correlation also counts each model's own stable deviation. The same equation shows why this does not stop the aggregate from transferring. Deploying w fitted on reader A to the sheets of reader B gives

$$u_B(x) = w_A^\top z_B(x) = \sum_k w_{Ak}\,\lambda_{Bk}\,\tilde{s}_k(x) + \sum_k w_{Ak}\,\tilde{\varepsilon}_{Bk}(x), \tag{4}$$

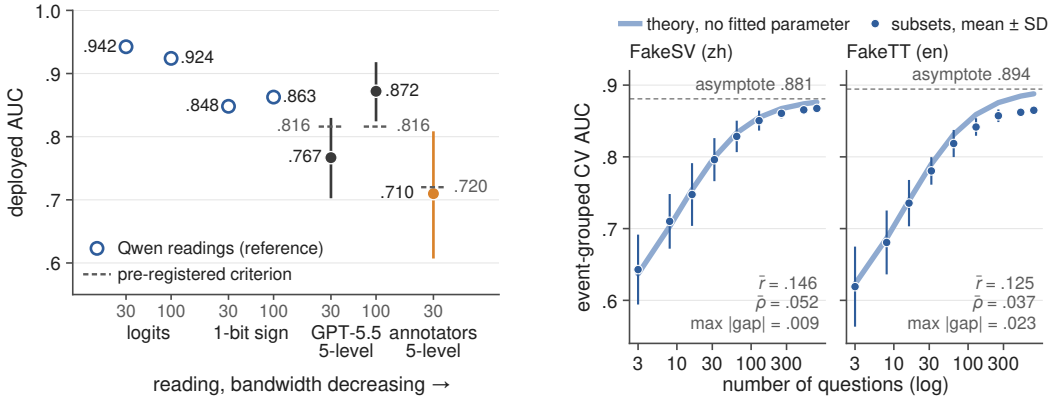


Figure 3: Left: coefficients fitted on Qwen2.5-VL-7B's one-bit signs for 30 or 100 questions (2,882 development videos), deployed on readings of decreasing bandwidth from 200 other videos (100 for the annotators): Qwen's own continuous logits and signs as references, GPT-5.5's five-level answers and two annotators' five-level answers, each standardized on its own pool, with 95% CIs over videos and the pre-registered criteria (.816 for GPT-5.5, .720 for the annotators). Right: the aggregation law, a curve with no parameter beyond $\bar{r}$ and $\bar{\rho}$, against random question subsets (mean $\pm$ SD) on FakeSV and FakeTT; max $|\text{gap}|$ is the largest deviation of a subset mean from the curve.

a signal term and a weighted sum of B's own deviations. If B's reliabilities are proportional to A's, $\lambda_{Bk} = c\,\lambda_{Ak}$ for every question, the signal parts of the two scores are proportional and rank content identically. Reliabilities need not match question by question; they only need to keep their relative size on the questions that carry weight. The deviation term is uncorrelated with the signal and stays small beside it when a model's deviations are weakly correlated across questions, so that they average out over hundreds of weighted questions; the transfer results are consistent with this assumption. This is how models that disagree on single questions can produce nearly the same ranking.

Saturation. The readout's power comes from aggregation. Let $\bar{r}$ be the mean correlation of standardized readings with the label and $\bar{\rho}$ the mean inter-question correlation, both estimated on the training side. If subset statistics are approximated by bank means, an equal-weight composite of k random questions, each signed toward the label, satisfies (Brown, 1910; Spearman, 1910)

$$R^2(k) = \frac{k\,\bar{r}^2}{1 + (k-1)\,\bar{\rho}}, \qquad K_{1/2} = \frac{1-\bar{\rho}}{\bar{\rho}}, \tag{5}$$

and the equal-variance binormal map turns correlation into AUC (Green & Swets, 1966),

$$\mathrm{AUC}(k) = \Phi\big(\delta(k)/\sqrt{2}\big), \qquad \delta(k) = \frac{R(k)}{\sqrt{\pi(1-\pi)}\,\sqrt{1-R(k)^2}}, \tag{6}$$

where π is the label prevalence and Φ the standard normal distribution function. Two statistics fix how performance saturates with the number of questions. On FakeSV $K_{1/2} = 18$; the curve comes within .02 of its asymptote at 200 questions, and random subsets stay within .009 of it (Fig. 3, right); FakeTT's statistics likewise reproduce FakeTT's curve. The law describes the equal-weight composite; the fitted readout reweights the same questions and gains a little more. The structure of the bank tells the same story. Removing any design family, any elicitation style or either generator's half changes in-domain AUC and cross-model loss by at most .005 and cross-dataset AUC by at most .018 relative to removing as many random questions. Apart from fusing a second reader, which gains about .01 AUC, less than the measurement model predicts, none of the attempted improvements, question selection among them, raises AUC over the full linear readout (App. A).

Where reading thins. Transfer weakens in two places. The first is the smallest model. InternVL2.5-1B reaches .887 fitted on itself but only .758 with the coefficients of Qwen2.5-VL-7B (Qwen below), and it loses .113 on average as a target against .041 as a source: its readings carry task information, read in a way not proportional to the other models', so whether a model can serve as a reader

depends on whether it reads like the others, not only on how strong it is alone. The second is the answer channel. An open model's logit difference carries direction and magnitude, whereas a closed interface or a questionnaire returns a few discrete levels, a many-to-one map (Cover & Thomas, 2006). On one model the channel can be isolated (App. E). At 30 questions, binning Qwen's own readings into five levels by per-question magnitude leaves its self-fit almost unchanged, while re-binning them to the answer distribution GPT-5.5 actually produces lowers it to .766, near GPT-5.5's own .737, because GPT-5.5 puts nearly all its answers on one level for many questions: what discards information is the collapse of answers, not the number of levels. At 100 questions the same re-binning gives .815, below GPT-5.5's .854, so there GPT-5.5's answers carry more than their distribution alone allows. With coefficients fitted on Qwen's binarized answers to the same questions and deployed unchanged (Fig. 3, left), GPT-5.5 reaches .767 with 30 questions and .872 with 100, and the two annotators reach .710 with 30. Against the criteria frozen in advance (App. E), the 30-question arm fails and the annotators are partial, while the 100-question arm, registered without a gate of its own, clears the 30-question criterion: a discrete answer carries less per question, and more questions recover part of it.

5 Questions Carry the Meaning

Correspondence. Between two models never aligned, the only thing dimension k has in common is the text of question k. Permuting the correspondence between questions and readout dimensions, while keeping the dimension count and every marginal distribution, drops cross-dataset AUC from .877 to .539 and cross-model transfer to near chance. Generating questions per video, the core of SAFE, leaves dimension k with no fixed meaning; a SAFE-style reimplementation on the same backbone and inputs reaches only .737 across datasets, where a readout on the first 204 questions of the bank reaches .850 on the same videos. Readings also follow the meaning of the question. On one model and 100 questions, paraphrases leave per-question readings correlated at .97, and negation flips 68% of them, to a degree ordered by design family and strongest for checkability (App. F).

Compatibility and validity. If fixed questions alone established shared coordinates, questions unrelated to veracity should transfer too. With the same two generators, input quotas and freezing procedure, we drafted a bank of questions with no relation to veracity, such as whether the frames show something blue or whether the title mentions food, with criteria pre-registered before scoring (App. G). Across models it still transfers, if less well than the real bank: on one pair of models reading identical text, its sheets are recognized at .863 and .786, against .971 and .964 for matched real-bank subsets, and raw transfer loses .033 AUC, against .017 (Fig. 4). Across datasets the gap is larger: the control falls to .599, against .807 for real-bank subsets of the same size and input pattern and .484 for its shuffled reference, and the stage-two questions, which read frames or excerpts, show the same ordering with a smaller gap (.674 against .834). Compatibility comes from fixing the coordinates; validity, measured as discrimination across datasets, comes from their content. Against the frozen criteria, the in-domain and per-question endpoints fail, because the control is nearly as discriminative in domain (.802 against .876), and the cross-dataset endpoint is partial. The control also shows something about the benchmark: with the unrelated bank at .802 in-domain and a character n-gram classifier at .839, in-domain AUC cannot tell veracity cues from topic where topic and label are associated; performance across datasets can, insofar as the two benchmarks do not share that topic structure.

Accuracy. With the corroboration channel, the same questions, with the backbone frozen, come within a point of a purpose-built fine-tuned detector and hold up better across events (Tab. 1). On the official temporal split, the readout reaches 89.3 accuracy and FakeSV-VLM (Wang et al., 2025b), reproduced with its released configuration, 90.0; the paired difference has a confidence interval spanning zero. The inputs differ: the detector reads one video with its event keyword, while the readout also reads earlier same-event videos. From the video alone the readout reaches 87.1, 2.9 points below the detector and slightly above a hidden-state probe on the same backbone (86.4), a lead that reverses under development cross-validation (App. H). Under the official event-disjoint protocol, which targets shortcut learning and event-specific leakage (Geirhos et al., 2020; Wang et al., 2018), FakeSV-VLM retrained per fold falls to 80.0 while the readout keeps 83.6 (82.3 without corroboration), close to NEED (Qi et al., 2023b) and DugFND (Gong et al., 2025), which use debunking annotations and social graphs. Across datasets, Chinese-fitted coefficients reach AUC .877

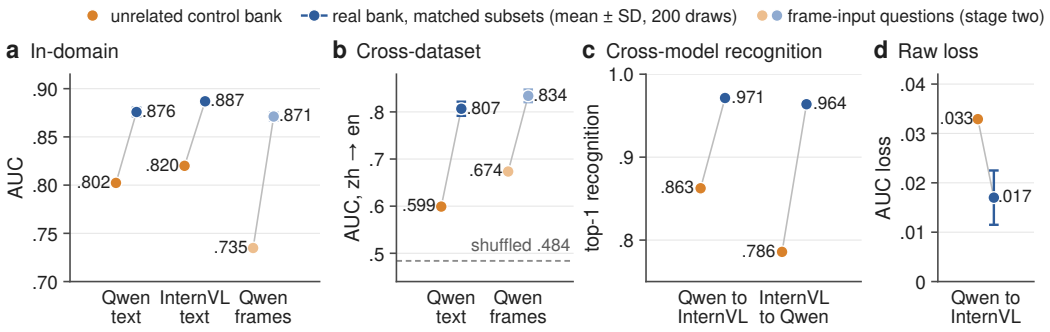


Figure 4: A bank unrelated to the task against matched subsets of the real bank (mean ± SD over 200 subsets; lighter markers: stage two, questions reading frames or same-event excerpts). (a) In-domain AUC; (b) cross-dataset AUC, Chinese-fitted coefficients on English FakeTT, with the shuffled reference; (c) cross-model recognition; (d) raw-reading transfer loss. Qwen: Qwen2.5-VL-7B; InternVL: InternVL2.5-8B. Recognition and transfer are largely shared; cross-dataset discrimination is mostly the real bank's.

Table 1: Task utility: the official FakeSV temporal test set (542 videos), the official event-disjoint five folds (event-cluster bootstrap CI for the readout, mean and spread over folds for the retrained baseline), and transfer between FakeSV and FakeTT on the full pools, with readout thresholds set by the source's label prior (RADAR uses test-time adaptation). Published figures are from the cited papers; n/a: not applicable or not reported.

System	Accuracy	AUC	Note
Temporal split, official test set (542 videos)			
FakeSV-VLM (Wang et al., 2025b)	90.0	n/a	reproduced; published 90.22
DugFND (Gong et al., 2025)	91.0	n/a	published; social graph
Readout w/ corroboration	**89.3**	**.944**	all 768 questions
Readout w/o corroboration	87.1	.922	single video
Hidden-state probe, same backbone	86.4	.920	linear probe
Direct answer, same backbone	63.5	n/a	text fields and frames
Event-disjoint, five folds			
FakeSV-VLM retrained per fold	80.0 ± 1.5	n/a	fine-tuned per fold
Readout w/ corroboration	**83.6** [81.1, 85.8]	n/a	all 768 questions
Readout w/o corroboration	82.3 [79.8, 84.5]	n/a	single video
NEED (Qi et al., 2023b)	84.6	n/a	published; debunking annotations
DugFND (Gong et al., 2025)	85.5	n/a	published; social graph
Across datasets, full pools, no target labels			
Readout w/ corroboration	77.7	.877	FakeSV→FakeTT
Readout w/o corroboration	75.0	.848	FakeSV→FakeTT
FakeSV-VLM, event-disjoint	78.2	n/a	FakeSV→FakeTT
RADAR (Lang et al., 2026)	64.7	n/a	FakeSV→FakeTT, published
Readout w/ corroboration	77.5	.881	FakeTT→FakeSV
RADAR (Lang et al., 2026)	67.6	n/a	FakeTT→FakeSV, published

on English FakeTT and .881 in reverse and, with a threshold set by the source prior, 10 to 13 accuracy points above the published test-time adaptation of RADAR (Lang et al., 2026); on FakeTT the corroboration channel reads all same-event videos, and without it the forward AUC is .848. The three kinds of reusable object differ in what they need to move. A fine-tuned detector's weights are the model; applied unchanged to FakeTT, the event-disjoint checkpoint reaches 78.2 accuracy, level with the readout's 77.7. A hidden-state probe on similarities to 300 unlabelled anchor videos (Moschella et al., 2023) crosses from Qwen to InternVL2.5-8B at .869 (.865 on Qwen, .870 for the question readout), but it needs both models' hidden states, and across datasets a probe reaches only .774. The question readout needs only answers, and so extends to a closed interface and to people.

A second construct. We rebuilt the recipe for harmful-meme detection on Chinese ToxiCN-MM (Lu et al., 2024) and English HarMeme (Pramanick et al., 2021), with 220 questions drafted by the same generators and frozen before scoring. With the backbone frozen and only the linear readout trained, the readout reaches AUC .848 and .901, placing it on each benchmark's own metric between the published zero-training methods and purpose-trained systems (on HarMeme, 84.2 accuracy against 83.9 for U-CoT+, about one test meme; App. I). Within each dataset the representation travels as before: coefficients move from 7B to 3B unchanged with a loss under .02 AUC, a change of prompting language costs under .02, and the aggregation law holds. Between the two meme datasets the readout does not transfer and performs at the level of the shuffled reference. A likely reason lies in the task. Verification cues for video news, such as whether corroboration exists or whether the narration matches the footage, are largely shared by the two platforms, whereas what makes a meme harmful depends on its language and culture. The representation travels on both sides; what does not cross datasets is the relation between the questions and the label.

6 Discussion and Conclusion

These results suggest another way to upgrade models. A downstream system needs to keep only three text files, the bank, the per-question statistics and the readout coefficients; replacing the model took about 100 unlabelled videos to estimate the new model's statistics in the pair we tested (App. C), and moving between the two video datasets took one threshold. More generally, the recognition and transfer of Sec. 4 suggest that natural-language questions may serve as a coordinate system shared by the vision-language models we tested.

This paper establishes these properties on one construct and replicates the within-dataset ones on a second; more constructs are the natural next step. Discrete-answer readers currently recover information through more questions, and answer formats of higher resolution are worth studying. Corroboration questions read whether corroborating material is available, and content-swap tests would establish more precisely what they measure. How the test sets were used and the checks not yet run are stated in App. J.

All that is needed is a fixed set of questions, a frozen model that answers them much as the others do, and one logistic regression. A representation built this way stays stable across model families, generations and prompting languages, and extends, within the bandwidth of the answer channel, to a closed interface and to people; its aggregation behaviour is described by two statistics; with corroboration, its accuracy comes within a point of heavy fine-tuning. When a representation is defined by language rather than by weights, the bank, the standardization statistics and the fitted coefficients survive the replacement of the model. Models are mortal. Measurements are not.

Reproducibility Statement

The full question bank (Chinese and English wording and input declarations), the freezing hashes, the reading matrices of all eleven readers, the readout coefficients, the negative-control bank and its pre-registration, and all evaluation scripts will be released. Official dataset splits are used as published, and App. J states which choices were made after test results had been seen. Reproducing the open-model results requires no question generator and no closed model; the GPT-5.5 arm used the snapshot gpt-5.5-2026-04-23.

Ethics Statement

The study uses two public misinformation benchmarks and two public meme benchmarks under their release terms and collects no data from platform users. Figure examples do not reproduce identifiable individuals. The two human annotators were informed volunteers and received neither labels nor model outputs. Misinformation detectors carry dual-use risk, including the suppression of legitimate speech; a representation whose every decision decomposes into named questions makes such decisions auditable, which we regard as a mitigation and not an exemption. One of the largest coefficients points from a self-declared dramatization toward fake (App. F), so an honest disclaimer can raise a video's score; coordinates of this kind need review before any deployment.

USE OF GENERATIVE AI

We used generative AI tools to generate candidate questions for the main, negative-control and meme question banks, screen negative-control candidates, identify relevant literature, and assist with translation from an initial Chinese draft and language polishing. All AI-assisted outputs were reviewed and revised by the authors, and all AI-suggested references were individually verified against their original sources. The authors take full responsibility for the final content of the paper.

REFERENCES

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14940–14949, 2022.

Ayushi Agarwal. Cross-architecture steering transfer in language models: A systematic empirical study. *arXiv preprint arXiv:2608.05164*, 2026.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. COSMOS: Catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14084–14092, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Vinamra Benara, Chandan Singh, John X. Morris, Richard Antonello, Ion Stoica, Alexander G. Huth, and Jianfeng Gao. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. FakingRecipe: Detecting fake news on short video platforms from the perspective of creative process. In *ACM International Conference on Multimedia (MM)*, 2024.

Ivo Bueno, Lea Hirlimann, and Enkelejda Kasneci. MemeScouts@LT-EDI 2026: Asking the right questions – prompted weak supervision for meme hate speech detection. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pp. 198–205, 2026.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 321–332, 2022.

Junyi Chen, Qian Liu, Jing Sun, and Yi Zhang. TrueLens: Video fake news detection with dual level evidence gathering and consolidation. In *Proceedings of the ACM Web Conference (WWW)*, 2026.

Xinhao Chen, Dongxin Wen, and Decheng Zuo. Towards detecting Chinese harmful memes with fine-grained explanatory augmentation. *Electronics*, 14(17):3504, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Xiang Cheng, Raveesh Mayya, and João Sedoc. To err is human; to annotate, SILICON? Toward robust reproducibility in LLM annotation. *arXiv preprint arXiv:2412.14461*, 2024.

Sangwoo Cho, Kushal Chawla, Pengshan Cai, Zefang Liu, Chenyang Zhu, Shi-Xiong Zhang, and Sambit Sahu. Ask, don't judge: Binary questions for interpretable LLM evaluation and self-improvement. In *Second Workshop on Compositional Learning at ICML (non-archival); arXiv:2606.27226*, 2026.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Jules L. Ellis and Klaas Sijtsma. Proof of reliability convergence to 1 at rate of Spearman–Brown formula for random test forms and irrespective of item pool dimensionality. *Psychometrika*, 89 (3):774–795, 2024.

Jean Feng, Avni Kothari, Luke Zier, Chandan Singh, and Yan Shuo Tan. Bayesian concept bottleneck models with LLM priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Haisong Gong, Bolan Su, Xinrong Zhang, Jing Li, Qiang Liu, Shu Wu, and Liang Wang. Mining the social fabric: Unveiling communities for fake news detection in short videos. *arXiv preprint arXiv:2508.07992*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *International Conference on Machine Learning (ICML)*, 2026.

Sid Ali Hamideche, Louis-Adrien Dufrène, Quentin Lampin, and Guillaume Larue. How far do simple transformations translate across text embedding models? *arXiv preprint arXiv:2608.05980*, 2026.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1803–1812, 2017.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM Web Conference (WWW)*, 2025.

Haonan Huang. The yes-no bias of large language models reflects answer order and wording, not shifts in moral judgment. *arXiv preprint arXiv:2607.05552*, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Karl G. Jöreskog. Statistical analysis of sets of congeneric tests. *Psychometrika*, 36(2):109–133, 1971.

Su-Hyeon Kim and Yo-Sub Han. Cross-family universality of behavioral axes via anchor-projected representations. *arXiv preprint arXiv:2605.09875*, 2026.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. Departmental Papers (ASC), University of Pennsylvania, 2011.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Jian Lang, Rongpei Hong, Ting Zhong, Yong Wang, and Fan Zhou. Nip rumors in the bud: Retrieval-guided topic-level adaptation for test-time fake news video detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. Weakly supervised veracity classification with LLM-predicted credibility signals. *EPJ Data Science*, 2025.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zehao Li, Hongwei Yu, Hao Jiang, Qiang Sheng, Yilong Xu, Baolong Bi, Yang Li, Zhenlong Yuan, Yujun Cai, and Zhaoqi Wang. FactGuard: Agentic video misinformation detection via reinforcement learning. *arXiv preprint arXiv:2602.22963*, 2026.

Hao Lin and Yongjun Zhang. Navigating the risks of using large language models for text annotation in social science research. *Social Science Computer Review*, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Fang Liu, Yili Li, Jian Lang, Rongpei Hong, and Fan Zhou. Enhancing fake news video detection with self-driven question-answer from LMMs. *Information Processing & Management*, 63(2), 2026.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Jingyuan Liu. What is actually being annotated? Inter-prompt reliability as a measurement problem in LLM-based social science labeling. *arXiv preprint arXiv:2604.16413*, 2026.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. Towards comprehensive detection of Chinese harmful memes. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 13302–13320, 2024.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6801–6817, 2021.

Pablo Marcos-Manchón and Lluís Fuentemilla. Shared representations in brains and models reveal a two-route cortical organization during scene perception. *Communications Biology*, 2026.

Pablo Marcos-Manchón, Rishi Jha, and Lluís Fuentemilla. Platonic representations in the human brain: Unsupervised recovery of universal geometry. *arXiv preprint arXiv:2605.20496*, 2026.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 23817–23839, 2025.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning (ICML)*, 2023.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. Activation space interventions can be transferred between large language models. In *International Conference on Machine Learning (ICML)*, 2025.

OpenAI. GPT-5.5 system card. `https://openai.com/index/gpt-5-5-system-card/`, 2026.

Fengjun Pan, Xiaobao Wu, Tho Quan, and Anh Tuan Luu. Read as you see: Guiding unimodal LLMs for low-resource explainable harmful meme detection. In *Proceedings of the ACM Web Conference (WWW)*, 2026.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tobia Poppi, Silvia Cappelletti, Sara Sarto, Florian Schiffers, Garin Kessler, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Do models share safety representations? Cross-model steering for safe visual generation. *arXiv preprint arXiv:2606.05290*, 2026.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2783–2796, 2021.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14444–14452, 2023a.

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. Two heads are better than one: Improving fake news video detection by correlating with neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11947–11959, 2023b.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.

Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19364–19373, 2022.

Filippo Rinaldi, Aniello Panariello, Giacomo Salici, Angelo Porrello, and Simone Calderara. Transporting task vectors across different architectures without training. In *International Conference on Machine Learning (ICML)*, 2026.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3): 271–295, 1910.

Oscar Thorsted Svendsen, Nikolaj Holst Jakobsen, Fabian Mager, and Hiba Nassar. Improving relative representations with learned anchors and whitened inner products. *arXiv preprint arXiv:2605.30596*, 2026.

Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Luke Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. A guide to misinformation detection data and evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5801–5809, 2025.

Shaydanay Urbani. Verifying online information: The absolute essentials. First Draft, 2019.

W3C Credible Web Community Group. Credibility signals. W3C Draft Community Group Report, 21 October 2018, 2018.

Jinyuan Wang, Ningyuan Deng, and Yi Yang. Assessing and mitigating miscalibration in LLM-based social science measurement. *arXiv preprint arXiv:2605.11954*, 2026.

Junxi Wang, Jize Liu, Na Zhang, and Yaxiong Wang. Consistency-aware fake videos detection on short video platforms. In *International Conference on Intelligent Computing (ICIC)*, 2025a.

Junxi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. FakeSV-VLM: Taming VLM for detecting fake short-video news via progressive mixture-of-experts adapter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4782–4798, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 849–857, 2018.

Yongdong Wang, David J. Veltkamp, and Bruce R. Kowalski. Multivariate instrument standardization. *Analytical Chemistry*, 63(23):2750–2756, 1991.

Jerome J. Workman. A review of calibration transfer practices and instrument differences in spectroscopy. *Applied Spectroscopy*, 72(3):340–365, 2018.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

Haochen Zhao, Yongxiu Xu, Xinkui Lin, Dong Xie, Jiarui Lu, Yuqi Qian, Yubin Wang, Hongbo Xu, and Gaopeng Gou. Sparse evidence can suffice: Agentic evidence seeking for multimodal video misinformation detection. *arXiv preprint arXiv:2607.18080*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Tyler Zhu, Tengda Han, Leonidas Guibas, Viorica Pătrăucean, and Maks Ovsjanikov. Dynamic reflections: Probing video representations with text alignment. In *International Conference on Learning Representations (ICLR)*, 2026.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10817–10826, 2024.

APPENDIX OVERVIEW

The appendix follows the order of the main text. The method of Sec. 3 is supported by App. A (how the questions were drafted and frozen, what each design choice contributes and why no questions are selected), App. B (the derivations behind Eqs. (3) to (6) and their assumptions) and App. C (readers, prompts, data, standardization statistics, the cross-dataset threshold, the corroboration channel and the reproduced baseline). Section 4 is supported by App. D (the full transfer matrices and the complete recognition and language numbers) and App. E (the closed API and the human annotators), and Sec. 5 by App. F (controls, paraphrase and negation, and what the readout reads), App. G (the negative-control bank), App. H (accuracy, cross-dataset details and cost) and App. I (the second construct). Appendix J states how the test sets were used and lists the checks not yet run.

A THE QUESTION BANK

This appendix describes how the bank of Sec. 3 was built and what each of its design choices contributes. It covers the literature behind the design families, drafting, freezing, coding against professional verification standards, design ablations and the evidence for using all questions.

Task-side literature. The design families come from prior work on short-video and multimodal misinformation detection, which falls into classes by the resources it uses. Content-only fine-tuning: FakeSV-VLM (Wang et al., 2025b), FakingRecipe (Bu et al., 2024) and CA-FVD (Wang et al., 2025a). Content plus social context such as comments and publisher features: SV-FEND (Qi et al., 2023a) and opinion-evolution modelling with prompting and diffusion (Zong et al., 2024). Retrieval-augmented chain-of-thought reasoning with a multimodal LLM: ExMRD (Hong et al., 2025). Event and social graphs: NEED (Qi et al., 2023b) and DugFND (Gong et al., 2025). Retrieval and agents with labelled precedents, evidence search or tool use: TrueLens (Chen et al., 2026), SIEVE (Zhao et al., 2026) and FactGuard (Li et al., 2026). Per-sample free-form question answering: SAFE (Liu et al., 2026). Event-adversarial training: EANN (Wang et al., 2018). Out-of-context image-text pairs: COSMOS (Aneja et al., 2023) and NewsCLIPpings (Luo et al., 2021). The classes use different resources; the main text compares accuracy with the content-only class and lists systems of other classes with their resources noted (Tab. 1).

Drafting. The 768 questions were drafted by two generators from different families, DeepSeek-V4-Pro and GPT-5.5 (snapshot gpt-5.5-2026-04-23); neither family supplies any of the eleven logit readers of Sec. 4, but the same GPT-5.5 snapshot later serves as the discrete-answer reader, so App. E tests whether it answers the questions it drafted differently. Drafting used 21 prompt cells: five design families crossed with four elicitation styles (direct, adversarial, contrastive, authenticating)

give 20 cells of 32 questions, plus one free exploration cell of 128, and each generator drafted half of every cell independently (Fig. A.1). Every question carries Chinese and English wording, a declaration of the inputs it reads (frames, title, on-screen text, speech transcript, earlier same-event excerpts) and a one-sentence rationale. The generators never saw any reader's output, and the prompt templates are released. Table A.1 states what each family asks about, with one example question each.

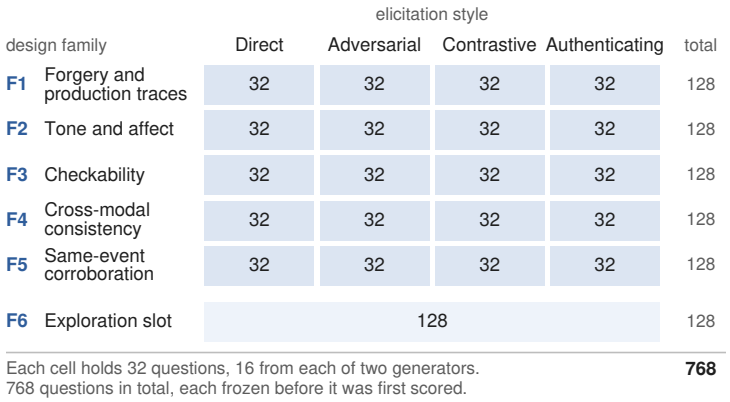


Figure A.1: The design of the bank: five design families crossed with four elicitation styles, 32 questions per cell (16 from each generator), and an exploration slot of 128; each of the 768 questions was frozen before it was first scored.

Table A.1: The five design families and the exploration slot, with one example question each (verbatim from the bank).

Family (128 each)	What it asks about	Example question
F1 forgery traces	Editing and synthesis evidence in frames and audio	"Are there multiple layers of overlaid text at the same time, with fonts, colors, or positions that clearly belong to different editing layers?"
F2 tone and affect	Inflammatory tone and affective expression patterns	"Does the title contain an imperative command, such as 'Watch' or 'Don't miss'?"
F3 checkability	Verifiable numbers, dates, places and identifiers	"Does the title or description contain a formal reference number, such as a case number, notice number, document number, flight number, train number, or order number?"
F4 cross-modal consistency	Whether title, speech and visuals support each other	"Does the speech transcript describe specific actions, damage, casualties, arrests, or outcomes that the video frames themselves do not visibly support?"
F5 corroboration	Whether earlier same-event content corroborates the video	"Do any of the other videos show emergency response vehicles (ambulances, fire trucks, police cars) at the event scene?"
F6 exploration slot	Open slots for questions outside the designed families	"Does the video contain a statement that it is a dramatization, reenactment, or staged recreation?"

Freezing. Question order was set by randomized blocks: cells take turns, and generator and call index alternate, so that batch number is not confounded with any cell. The order and full wording were registered by SHA-256 before the questions concerned were scored: first the 680 base questions, then 88 exploration-slot questions appended at the end under the same rule. Using the first k questions is a budget parameter only.

Coding against professional standards. Every question was coded against the W3C Credibility Signals list (W3C Credible Web Community Group, 2018) and First Draft's five verification pillars

(Urbani, 2019) as exact, partial or no match, with a reverse blind-spot table and a 10% spot check. Of the 640 design-family questions, 549 (85.8%) match at least one standard, against a pre-frozen threshold of 60%; the forgery family matches provenance signals at 100%, and the exploration slot matches at 50.8% by design. The five First Draft pillars all have counterparts in the bank; 16 of the 36 W3C signals have none, seven of them web-page, advertising and academic-citation signals that do not exist for short video. Of the 154 questions matching neither standard, 55 concern platform-internal corroboration and checkable identifiers, practices that exist in verification work but that neither standard lists.

Design ablations. Each design choice of Sec. 3 can be tested on the readings already collected: every question is answered in its own forward pass, so removing questions removes columns, and only the standardization statistics and the readout are recomputed. The protocol is that of the negative-control references: the 1,200-video panel, event-grouped five-fold CV, per-question standardization and $C = .01$; cross-model deployment from Qwen2.5-VL-7B to InternVL2.5-8B after standardization; Chinese-fitted coefficients deployed raw on the 598-video FakeTT panel. Each subset is compared with 20 random subsets of the full bank of the same size. Table A.2 removes one family, one elicitation style or one generator's half, and Tab. A.3 keeps one family or one style alone. Removing a family or a style moves in-domain AUC and cross-model loss by at most .005 and cross-dataset AUC by at most .011 relative to the random reference, so no single block is a pillar. Used alone, the families differ widely, which is what a design grid is for: checkability alone matches a random subset of its size in-domain and across models, corroboration alone barely recognizes videos because it reads excerpts of other videos, and forgery traces alone transfer worst across models and datasets. Every single style stays within about two standard deviations of a random subset of its size, so several styles serve coverage rather than accuracy. The DeepSeek and GPT-5.5 halves agree in-domain and across models and sit on either side of the reference across datasets (.858 and .823, about 1.5 SD each).

Table A.2: Design ablations, removing one family or style, or keeping one generator's half. In-domain: Qwen2.5-VL-7B event-grouped CV AUC. Cross-model: AUC deployed on InternVL2.5-8B after standardization, and the loss against its own fit. Cross-dataset: raw deployment on FakeTT. Reference: 20 random subsets of the same size (mean ± SD).

		In-domain		Cross-model			Cross-dataset	
Subset	k	AUC	Ref	AUC	Loss	Ref loss	AUC	Ref
Full bank	768	.885	n/a	.870	.024	n/a	.832	n/a
w/o F1 forgery traces	640	.888	$.887 \pm .001$	.872	.024	$.022 \pm .002$	.830	$.836 \pm .006$
w/o F2 tone and affect	640	.884	$.887 \pm .001$	.870	.026	$.022 \pm .002$	.838	$.836 \pm .006$
w/o F3 checkability	640	.882	$.887 \pm .001$	.868	.026	$.022 \pm .002$	.847	$.836 \pm .006$
w/o F4 cross-modal consistency	640	.887	$.887 \pm .001$	.876	.018	$.022 \pm .002$	.826	$.836 \pm .006$
w/o F5 corroboration	640	.883	$.887 \pm .001$	.872	.021	$.022 \pm .002$	.831	$.836 \pm .006$
w/o F6 exploration slot	640	.890	$.887 \pm .001$	.875	.023	$.022 \pm .002$	.831	$.836 \pm .006$
w/o direct style	608	.888	$.886 \pm .001$	.876	.023	$.022 \pm .003$	.829	$.838 \pm .007$
w/o adversarial style	608	.885	$.886 \pm .001$	.872	.023	$.022 \pm .003$	.840	$.838 \pm .007$
w/o contrastive style	608	.885	$.886 \pm .001$	.873	.020	$.022 \pm .003$	.841	$.838 \pm .007$
w/o authenticating style	608	.885	$.886 \pm .001$	.872	.021	$.022 \pm .003$	.836	$.838 \pm .007$
DeepSeek half only	384	.886	$.889 \pm .003$	.875	.021	$.018 \pm .003$	.858	$.840 \pm .011$
GPT-5.5 half only	384	.891	$.889 \pm .003$	.883	.016	$.018 \pm .003$	.823	$.840 \pm .011$

Using every question. Two kinds of evidence support a readout that uses every question. Of the attempted improvements in Tab. A.4, only fusing a second reader raised AUC, by about .01 over the better single reader and less than the measurement model predicts (App. B), and four selection-bias experiments show why selected subsets look better than they are (Tab. A.5). Greedy source-side selection of 27 questions lowers cross-dataset AUC by .033 (bootstrap 95% CI $[-.051, -.011]$); one reader's top-30 lists have a fold-to-fold Jaccard of only .37 to .52; the coefficient mass of the full readout is spread over hundreds of questions, the heaviest 30 carrying 12.5 to 12.8%; and letting selection touch the evaluation pool inflates one comparison from $+.022$ (CI containing zero) to $+.061$.

Table A.3: Design ablations, keeping one block alone. In-domain AUC, cross-model loss and cross-dataset AUC as in Tab. A.2, without the deployed cross-model AUC; the last column is top-1 recognition from Qwen2.5-VL-7B to InternVL2.5-8B.

		In-domain		Cross-model		Cross-dataset		
Subset kept	k	AUC	Ref	Loss	Ref loss	AUC	Ref	Recognition
F1 forgery traces	128	.845	$.886 \pm .004$	.056	$.015 \pm .005$	.719	$.832 \pm .017$	.553
F2 tone and affect	128	.857	$.886 \pm .004$	.025	$.015 \pm .005$	.764	$.832 \pm .017$	.894
F3 checkability	128	.886	$.886 \pm .004$	.017	$.015 \pm .005$	.784	$.832 \pm .017$	.876
F4 cross-modal consistency	128	.872	$.886 \pm .004$	.031	$.015 \pm .005$	.831	$.832 \pm .017$	.746
F5 corroboration	128	.827	$.886 \pm .004$	.049	$.015 \pm .005$	.826	$.832 \pm .017$	.494
F6 exploration slot	128	.868	$.886 \pm .004$	.014	$.015 \pm .005$	.815	$.832 \pm .017$	.969
direct style	160	.881	$.888 \pm .004$	.023	$.017 \pm .003$	.837	$.836 \pm .013$	.969
adversarial style	160	.890	$.888 \pm .004$	.017	$.017 \pm .003$	.810	$.836 \pm .013$	.947
contrastive style	160	.892	$.888 \pm .004$	.015	$.017 \pm .003$	.834	$.836 \pm .013$	.962
authenticating style	160	.893	$.888 \pm .004$	.016	$.017 \pm .003$	.841	$.836 \pm .013$	.962

Table A.4: Attempted improvements over the full bank with a linear readout, none adopted.

Attempted improvement	Outcome
Non-linear readouts (GAM, boosting, trees)	no gain
Six in-domain pruning schemes	accuracy up, AUC down (threshold effect; Tab. A.5)
Greedy source-side selection	cross-dataset AUC $-.033$
External-arm abstention	no gain
Paraphrase-averaged rescoring	no gain
Knowledge-axis injection	no gain
EM prior correction	degenerates under covariate shift (App. C)
Reader ensembling	about $+.01$ AUC over the better single reader, below the registered predictions (App. B)

Table A.5: Four selection-bias experiments.

Selection-bias experiment	Result
(i) six in-domain pruning schemes	all raise accuracy while lowering AUC: threshold effect, not information gain
(ii) greedy 27-question source-side selection	cross-dataset AUC $-.033$ (bootstrap 95% CI $[-.051, -.011]$)
(iii) stability of top-30 lists	fold-to-fold Jaccard only .37 to .52: a "best subset" is not a stable object
(iv) selection touching the evaluation pool	inflates the same AUC comparison to $+.061$; isolation shrinks it to $+.022$ (CI contains 0)

B Derivations

This appendix derives the equations used in Secs. 3 and 4 and states what each needs: Eq. (3) from Eq. (2), the proportional-signal statement about Eq. (4), the aggregation law and its approximation, why the AUC of .944 in Tab. 1 exceeds the asymptote of the law, and the fusion prediction that follows from the measurement model.

Equation (3). Under Eq. (2), with the deviation uncorrelated with the tendency over videos, the standardized reading is $z_{mk} = (g_{mk}(s_k - \mu_{s_k}) + \varepsilon_{mk})/(g_{mk}^2\sigma_{s_k}^2 + \sigma_{\varepsilon,mk}^2)^{1/2} = \lambda_{mk}\tilde{s}_k + \tilde{\varepsilon}_{mk}$, where μ_{s_k} and σ_{s_k} are the mean and standard deviation of the tendency over videos and $\lambda_{mk} = g_{mk}\sigma_{s_k}/(g_{mk}^2\sigma_{s_k}^2 + \sigma_{\varepsilon,mk}^2)^{1/2}$; λ_{mk}^2 is the share of the reading's variance that the tendency explains, the indicator reliability of structural equation modelling. With deviations uncorrelated across

readers, two readers' readings on one question correlate at $\lambda_{mk}\lambda_{m'k}$, the attenuation formula of Spearman (1904). A paraphrase correlation within one reader also counts the reader's own deviation, which paraphrases share, and so bounds λ_{mk}^2 from above. The paragraph on single questions and aggregates in Sec. 4 uses both relations.

Proportional signal. Deploying w_A on reader B gives the signal $\sum_k w_{Ak}\lambda_{Bk}\tilde{s}_k(x)$. If $\lambda_{Bk} = c\,\lambda_{Ak}$ for every question, this equals c times the source signal $\sum_k w_{Ak}\lambda_{Ak}\tilde{s}_k(x)$, and the two rank content identically. The deviation term $\sum_k w_{Ak}\tilde{\varepsilon}_{Bk}$ is uncorrelated with the signal over videos, and its variance, $\sum_{k,l} w_{Ak}w_{Al}\,\mathrm{Cov}(\tilde{\varepsilon}_{Bk}, \tilde{\varepsilon}_{Bl})$, stays small beside the signal's when B's deviations are weakly correlated across questions; the ranking then follows the signal. In general the deployed signal equals c times the source signal plus $\sum_k w_{Ak}(\lambda_{Bk} - c\,\lambda_{Ak})\tilde{s}_k(x)$, a term that is small when the questions carrying weight keep their relative reliability, and whose parts also cancel when their deviations have different signs.

The aggregation law and its assumptions. Under standardized readings, the equal-weight composite of k questions has variance $k + k(k-1)\bar{\rho}$ and covariance $k\bar{r}$ with the label, which gives $R(k) = k\bar{r}/\sqrt{k + k(k-1)\bar{\rho}}$, Eq. (5). For a fixed subset with its own statistics this is an identity. Replacing subset statistics by bank means is exact in expectation for numerator and denominator separately but not for their ratio; the error comes from heterogeneity of question validity, is largest at small k (at $k = 1$, $\mathbb{E}[R^2] = \bar{r}^2 + \mathrm{Var}(r)$), and shrinks as the sums average over questions. The other assumptions are random draws of questions, equal weights, direction alignment by the sign of r_k, and the equal-variance Gaussian score distributions of the binormal map (Green & Swets, 1966). $K_{1/2}$ is the half-saturation constant of $R^2(k)$, not of the AUC curve. In Fig. 3 the statistics, the direction alignment and the empirical curve come from the same development pool, so the agreement checks the functional form; the out-of-sample test is extrapolation across datasets, where predicting one dataset's curve from the other's $(\bar{r}, \bar{\rho})$ deviates by at most .03 in both directions (.009 and .023 within pool).

Reconciling .944 and .881. The asymptote .881 bounds equal-weight random composites under the exchangeable model. Three numbers line up under one protocol. The equal-weight composite of all 768 questions under event-grouped CV reaches .868 (theory .876). Fitted weights under the same protocol (L2 logistic regression, $C = .003$) reach .891, a gain of .023 from the heterogeneity of question validity and the inter-question covariance that the exchangeable model flattens. The official temporal test set, a different pool and protocol in which events cross the train and test boundary, gives .944. The main-table figure exceeds the asymptote because of fitted weights and a change of protocol.

Fusion prediction. Averaging two readers' standardized readings raises the effective validity of question k by $\sqrt{2/(1+\rho_k)}$ when the two readers' deviations are uncorrelated, where ρ_k is the two readers' correlation on that question; applying the mean uplift to the single-reader discriminability gives the prediction. The predictions were written to hashed, timestamped files before the fused readouts were computed, with no pass band. For Qwen2.5-VL-7B and InternVL2.5-8B the prediction was .919 and the observed value .904. For the second pair, Qwen2.5-VL-7B and Qwen3-VL-8B, the registered value .9015 equals the raw prediction .9165 minus a .015 correction fixed from the first pair's error; the observed value was .8995, so the raw error is $-.017$ and the corrected error $-.002$. For GPT-5.5 with Qwen's signs a numerical bug corrupted the registered prediction; computed after the fact, the formula predicts .862 against an observed .842. A composite-level version scales the mean validity by γ, the mean over questions of $\sqrt{2/(1+\rho_k)}$, and the mean inter-question correlation by γ^2 inside Eq. (5), which leaves the asymptote unchanged. On the 1,200-video panel with 675 common questions (Qwen2.5-VL-7B and InternVL2.5-8B, median $\rho_k = .73$, $\gamma = 1.09$), the per-question validity of the fused sheet is .161 against the predicted $\gamma\bar{r} = .160$, while the fused inter-question correlation rises less than predicted (.061 against $\gamma^2\bar{\rho} = .066$). The equal-weight fused AUC is .883, the composite-level prediction .867 and the single-question uplift applied to δ .887: the per-question scaling holds, and the two forms bracket the observed gain from below and above.

C READERS, DATA AND PROTOCOL

This appendix gives the protocol behind Sec. 3: the readers and what they see, the prompt templates, the data and splits, how the standardization statistics are estimated and how many unlabelled videos they need, the cross-dataset threshold, the construction of the corroboration channel, and the reproduction of the fine-tuned baseline.

Readers. Table C.1 lists every reader and its ingestion mode.

Table C.1: The eleven frozen VLM readers, with their language models, and the discrete-answer readers (a closed API and two annotators). All read the same underlying material (title truncated at 300 characters, on-screen text and transcript at 400 each, eight keyframes, at most six causal same-event excerpts). The VLMs' readings are first-token logit differences; GPT-5.5 and the annotators answer on five levels and have no self-fit on the panel. The self-fit AUC is the diagonal of Tab. D.1; † marks the six readers that cover all 3,082 development videos.

Reader	Family, gen.	Language model	Visual ingestion	Questions	Self-fit
Qwen2-VL 2B	Qwen, 2	Qwen2-1.5B	8 frames	768 (zh)	.886
Qwen2-VL 7B†	Qwen, 2	Qwen2-7B	8 frames	768 (zh)	.899
Qwen2.5-VL 3B†	Qwen, 2.5	Qwen2.5-3B	8 frames	768 (zh)	.892
Qwen2.5-VL 7B†	Qwen, 2.5	Qwen2.5-7B	8 frames	768 (zh)	.885
Qwen3-VL 8B†	Qwen, 3	Qwen3-8B	8 frames	768 (zh)	.895
InternVL2.5 1B	InternVL, 2.5	Qwen2.5-0.5B	8 frames, 448 px	768 (zh)	.887
InternVL2.5 2B	InternVL, 2.5	InternLM2.5-1.8B	8 frames, 448 px	768 (zh)	.871
InternVL2.5 4B	InternVL, 2.5	Qwen2.5-3B	8 frames, 448 px	768 (zh)	.885
InternVL2.5 8B†	InternVL, 2.5	InternLM2.5-7B	8 frames, 448 px	768 (zh)	.894
InternVL3 8B†	InternVL, 3	Qwen2.5-7B	8 frames, 448 px	768 (zh)	.896
Pixtral 12B	Mistral (Pixtral)	Mistral NeMo 12B	2×4 mosaic	768 (en)	.880
GPT-5.5 (closed API)	OpenAI	n/a	8 frames (448 px), declared inputs	30 / 100 (zh)	n/a
Two human annotators	n/a	n/a	full materials, scoring guide	30 (zh)	n/a

Inputs and prompts. The Qwen and InternVL models receive the eight frames as a sequence; Pixtral 12B receives them as one 2×4 mosaic, declared as its ingestion mode, so it differs from the others in input format as well as in family and language. On-screen text is the OCR text distributed with FakeSV and, for FakeTT, text extracted with RapidOCR; transcripts come from Whisper (Radford et al., 2023) (base model, through faster-whisper) for both datasets. Readings are the first-token logit difference between *yes* and *no*; when a tokenizer shares a prefix across the two words, the first distinguishing token is compared. Each question receives only its declared material, the material precedes the question, and an answer instruction closes the prompt. The English template, used by Pixtral, reads "[frames] Title: … Answer based only on the content above: ⟨question⟩ Answer only 'Yes' or 'No'. Answer:", with "Answer based only on the video frames:" when a question reads frames only; the Chinese template has the same structure and asks for *shi* or *fou*. The material itself stays in its original language in both templates. The exact templates are released with the code.

Data and splits. FakeSV (Chinese Douyin) is used with its official temporal split (train 2,536, validation 546, test 542) and its official event-disjoint five-fold protocol, with no change to the lists. FakeTT (English TikTok) has 1,991 videos; under its official temporal split the train and validation side is 63% fake and the test side 33%. Reporting both protocols side by side is one way to state a benchmark's assurance limitations, as Thibault et al. (2025) recommend.

Standardization statistics. Per-question standardization plays the role that calibration transfer plays between spectrometers (Workman, 2018): a per-channel affine map estimated on unlabelled material. All statistics are estimated on each fold's training side and on each reader's own readings, without labels and without borrowing across readers. On the 296-question block, with Qwen2.5-VL-7B's coefficients deployed on InternVL2.5-8B, the one pair tested, 100 unlabelled deployment-side videos sufficed: the deployed AUC comes within .0014 of its value with full-pool statistics, while 50 videos only match raw readings and 25 lose .006. Readouts use $C = .003$ on the 3,082-video pool

and the official temporal test set and $C = .01$ on the 1,200-video panel and in the transfer matrices; under the event-disjoint protocol C is chosen on each fold's validation split. The eleven-reader panel uses all 768 questions.

The cross-dataset threshold. Cross-dataset comparison follows RADAR's accounting: the full target pool, accuracy and macro-F1, and unlabelled target information allowed. We use one scalar, which sets the predicted positive rate of the target pool to the source's label prior, whereas RADAR performs test-time adaptation; pool sizes differ by under 2%. When the two datasets have different class priors this rule is a heuristic, which is why cross-dataset conclusions are stated in AUC. Expectation-maximization (EM) prior correction (Saerens et al., 2002), the label-shift counterpart of importance weighting under covariate shift (Shimodaira, 2000), fails in both directions (204-question transfer arms on the official test splits; estimated priors of .893 and .000 against true .331 and .561), because the per-question offset is a covariate shift rather than a pure label shift; a synthetic check confirms the implementation.

The corroboration channel. Corroboration questions read the title and on-screen text of at most six same-event videos published earlier than the query video, the causal pool, and never labels. Same-event membership is the event field that FakeSV and FakeTT annotate for every video; a deployed system would need its own grouping, by retrieval or clustering, which we did not evaluate. On FakeTT the channel reads all same-event videos of the pool, earlier or later, because no causal pool was built for it; without the 163 corroboration questions, FakeSV-fitted coefficients reach AUC .848 and 75.0 accuracy on the full FakeTT pool, against .877 and 77.7 with them. On FakeSV four pools were compared: all same-event videos including later ones (not deployable), the causal pool, the training set only (TrueLens's accounting, which removes the evidence channel for 34% of test videos) and none (SIEVE's single-video setting). Under the 768-question protocol with no selection, the causal pool gives 89.3 accuracy and the single-video setting 87.1; on the 680-question base bank the all-pool arm reached 90.6, but it uses later videos and cannot be deployed. NEED uses same-event neighbours and debunking videos, together with debunking annotations; TrueLens retrieves labelled precedents from the training and validation sets. Under the temporal split, 56.6% of a test video's causal neighbours belong to the official training and validation sets, a flow of content across splits that we treat as transductive; neighbour labels are homogeneous within events at 89 to 96%.

The reproduced baseline. FakeSV-VLM was reproduced from its released configuration, trained on the temporal split and evaluated on the full 542 test videos: 90.04 accuracy and 89.8 macro-F1, against a published three-run mean of 90.22. Under the event-disjoint protocol the same recipe was retrained per fold, with the best checkpoint chosen on each fold's validation split, giving 80.0 ± 1.5, with plain and balanced accuracy in agreement.

D READERS: COMPLETE RESULTS

This appendix gives the complete numbers behind the paragraphs on transfer and recognition in Sec. 4: one readout on every reader, the full transfer matrices, the early pre-registered calibration and the cross-generation intervals, recognition in detail, and the prompting-language axis.

One readout on every reader. Figure D.1 deploys the coefficients fitted once on Qwen2.5-VL-7B on every reader and arm. For every VLM of at least 3B parameters except Pixtral the deployed value stays within .025 of the reader's own fit; Pixtral (.043 below), the 1B and 2B models and the discrete-answer arms at 30 questions fall further below it, while GPT-5.5 at 100 questions exceeds its reference.

The full transfer matrices. Table D.1 gives the standardized transfer matrix and Fig. D.2 the same matrix on raw readings. Every cell that involves neither a 1B nor a 2B model lies within .07 of its column's self-fit.

Pre-registered calibration and cross-generation intervals. On the original 3×3 matrix the pre-registered criteria were a maximum pairwise asymmetry of at most .015, a mean transfer loss of at most .020 and a reference-row degradation of at most .005 after calibration. The point estimates were .011, .017 and .000, a pass; the bootstrap upper bounds of .028 and .022 lie in the partial band of the pre-registered criteria. For Qwen2.5-VL-7B and Qwen3-VL-8B on the 296-question block of the 1,200-video panel, self-fits are .891 and .899 and cross-fits .889 and .880, so each direction loses about .010 AUC.

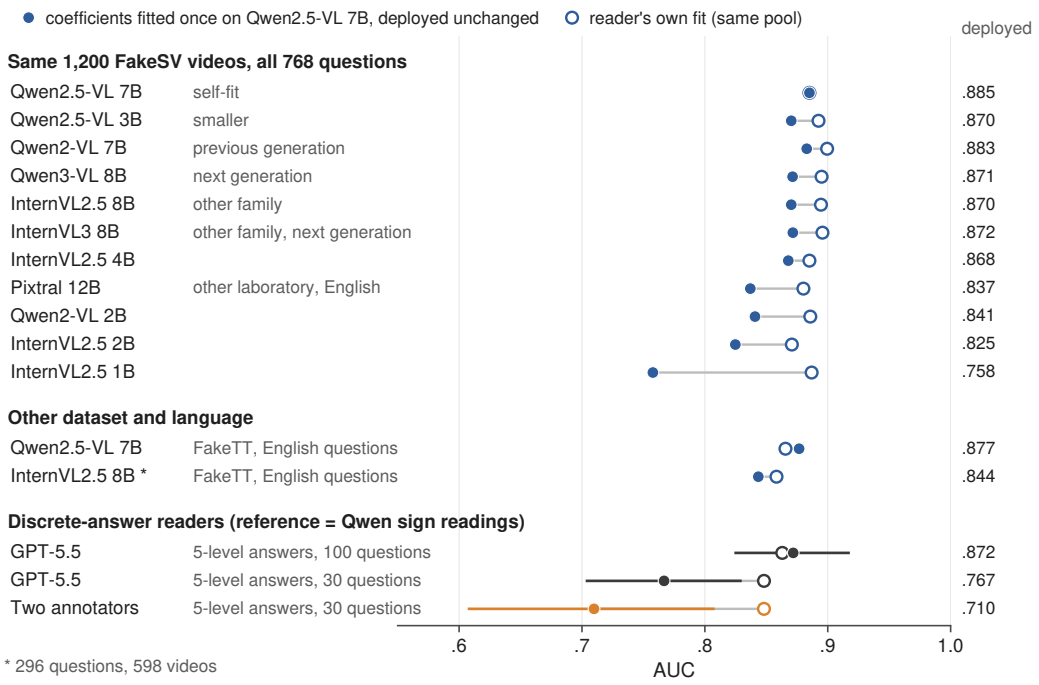


Figure D.1: One readout, every reader. Coefficients fitted once on Qwen2.5-VL-7B with Chinese questions, deployed unchanged (filled) on every other reader and arm; for the discrete-answer arms they are fitted on Qwen's one-bit signs for the same 30 or 100 questions. Hollow markers are each reader's own fit on the same pool and protocol and, for the discrete-answer arms, Qwen's own signs read by those coefficients. Top: eleven frozen VLMs, 1,200 videos, 768 questions. Middle: another dataset and language (English FakeTT). Bottom: GPT-5.5 (200 videos) and two annotators (100 videos), with 95% CIs over videos.

Table D.1: Fit on the row reader, deploy on the column reader: out-of-fold AUC after per-question standardization (1,200-video panel, 768 questions, event-grouped five-fold CV). The diagonal is the self-fit.

	Qwen-VL					InternVL					Pixtral
Fitted on	2 2B	2 7B	2.5 3B	2.5 7B	3 8B	2.5 1B	2.5 2B	2.5 4B	2.5 8B	3 8B	12B
Qwen2-VL 2B	**.886**	.874	.853	.862	.858	.767	.805	.865	.885	.868	.798
Qwen2-VL 7B	.825	**.899**	.880	.877	.882	.765	.799	.870	.885	.883	.837
Qwen2.5-VL 3B	.833	.883	**.892**	.868	.879	.788	.827	.870	.875	.859	.840
Qwen2.5-VL 7B	.841	.883	.870	**.885**	.871	.758	.825	.868	.870	.872	.837
Qwen3-VL 8B	.831	.886	.875	.873	**.895**	.736	.800	.868	.884	.882	.824
InternVL2.5 1B	.826	.844	.868	.854	.832	**.887**	.829	.856	.870	.863	.828
InternVL2.5 2B	.833	.847	.859	.855	.821	.812	**.871**	.866	.868	.870	.799
InternVL2.5 4B	.829	.882	.863	.867	.863	.829	.810	**.885**	.885	.876	.838
InternVL2.5 8B	.828	.879	.849	.866	.864	.755	.786	.866	**.894**	.876	.813
InternVL3 8B	.830	.884	.857	.868	.876	.820	.822	.876	.892	**.896**	.830
Pixtral 12B	.816	.881	.873	.872	.878	.714	.811	.871	.876	.874	**.880**

Recognition in detail. With all 768 questions, the eight readers of at least 3B parameters recognize each other at a mean of .974 (median .981, minimum .918); all 110 ordered pairs average .793 (median .926); the 54 pairs involving 1B or 2B models average .605 (minimum .290). With the 353 frame-input questions only, the mean over all 110 pairs is .577; with the 129 frame-input questions of the 296-question block, Qwen recognizes InternVL at .677, about 800 times chance. On unified

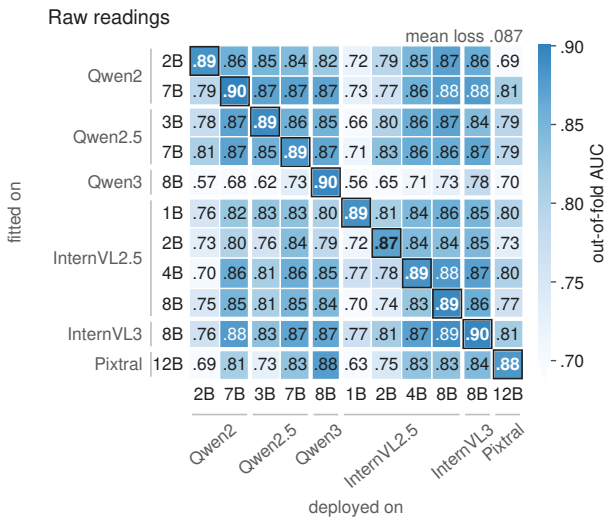


Figure D.2: The eleven-reader transfer matrix on raw readings, in the layout and colour scale of Fig. 2(a); values below .70 take the lightest colour.

pools of 300 videos the four axes order as language 1.000, generation .984, family .937 and size .927, and a double control at 38 questions keeps the language axis at 1.000 where the other axes reach .68 to .75. Shallow metadata explains 5.5 to 5.8% of sheet variance out of fold, and the residual after removing it still recognizes at .909.

Prompting language. With 300 videos and the model held fixed, switching 38 questions from Chinese to English leaves per-question rankings nearly unchanged (median Spearman .881), costs a Chinese-fitted readout .022 AUC on English sheets (95% CI [.003, .043]), and leaves the Chinese and English sheets of the same video recognizing each other (300 of 300 and 299 of 300).

Per-question agreement. Question by question, cross-reader readings correlate at .645 in the median (Spearman), while paraphrases of a question on one reader correlate at .97. Under Eq. (3) the first estimates $\lambda_{Ak}\lambda_{Bk}$ and the second also counts each reader's own deviation, which is stable under paraphrase, so about a third of a single reading's variance is specific to the reader.

E Discrete-Answer Readers

This appendix gives the details behind the second place where reading thins in Sec. 4. It first isolates the answer channel on one model, then describes GPT-5.5 and the two annotators. The criteria were frozen before any answer was collected. For GPT-5.5 at 30 questions, a pass required a deployed AUC of at least .75 and of at least Qwen's sign reference minus .03, which fixed the threshold at .816; the arm fails, at .767 [.703, .830]. The 100-question arm was registered as a measurement without a gate of its own and clears .816 (.872 [.824, .918]). For the annotators, a pass required an AUC of at least .72 with a lower confidence bound above .60, after a $\kappa \geq .6$ gate; they are partial (.710 [.607, .808]).

Isolating the channel on one model. On the 200-video pool with 30 questions and event-grouped five-fold cross-validation, binning Qwen's continuous readings into five levels by per-question magnitude lowers its self-fit only from .913 to .902, while re-binning the same readings per question to the answer distribution GPT-5.5 actually produces lowers it to .766; GPT-5.5's own self-fit is .737. At 100 questions the four values are .894, .891, .815 and .854: at the larger budget GPT-5.5's answers carry more than the collapse of their distribution alone would leave. At 30 questions, what discards information is the concentration of answers on one level; the number of levels matters little.

GPT-5.5. GPT-5.5 (snapshot gpt-5.5-2026-04-23) answered through the Batch API on a fixed five-level scale and received the same declared inputs as the open models, including eight 448-pixel keyframes. The model does not accept a non-default temperature, so default decoding with low

reasoning effort was used. All 6,000 cells at 30 questions and all 20,000 at 100 questions returned and parsed; the log records seven cells where a small pilot run and the full run disagreed because of non-determinism, and one question on which every answer was an abstention. At 30 questions, 11 questions have a modal answer above 92% and are near-constant (Tab. E.1). Questions that GPT-5.5 itself drafted and questions DeepSeek drafted do not differ at 100 questions on any of three tests after Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) (sign agreement with Qwen, modal share, single-question discriminability; all $q > .36$), nor at 30 questions (Tab. E.2).

Table E.1: GPT-5.5 at 30 questions, per question: drafter family, modal answer share, collapse (modal share above .92), sign agreement with Qwen and single-question discriminability. IDs are drafting identifiers of the released bank, not positions in the frozen order.

ID	Drafter	Modal	Coll.	Sign	$\lvert\text{AUC}-.5\rvert$	ID	Drafter	Modal	Coll.	Sign	$\lvert\text{AUC}-.5\rvert$
c0659	DeepSeek	.99	yes	.92	.005	c0115	GPT	.81		.87	.158
c0679	GPT	.99	yes	1.00	.010	c0232	GPT	.77		.83	.022
c0074	GPT	.98	yes	.95	.021	c0566	DeepSeek	.73		.09	.091
c0446	DeepSeek	.97	yes	.97	.004	c0640	DeepSeek	.73		.92	.067
c0342	DeepSeek	.95	yes	.93	.047	c0482	GPT	.70		.86	.014
c0658	DeepSeek	.95	yes	.99	.034	c0633	DeepSeek	.70		.79	.031
c0366	GPT	.95	yes	.96	.003	c0299	GPT	.67		.85	.065
c0392	GPT	.95	yes	.79	.047	c0507	DeepSeek	.65		.57	.124
c0473	DeepSeek	.95	yes	.70	.007	c0357	GPT	.63		.82	.021
c0322	GPT	.95	yes	.90	.052	c0570	DeepSeek	.62		.63	.044
c0632	DeepSeek	.94	yes	.92	.022	c0626	DeepSeek	.53		.81	.076
c0676	GPT	.91		.93	.069	c0502	DeepSeek	.51		.64	.092
c0316	DeepSeek	.90		.94	.068	c0258	GPT	.49		.71	.009
c0245	DeepSeek	.89		.92	.065	c0518	GPT	.47		.83	.030
c0200	GPT	.85		.90	.059	c0627	DeepSeek	.39		.80	.125

Table E.2: GPT-5.5 by drafter family, 30 questions.

Metric (per question)	DeepSeek-drafted ($n = 16$)	GPT-drafted ($n = 14$)	Mann–Whitney p
sign agreement with Qwen (median)	.862	.865	.47
collapsed (modal share $> .92$)	6/16	5/14	.80
$\lvert\text{AUC} - .5\rvert$ (median)	.056	.026	.23

The 30 consensus questions. The discrete-answer arms use 30 questions selected among the first 204 of the frozen order by the average rank of coefficient magnitude across three VLM readers, fitted with labels on the 1,200-video development panel. That panel contains 85 of the 200 videos of the GPT-5.5 pool and 44 of the 100 annotated videos, so the selection saw the labels of part of the evaluation pools, and these arms are not selection-free. The 70 further questions of the 100-question arm were drawn at random, stratified by generator, family and input type, with the draw frozen before any request. Re-selecting from all 768 questions with selection and evaluation pools isolated changes the AUC of Qwen's own signs by $+.022$, with a confidence interval containing zero. The intervals of both arms resample videos, as registered; the 200-video pool spans 42 events, and resampling events instead widens GPT-5.5's intervals to [.667, .857] at 30 questions and [.782, .942] at 100.

The two annotators. Neither annotator is an author; annotator A knew the hypothesis and annotator B did not. Each answered the same 30 consensus questions on 70 videos (40 overlapping) on a five-level scale, in an interface that highlights the inputs each question declares, blind to the labels. Agreement is reported as Cohen's κ (Cohen, 1960), with the .6 gate following the conventional bands (Landis & Koch, 1977), and as Krippendorff's α (Krippendorff, 2011). For empty material the annotators followed three rules: existence questions are answered no, attribute questions cannot be judged and relation questions cannot be judged if either side is empty; model readers see the bare question. Table E.3 gives the three collection rounds. Cohen's κ rose from .280 to .624 while the deployed AUC moved only from .695 to .710, so harmonization changed agreement and not the signal.

Table E.3: The human arm over three collection rounds. Agreement (κ, α) is computed on the 1,200 overlapping cells (40 videos $\times$ 30 questions), the deployed AUC on the union of 100 annotated videos. Rounds two and three were triggered by the κ gate. The third re-annotated only the 101 cells of four questions on which the annotators disagreed, so its κ is an upper bound; on the 26 untouched questions κ was already .607 after round two. Against the pre-registered criterion (AUC $\geq$ .72 with CI lower bound $>$ .60) every round is partial; A / B: AUC of each annotator alone; bold: the round reported in the main text.

Round	Trigger	κ 5-level / weighted / sign	α	Deployed AUC [95% CI]	A / B
1, initial	none	.280 / .335 / .309	.362	.695 [.584, .796]	.623 / .718
2, A re-annotates all	gate not met; rules for empty material	.506 / .504 / .577	.479	.708 [.603, .805]	.638 / .718
3, four questions harmonized	round-2 gate not met; 4 disputed questions re-annotated	**.624 / .613 / .703**	**.572**	**.710 [.607, .808]**	.640 / .723

In round three, 41 of the 44 cells annotator A changed went from "definitely no" to "cannot judge" (relation questions with empty material), and 35 of the 42 cells B changed went from "cannot judge" to "definitely no" (existence questions with absent material). The disagreement was a rule-level disagreement about empty material, consistent with A's test-retest agreement of .99 on 26 stable questions. Coefficients fitted on the human sheets agree with Qwen's coefficients in sign on 70% of questions (rank correlation .51). The median pace was 3.1 s (A) and 4.0 s (B) per cell, with no straight-lining.

F Question Text and Meaning

This appendix supports the paragraph on correspondence in Sec. 5: three controls, each removing one alternative explanation, the paraphrase and negation tests, what the readout reads, and two decisions decomposed question by question.

Three controls. Figure F.1 collects three controls. The first concerns input modality: equal numbers of text-input and frame-input questions, 19 each, transfer across prompting languages with the losses of Tab. F.1. The second concerns dimensional coincidence: keeping the dimension count and every marginal distribution and permuting only the correspondence between questions and readout dimensions, 20 times, drops cross-dataset AUC from .877 to a mean of .539 (SD .073) and cross-reader AUC on the 296-question block from .891 to .488 (SD .112). The third concerns per-sample question generation: a SAFE-style reimplementation that generates questions per video and pools the answers, on the same backbone and inputs, reaches .880 on the FakeSV test set but only .737 across datasets, while a readout on the first 204 questions of the frozen order, the budget an earlier protocol chose by development cross-validation, reaches .850 on the same videos, a paired difference of $+.113$ (95% CI [.070, .156]).

Table F.1: Equal-count prompting-language control (19 text-input against 19 frame-input questions, same videos).

Question group (19 each)	Transfer loss, zh$\rightarrow$en prompting	Recognition (top-1)
text-input questions	$+.042$	.863
frame-input questions	$-.019$ (deployed side higher)	.893

Paraphrase and negation. On Qwen2.5-VL-7B, 100 questions of the bank were rescored on 150 videos in their original wording, two paraphrases and one negation. Paraphrase invariance is near-perfect: the median per-question correlation is .97, all correlations are positive and 99% exceed .5. Negation flips most questions (median $-.49$, 68% negative), ordered by family: checkability $-.72$, exploration slot $-.71$, forgery traces $-.64$, tone and affect $-.58$, cross-modal consistency $-.28$ and corroboration $+.10$, still $+.06$ when corroboration questions receive their full declared materials.

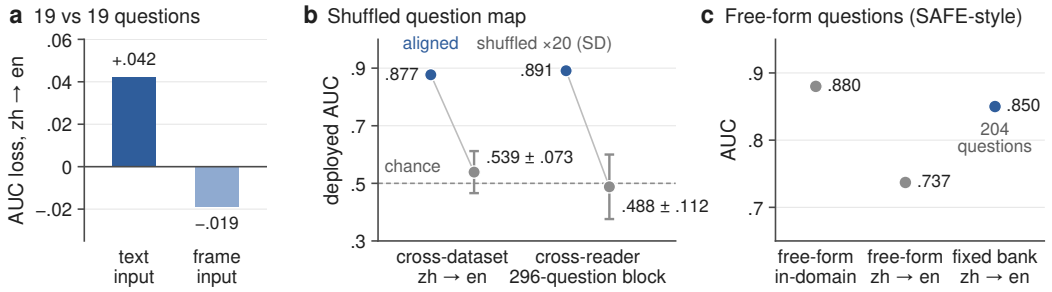


Figure F.1: Three controls. (a) Transfer loss when the prompting language changes from Chinese to English, for 19 text-input and 19 frame-input questions; (b) permuting the question-to-dimension correspondence, 20 draws (bars: SD across draws), for cross-dataset and cross-reader deployment; (c) free-form questions generated per video (SAFE-style reimplementation) against a readout on the first 204 questions of the bank.

This is why corroboration coordinates are read as the availability of corroborating material. Per-question results and the 55 questions that cannot be negated are released.

What the readout reads. By input source, counting a question toward every source it declares, on-screen text has the strongest single questions (median $|\mathrm{AUC} - .5|$ of .118), the title and description come next (.110), then same-event videos (.069), frames (.057) and the speech transcript (.045, with 26% of questions within two standard errors of chance). Removing any single source costs at most .010 AUC, whereas removing frames and on-screen text together (463 questions) costs .026: single sources can replace each other, the visual channel as a whole cannot. By design family, checkability is strongest (.098) and forgery traces weakest (.042, 22% near chance). Figure F.2 draws these distributions. The exploration slot holds 13 of the 20 strongest single questions, including the strongest of all ("is this mainly a music video, dance or lip-sync rather than a news record", AUC .803 toward fake) and the two largest readout coefficients (Tab. F.2). Several of them transfer only partly to FakeTT (the two largest coefficients keep .654 and .714 there); they expose self-declared non-news content inside the fake class. On the FakeSV test set, removing the whole slot of 128 questions raises accuracy by 0.7 points and lowers AUC by .005 (on the development panel its removal raises AUC by .005, Tab. A.2); removing the six questions that declare staging leaves accuracy unchanged and removing a broader eleven-question staging cluster raises it by 0.2 points, so the main result does not rest on them. Two directions replicate across datasets and one reverses. "Are the other sources mostly full copies of this one" points to fake (.633 on FakeSV, .690 on FakeTT); "speech describes concrete outcomes the frames do not show" points to real (.392 and .397, since ordinary broadcasts narrate outcomes over file footage); and "earlier same-event videos contain raw phone footage" points to real on Douyin (.394) and to fake on TikTok (.535), so production-style cues are platform-specific. On a benchmark where topic and label are associated, these numbers describe the association between questions and labels, topic structure included (Sec. 5).

Two decisions, question by question. Figure F.3 takes two test videos, one real and one fake, and shows eight keyframes and the eight largest contributions of the 768-question readout, each the coefficient times the standardized reading, positive toward fake. The decomposition is exact, not a post-hoc attribution (Rudin, 2019). The two videos are the most confident correct real and fake test videos whose eight largest contributions span at least three design families; the fake one is a satirical clip with recognizable people whose evidence is textual, and its frames are pixelated.

G The Negative-Control Bank

This appendix gives the design, the pre-registered criteria and the results of the negative-control bank used in the paragraph on compatibility and validity in Sec. 5.

Design. The bank separates the contribution of fixing the coordinates from that of their content. It has six incidental-attribute families: colour and appearance; objects, living things and food; spatial layout; incidental content of the transcript; incidental vocabulary of the title and on-screen text;

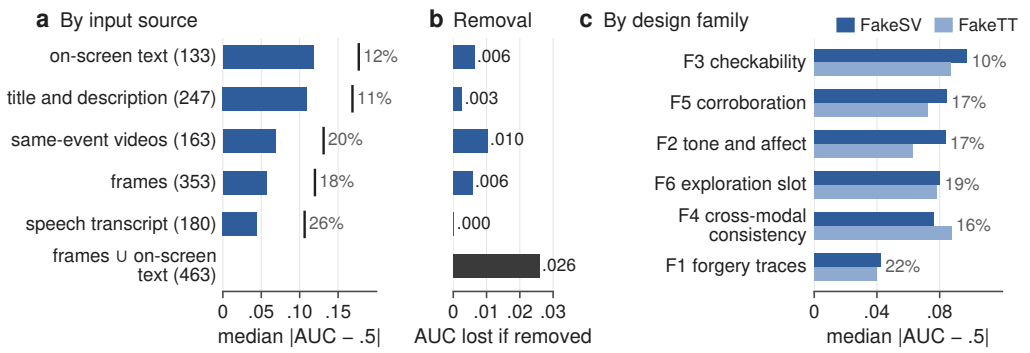


Figure F.2: What the readout reads on FakeSV (3,082 development videos). (a) Single-question discriminability by input source, counting a question toward every source it declares (bars: median; ticks: 75th percentile); (b) AUC lost when a source is removed, the last row removing frames and on-screen text together; (c) discriminability by design family on FakeSV and FakeTT. Percentages in (a) and (c): share of questions with $|\mathrm{AUC} - .5| < .02$, about two standard errors.

incidental content of earlier same-event videos. Each is crossed with four surface-form styles and drafted by the same two generators from the same prompt skeleton, with the task description removed and evaluative, source, production, style, specificity, consistency, corroboration, engagement and high-risk-topic vocabulary banned. The 752 candidates became 708 after rule screening, 673 after an independent LLM judge applied nine rejection rules and 669 after deduplication; 256 were then drawn stratified by the real bank's input-pattern distribution and frozen by SHA-256, with the criteria registered before scoring. Stage one scored the 138 text-only questions on Qwen2.5-VL-7B and InternVL2.5-8B, on the FakeSV panel (1,200 videos) and on FakeTT (598). Stage two scored the remaining 118 questions, which read frames or same-event excerpts (102 frame-only or frame-plus-text, 16 with the causal same-event pool), on Qwen2.5-VL-7B, on the same videos and with the reader configuration of the real bank. References are 200 random subsets of the real bank matched in size and input pattern and evaluated under the same protocol; for stage two the quotas are 88 frame-only, 16 causal-pool, 8 title-plus-frame and 6 on-screen-text-plus-frame questions.

Criteria. The recovery ratio is $\mathrm{rec} = (\mathrm{AUC}_{\mathrm{ctl}} - .5)/(\mathrm{AUC}_{\mathrm{ref}} - .5)$. The frozen bands are: in-domain, pass at $\mathrm{rec} \leq .50$ and fail at $\mathrm{rec} \geq .80$; cross-dataset, pass at $\mathrm{rec} \leq .40$ with control minus shuffled below .05, fail at $\mathrm{rec} \geq .70$; per-question median $|\mathrm{AUC} - .5|$, pass at $\leq .035$ and fail at $\geq .060$.

Results. Table G.1 gives both stages against the frozen criteria. The text-only control questions recover about four fifths of the real bank's in-domain discrimination and about a third of its cross-dataset discrimination. The stage-two control questions show the same ordering with a smaller gap, since scene attributes covary with how news-like a video looks on both platforms. On FakeTT the real-bank reference subsets read their same-event questions from all same-event videos, while the control bank read the causal pool; restricted to questions without same-event input on both sides, the stage-one cross-dataset comparison is .588 against .776 (rec .32), with the same verdict. As written before unblinding, the in-domain result means that question content cannot be invoked to explain in-domain performance and that single-question readings are not evidence of veracity, and Sec. 5 and App. F are written under this rule; the cross-dataset result is reported as incremental content validity, about .21 AUC over a form-matched unrelated bank in stage one and .16 in stage two.

Reference floors. Under the same protocol, a character n-gram classifier on the title, on-screen text and transcript reaches .839 in-domain, and the number of causal neighbours alone reaches .416, a $|\mathrm{AUC} - .5|$ of about .08. The control bank's mean yes-rate is .029 against .177 for the real bank; a yes-rate-matched reference leaves the in-domain verdict unchanged (rec .82), and the cross-dataset endpoint was not re-run with it. In stage two the yes-rates match (.196 and .197).

Table F.2: The twelve largest coefficients of the main readout of Tab. 1 (fitted on the 3,082 development videos), in order of magnitude. "Points to" is the direction of the coefficient, a property of the readout; the single-question AUCs on the FakeSV and FakeTT development pools are taken in that direction, so a value below .5 means the question alone points the other way. ‡ marks suppressor variables (large coefficient, near-chance single-question AUC). IDs are drafting identifiers of the released bank, not positions in the frozen order.

				Single-question AUC	
ID	Question (English wording)	Inputs	Points to	FakeSV dev	FakeTT dev
c0712	Are animals or wildlife the main visual subject of the video?	frames	fake	.664	.654
c0658	Does the video contain a statement that it is a dramatization, reenactment, or staged recreation?	frames, on-screen text, transcript	fake	.763	.714
c0366‡	Does the title or description contain a formal reference number, such as a case number, notice number, document number, flight number, train number, or order number?	title	real	.472	.566
c0704‡	Does the video use a noticeably tilted camera angle, often called a Dutch angle, where the horizon is slanted?	frames	real	.536	.427
c0429	Does the title describe a weather, disaster, or environmental condition such as flooding, fire, smoke, snow, or storm that the footage does not show?	title, frames	real	.572	.439
c0263	Does the title frame the claim as a specific incident or case rather than drawing a conclusion about an entire group or trend?	title	real	.719	.619
c0299	Do the video frames show a distinctive landmark, building exterior, or terrain feature that could help identify the location, rather than only a generic street or indoor setting?	frames	real	.621	.566
c0574	Does this video's title or on-screen text state a specific time or date, while another video about the event shows a visible clock or timestamp indicating a different time?	title, on-screen text, same-event excerpts	real	.731	.634
c0551	Do other videos show an event-related on-site announcement, notice, or staff explanation from an institution such as a government department, police, fire service, hospital, school, or company?	same-event excerpts	real	.737	.697
c0288	Among other videos about the same event, do any include direct references to official documents, databases, or primary sources?	same-event excerpts	real	.710	.680
c0640	Do any of the other videos show emergency response vehicles (ambulances, fire trucks, police cars) at the event scene?	same-event excerpts	real	.741	.643
c0679	Does the material show or describe supernatural or paranormal phenomena, such as ghosts, curses, miracles, or psychic abilities?	frames, title, on-screen text	fake	.712	.802

H Accuracy, Cross-Dataset Transfer and Cost

This appendix supports the paragraph on accuracy in Sec. 5: the gap between readouts on the same backbone, whether the hidden state holds information beyond the sheet, the probe and the readout under development cross-validation, the audit of the corroboration channel, cross-dataset details, label efficiency and selective prediction, and cost.

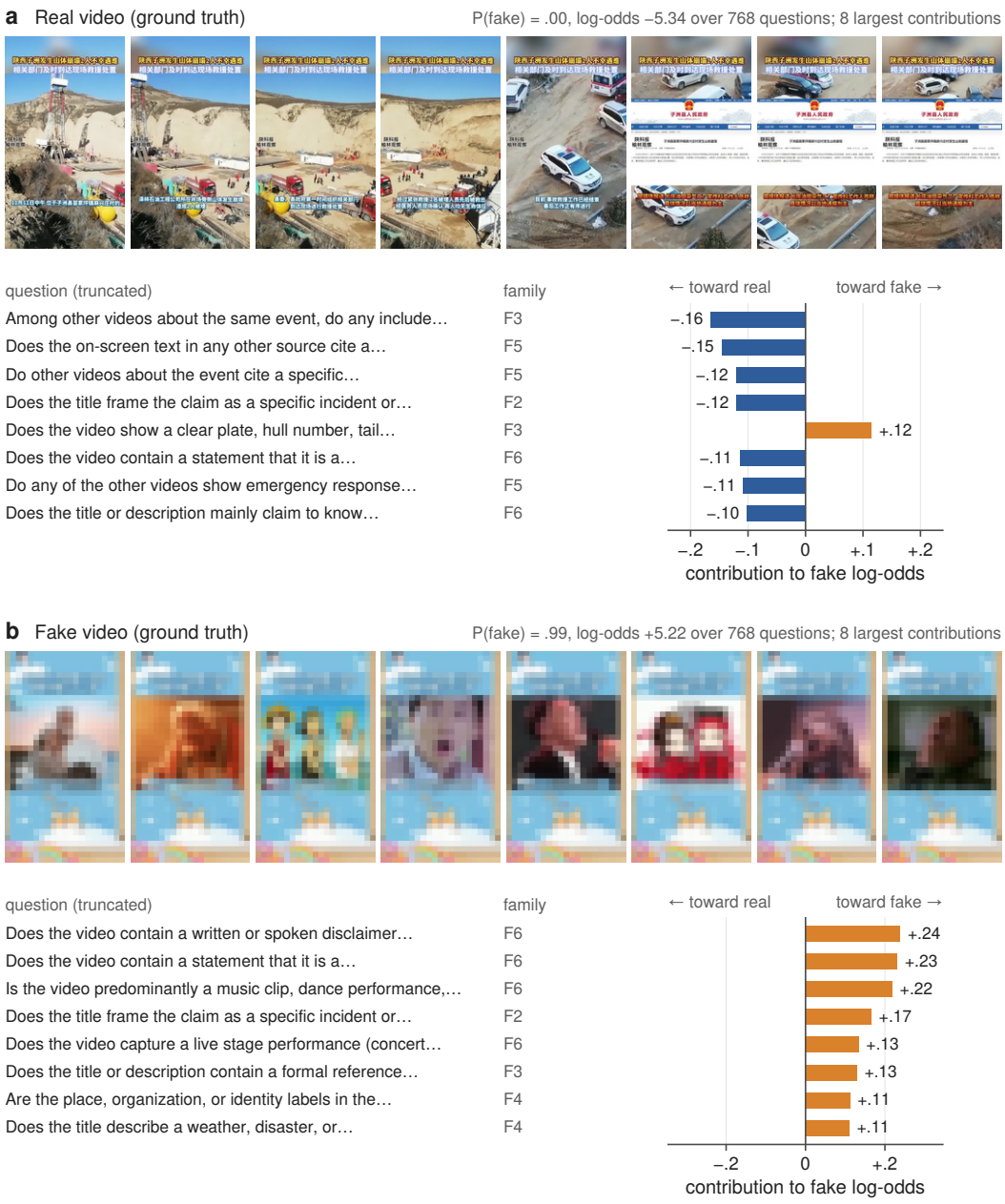


Figure F.3: Two test videos decomposed question by question: eight keyframes and the eight largest contributions to the fake log-odds (orange toward fake, blue toward real; family codes as in Tab. A.1). The fake video's frames are pixelated for privacy.

Readouts on the same backbone. Table H.1 compares five readouts of one frozen backbone. Asking the model directly reaches 63.5 accuracy and chain-of-thought prompting reaches less; the hidden-state probe reaches 86.4 and the question readout 89.3.

Information beyond the sheet. The hidden state is the mean over visual tokens at layer 20 of Qwen2.5-VL-7B. A ridge regression (penalty 10, fixed) on training folds maps the sheet z to the hidden state h, and all probes are L2-regularized logistic regressions with $C = .003$ on standardized features under five event-grouped folds of the 3,082 development videos. The residual $e = h - \hat{h}(z)$ predicts the label out of fold at .573 alone and at .890 jointly with the sheet, against .891 for the sheet alone. A linear model on (z, e) has the same expressive power as one on (z, h), so the test asks

Table G.1: The negative-control bank against the frozen criteria. Stage one: 138 text-only questions, two readers. Stage two: 118 questions reading frames or same-event excerpts, Qwen2.5-VL-7B. References: 200 real-bank subsets matched in size and input pattern.

Endpoint	Control	Shuffled control	Reference mean ± SD	rec [95% CI]	Verdict
Stage one: 138 text-only control questions, both readers					
In-domain AUC, Qwen (panel CV)	.802		.876 ± .004	.80 [.76, .85]	fail
In-domain AUC, InternVL	.820		.887 ± .003	.83 [.78, .87]	fail
Cross-dataset AUC, FakeSV→FakeTT, Qwen	.599	.484 ± .055	.807 ± .015	.32 [.18, .46]	partial
Per-question median |AUC − .5|	.073 [.059, .085]		.087 ± .007	.84	fail
Recognition, Qwen→InternVL / back	.863 / .786		.971 / .964	n/a	as predicted
Raw transfer loss, Qwen→InternVL	+.033		+.017	n/a	as predicted
Stage two: 118 control questions reading frames or same-event excerpts, Qwen only					
In-domain AUC, Qwen (panel CV)	.735		.871 ± .004	.63 [.57, .70]	partial
Cross-dataset AUC, FakeSV→FakeTT, Qwen	.674	.484 ± .057	.834 ± .013	.52 [.39, .65]	partial
Per-question median |AUC − .5|	.056		.063 ± .006	.89	partial

Table H.1: Readouts of the same frozen backbone, FakeSV official test set (542 videos).

Readout	Accuracy	AUC
Direct answer, text fields and frames	63.5	n/a
Direct answer, frames only	67.9	n/a
Chain-of-thought answer (greedy / self-consistency 5)	56.6 / 58.5	n/a
Hidden-state probe	86.4	.920
Question readout, 768 questions	89.3	.944

whether regularized linear probes gain from the hidden state. On the FakeSV test set the residual scores .513, with no difference between seen and unseen events (.501 and .526). The design is in the spirit of amnesic probing (Elazar et al., 2021), which removes a property from a representation and measures what is lost; here the sheet is regressed out of the hidden state.

Development cross-validation. Under the same five event-grouped folds on the 3,082 development videos, the hidden-state probe reaches 83.3 accuracy and .903 AUC, and the readout with corroboration 81.7 and .891.

Moving a probe. A probe can move between open models through relative representations (Moschella et al., 2023). Fitted on Qwen2.5-VL-7B's cosine similarities to 300 unlabelled development videos used as anchors, a hidden-state probe reaches .865 on Qwen and .869 when deployed on InternVL2.5-8B's similarities to the same anchors (1,200-video panel, event-grouped CV; shuffled anchors .60), against .870 for the question readout on this pair; fitted on the development set, it reaches .915 on the test set on Qwen and .908 across the two models. Across datasets, the same-backbone probe fitted on FakeSV reaches AUC .774 and 69.0 accuracy on the full FakeTT pool, against .877 and 77.7 for the question readout. The probe needs the hidden states of both models, so it cannot read a closed interface or a person.

The corroboration audit. Under the event-disjoint folds, the accuracy gain from corroboration grows with the number of causal neighbours: -1.0 points (not significant) with none, $+1.5$ with one or two and $+1.9$ ($p = .0009$) with three or more, while the all-pool arm, which includes later videos, still gains $+3.3$ on videos with no causal neighbour, the signature of leakage from future corroboration and the reason the causal pool is the main accounting. On the temporal test set, the readout with the channel removed entirely reaches 87.1 accuracy and .922 AUC, 2.2 points below the causal pool.

Cross-dataset details. FakeTT readings use the English wording of every question. Chinese-fitted coefficients reach .877 on the full English pool, above the English self-fit under development CV (.866), and .896 on the official FakeTT test split (299 videos), with 83.3 accuracy under the pool-level threshold and 76.3 under a threshold set inside the split. A FakeSV-VLM checkpoint trained under the event-disjoint protocol, applied unchanged to FakeTT, reaches 78.2 accuracy, 75.1 macro-F1 and a predicted-fake rate of .764 on the full pool (1,989 decodable videos), level with the readout in accuracy (77.7) and behind it in macro-F1 (77.5), and 69.6 accuracy with a predicted-fake rate of .595 on the official test split. RADAR's figures are quoted from its paper, whose pools differ from ours by 2 and 60 videos. Because FakeTT's official split is 63% fake on the train and validation side and 33% on the test side, every accuracy reported on it mixes in the choice of threshold, which is why AUC is our primary metric.

Label efficiency and selective prediction. Figure H.1 shows accuracy as a function of the number of labelled videos for the question readout and the same-backbone probe, against the zero-shot answer and the fine-tuned detector (30 draws per point), and selective prediction on both test sets (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017).

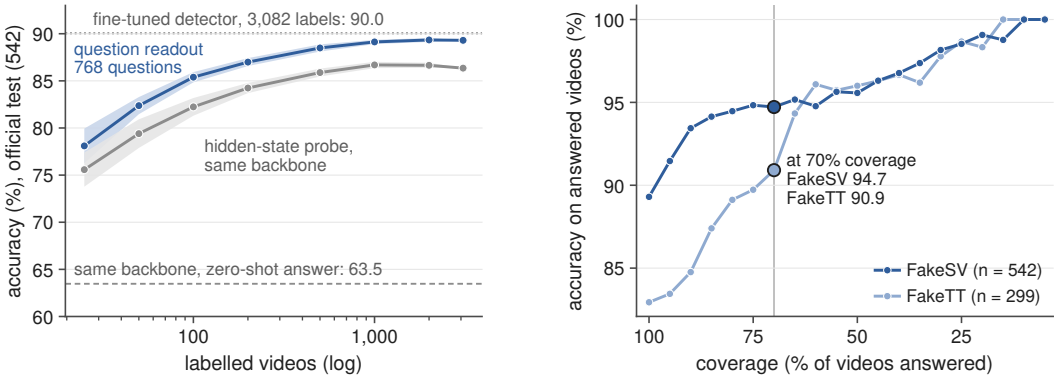


Figure H.1: Left: label efficiency on the FakeSV official test set (30 draws per number of labels; bands: 95% intervals). Right: accuracy on the videos answered as coverage decreases, on both test sets.

Cost. Table H.2 gives the measured cost per video. A text-only question is one forward pass of 17.6 ms and a frame-input question takes 424 ms, so the current implementation spends 157 s per video; by our estimate, a shared visual prefix would bring it to about 14 s. InternVL's 448-pixel inputs make frame-input questions about three times cheaper. The aggregation law prices the budget: with $K_{1/2} = 18$, Eqs. (5) and (6) give 96.7, 98.6, 99.2 and 99.6% of the full-bank AUC at 96, 204, 296 and 430 questions.

Table H.2: Inference cost per video (H200, Qwen2.5-VL-7B, measured per forward pass). The instrument is built once and reused unchanged across models and languages.

Configuration	Cost per video	Note
One text-only question	17.6 ms	one forward pass, one logit
One frame-input question	424 ms	visual encoding not shared
768 questions, current implementation	157 s	353 frame-input and 415 text-only questions
768 questions, shared visual prefix (bound)	about 14 s	one visual encoding plus 768 short answers
first 204 questions (budget variant)	about 39 s	87 frame-input and 117 text-only questions; AUC .936, accuracy 90.2
Fine-tuned detector, inference	one forward pass with generation	retrained at every change of model, language or dataset

I THE SECOND CONSTRUCT: HARMFUL MEMES

This appendix gives the complete results of the second construct in Sec. 5. The recipe is the same as for video: 220 bilingual questions drafted by the same two generators and frozen before scoring, a frozen backbone and a trained linear readout, evaluated on the official test sets of ToxiCN-MM (Lu et al., 2024) and HarMeme (Pramanick et al., 2021). Table I.1 places the readout next to the published zero-training and purpose-trained systems, with comparison figures taken from the cited papers.

Table I.1: Harmful-meme detection with the same recipe: 220 questions, Qwen2.5-VL 7B and 3B, official test sets. Comparison figures are taken from the cited papers.

	ToxiCN-MM (zh, test 2,400)	HarMeme Harm-C (en, test 354)
Question readout, ours (frozen backbone, linear readout)	AUC .848, acc 78.3, macro-F1 75.7	AUC .901, acc 84.2
Same backbone, direct answer	macro-F1 9.4 points below ours	acc 68.1 (harm recall collapses)
Best published zero-training	GPT-4 zero-shot, macro-F1 70.1 (Lu et al., 2024)	U-CoT+ acc 83.9 (Pan et al., 2026); GPT-4o acc 72.9 (Mei et al., 2025)
Purpose-trained systems	macro-F1: CLIP+MKE 80.2, RoBERTa 76.4 (Lu et al., 2024); FG-E2HMD 83.4 (Chen et al., 2025)	PromptHate acc 84.47 ± 1.75 / AUC .910 (Cao et al., 2022); RA-HMD acc 88.2 / AUC .935 (Mei et al., 2025)

Within each dataset the representation travels as it does for video. Cross-size recognition is .627 and .601 on Chinese (4,800 memes) and .745 and .749 on English (3,544 memes), with chance below .001. Coefficients moved from 7B to 3B unchanged reach .816 against a self-fit of .832 on Chinese and .890 against .897 on English, and none of 100 shuffles exceeds them in either direction. On the same 4,800 memes with only the question language changed, recognition is .985 and .991 and the coefficient loss .012 and .016 AUC; raw deployment nearly equals standardized deployment (.833 against .837) despite a median per-question shift of 1.19 standard deviations, because a pure shift moves no ranking. The aggregation law deviates by .006 (Chinese) and .010 (English). About 100 labels pass the GPT-4 zero-shot line on the Chinese benchmark (Lu et al., 2024) and GPT-4o on the English benchmark (Mei et al., 2025), while matching U-CoT+ (Pan et al., 2026) requires nearly the full label set; on HarMeme the readout's 84.2 is above U-CoT+'s 83.9 by about one test meme, and purpose-trained systems remain ahead. Between the two meme datasets transfer (.62 and .66) matches the shuffled reference; since changing the question language on the same memes barely moves the readout, the gap comes from the datasets and not from the language.

J TEST-SET USE AND CHECKS NOT YET RUN

Test-set use. The official splits are used as published. The protocol of the headline readout (all questions, $C = .003$, fitted on the training and validation sets) was fixed on development data. Three choices were made after test results had been seen: the in-domain threshold of .5, which out-of-fold validation also supports; the causal corroboration pool, chosen because it uses no later videos, although the pool that includes later videos scored higher; and the linear head, which development cross-validation also favours (Tab. A.4).

Checks not yet run. The second stage of the negative control was scored on Qwen2.5-VL-7B only; the sources of the in-domain association that unrelated questions exploit (topic, style, production) were not separated one by one; the free-form comparison changes question generation and pooling together, and its fixed-bank side is a 204-question readout. The paired difference against the fine-tuned detector has an interval spanning zero, which does not prove equivalence. Corroboration coordinates are read as availability of corroborating material, and a content-swap test (the same number of excerpts from unrelated or contradicting events) that would establish what they measure has not been run. The design ablations answer whether a block is needed given the rest of the bank; whether drafting by a design grid beats unstructured drafting of the same number of questions would

need a new bank and has not been tested. The declaration of inputs was not ablated either; the only evidence for it comes from an earlier generation of consistency questions, which saw both sides symmetrically and turned out to measure topic similarity.