DOES HIGH ACCURACY IMPLY TRUSTWORTHINESS? A LARGE-SCALE EVALUATION OF CONFIDENCE AND RISK IN FROZEN VISION AND VIDEO FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As large-scale vision and video foundation models enter high-risk settings, a standard safeguard is selective prediction, where a model acts only on high-confidence outputs while deferring uncertain ones. However, accuracy alone does not reveal whether a model's most confident predictions are reliably correct. Common uncertainty metrics also do not directly address this issue. AUROC and expected calibration error (ECE) evaluate confidence across the full prediction set, FPR@95 evaluates at a fixed error-rejection target, and other metrics remain confounded by model accuracy. In this work, we focus on the trustworthiness of a model's confidence levels among predictions selected for action. We propose that the positive likelihood ratio (LR+), originally defined for diagnostic testing, can be adapted to measure how much more often correct than incorrect predictions are included among a fixed fraction of the most-confident predictions. We evaluate LR+ across 18 frozen vision and video foundation-model backbones from six families and three scales, spanning image classification, semantic segmentation, video classification, and action anticipation. On ImageNet-1K, DINOv3 ViT-H+ and SigLIP2 g-opt achieve nearly identical top-1 accuracy (88.4% vs. 88.1%), yet their LR+ values among the top 20% most-confident predictions differ by more than twofold (24.0 vs. 49.7). Across all 18 models, AUROC and FPR@95 show little variation and rank models differently from LR+, which reveals substantial differences. Across downstream tasks, the relationship between accuracy and LR+ changes, and different models achieve the highest LR+. These results show that the trustworthiness of a model based on its confidence scores cannot be inferred from accuracy alone or assumed to transfer across benchmarks, and should be evaluated for each task at the fraction of predictions intended for action.

1 INTRODUCTION

Large-scale vision and video foundation models achieve remarkable accuracy across image classification, semantic segmentation, video classification, and action anticipation (Siméoni et al., 2025; Mur-Labadia et al., 2026; Wang et al., 2023; 2024; Tschannen et al., 2025; Chuang et al., 2025). As these models move into safety-critical settings, from autonomous systems and robotics to medical imaging, accuracy alone is insufficient: a model must also signal when its predictions should be trusted. A common risk-averse design allows the system to abstain, acting on its most-confident predictions and deferring the rest to a human operator, fallback controller, or expert review (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Nair et al., 2020). Two models with the same overall accuracy can therefore behave differently during deployment if one makes fewer errors among its most-confident predictions. What matters, then, is whether the predictions chosen for action are reliably correct, not simply whether confidence ranks predictions well overall. We therefore need uncertainty metrics that evaluate the trustworthiness of a model's confidence levels specifically where a system chooses to act, enabling systematic comparisons across frozen foundation models.

The metrics commonly used to evaluate confidence scores either confound confidence quality with overall accuracy, or do not directly capture trustworthiness where the system acts. Negative log-

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.

likelihood (NLL) and Brier score depend partly on whether predictions are correct. Risk-coverage curves plot error rate as low-confidence predictions are removed; because the curve starts at the model's overall error rate, models with different accuracies are not directly comparable without normalization (Geifman & El-Yaniv, 2017; Galil et al., 2023). Apparent differences in trustworthiness can therefore reflect overall model accuracy rather than confidence quality. Metrics such as AUROC for failure detection (Hendrycks & Gimpel, 2017) avoid this direct dependence on accuracy, but summarize confidence over the full prediction set rather than the high-confidence region used for action. A model can therefore have strong AUROC even if some of its most-confident predictions are incorrect, which matters when only high-confidence predictions are accepted. FPR@95 evaluates a threshold chosen to reject 95% of errors and reports the fraction of correct predictions that are also rejected. Expected calibration error (ECE) (Guo et al., 2017) measures whether confidence matches empirical accuracy on average (lower is better), but a low ECE can still hide high-confidence errors. Therefore, these metrics either remain influenced by overall accuracy or do not directly evaluate trustworthiness at a chosen fraction of the most-confident predictions.

To evaluate the trustworthiness of confidence directly, we adapt the positive likelihood ratio (LR+) from diagnostic testing (Deeks & Altman, 2004). At a chosen confidence cutoff, LR+ is the ratio of the fraction of correct predictions retained to the fraction of incorrect predictions retained. An LR+ of 10 means that, after filtering out low-confidence predictions, correct predictions remain ten times more often than errors. This directly measures how well confidence separates correct from incorrect predictions where the system acts. LR+ directly addresses both limitations described above. Unlike risk-coverage curves, it separately measures the proportion of correct and incorrect predictions kept, allowing confidence quality to be compared across models with different accuracies. LR+ is also evaluated at the specific fraction selected for action rather than averaged across all thresholds like AUROC, so it can reveal differences concentrated among the most-confident predictions.

In this work, we present the first large-scale evaluation of the trustworthiness of a model's confidence scores across modern frozen vision and video backbones, where each pretrained backbone is kept fixed and a task-specific probe is trained on its representations (Siméoni et al., 2025; Mur-Labadia et al., 2026; Tschannen et al., 2025). Previous work has studied uncertainty and calibration in neural networks (Guo et al., 2017), including under distribution shift (Ovadia et al., 2019) and across modern image classifiers (Minderer et al., 2021). However, confidence trustworthiness has not been systematically evaluated across the current generation of large-scale frozen vision and video backbones, which span distinct pretraining paradigms, model scales, and downstream tasks. We evaluate the maximum softmax probability scores produced by these probes without post-hoc recalibration. We evaluate LR+ while filtering out less-confident predictions and use the top 20% most-confident predictions as our main comparison point. We report AUROC, FPR@95, ECE, and signed ECE alongside LR+ to determine where standard uncertainty measures agree with or differ from LR+.

Our central finding is that higher accuracy does not imply more trustworthy confidence, and standard uncertainty measures can hide substantial differences between models. On ImageNet-1K, the most accurate backbone, DINOv3 ViT-H+ (88.4%), has an LR+ roughly half that of SigLIP2 g-opt at the top-20% setting (24.0 versus 49.7) despite nearly identical accuracies. At that fraction, DINOv3 therefore has approximately twice as many errors as SigLIP2. Yet standard uncertainty metrics show little variation across models and do not consistently agree with LR+ on which models perform better. Furthermore, the relationship between accuracy and LR+ is task-dependent: higher-accuracy models tend to achieve higher LR+ on some tasks, but not on others. Together, these results show that confidence trustworthiness cannot be inferred from accuracy or assumed to transfer across tasks.

We establish these findings across six model families spanning video self-supervision (V-JEPA 2.1, VideoMAE v2, InternVideo2), image self-supervision (DINOv3), and vision-language contrastive pretraining (SigLIP2, MetaCLIP2), at three scales each, across image classification, semantic segmentation, video classification, and action anticipation. Our contributions are as follows:

1. We present the first large-scale evaluation of the trustworthiness of model confidence scores across frozen vision and video foundation models, spanning six families, three scales, and four task domains, and adapt LR+ as a metric of trustworthy confidence under abstention.
2. We show that higher accuracy among models does not imply more trustworthy confidence levels as measured by LR+: on ImageNet-1K, the most accurate model has one of the lowest LR+ values when acting on its most-confident predictions.

3. We show that standard uncertainty measures can understate differences between models at the fraction selected for action and rank models differently from LR+: AUROC and FPR@95 vary little across models where LR+ reveals an approximately twofold gap, while ECE measures average calibration rather than reliability among selected predictions.
4. Using LR+, we show that the relationship between accuracy and trustworthy confidence varies across tasks, with different models achieving the highest LR+.

By directly evaluating confidence where a system acts, LR+ provides a practical step toward safer high-risk deployment. This can support risk-averse policies that act on confident predictions and defer uncertain cases. We also provide recommendations for evaluating abstention across models, tasks, and selected fractions.

2 Related Work

Calibration. Calibration measures whether a model's confidence scores match how often its predictions are correct. Guo et al. (2017) showed that widely used neural networks were often overconfident and that post-hoc temperature scaling could substantially improve calibration. Ovadia et al. (2019) subsequently showed that calibration often deteriorates under distribution shift. Minderer et al. (2021) found that more recent ImageNet classifiers, particularly non-convolutional architectures, were often better calibrated than earlier models. More recent studies of foundation models have reported different calibration patterns. Adler et al. (2026) found improved calibration in time-series foundation models relative to conventional baselines, while Hekler et al. (2025) reported underconfidence in several current-generation vision models. We report ECE and signed ECE for comparison, but our focus differs from calibration: calibration evaluates whether confidence matches accuracy, whereas we ask whether confidence identifies predictions that are trustworthy for action.

Selective prediction and failure detection. Selective prediction formalizes abstention by allowing a model to reject low-confidence predictions rather than predicting on every example (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Risk-coverage curves rank predictions by confidence and plot the error rate as progressively more low-confidence predictions are removed. Although they show performance across coverage levels, selective risk also depends on the model's underlying accuracy. A more accurate model can have lower risk even when its confidence is no better at distinguishing correct predictions from errors. Raw risk-coverage curves are therefore difficult to use for comparing confidence quality across models with different accuracies. AURC summarizes the full curve, while E-AURC (Geifman et al., 2019) and NAURC (Cattelan & Silva, 2024) aim to improve cross-model comparisons or reduce dependence on predictive accuracy. These metrics reduce the effect of accuracy differences but summarize performance across the full risk-coverage curve rather than specifically at the high-confidence fraction selected for action.

Failure detection asks how well confidence separates correct from incorrect predictions. Hendrycks & Gimpel (2017) established maximum softmax probability (MSP) as a baseline and evaluated it using AUROC and AUPR. FPR@95 is also common in failure prediction (Corbière et al., 2019). Under our convention, it rejects 95% of incorrect predictions and reports the fraction of correct predictions also rejected. Ding et al. (2020) argued that risk-coverage evaluation is better suited to selective prediction than AUROC and AUPR. They also showed that conventional ECE can hide substantial calibration errors because of its binning procedure, including errors in the high-confidence region. Together, these limitations show that strong confidence ranking alone does not necessarily tell us how trustworthy the predictions selected for action will be.

Most closely related to our evaluation, Galil et al. (2023) benchmark uncertainty across 523 pretrained ImageNet classifiers and introduce the Selective Accuracy Constraint (SAC). SAC reports the maximum fraction retained while meeting a specified accuracy. It therefore fixes a target performance level and asks how many predictions can be retained. Our deployment question is complementary: we evaluate whether confidence can be trusted among the predictions chosen for action. Galil et al. also showed that AURC and E-AURC remain sensitive to accuracy, while AUROC summarizes confidence over the full ranking.

Uncertainty under the frozen protocol. Frozen evaluation measures foundation-model transfer by holding pretrained backbones fixed and training task-specific probes (Siméoni et al., 2025; Mur-Labadia et al., 2026; Tschannen et al., 2025). However, these evaluations emphasize predictive

performance rather than confidence. GroVE (Venkataramanan et al., 2025) studied uncertainty in frozen vision-language models by learning an additional probabilistic model over their embeddings. We instead evaluate confidence already produced from the standard downstream probe. To our knowledge, the trustworthiness of confidence under the frozen-probe protocol has not been systematically evaluated across modern foundation-model families, scales, and image and video tasks.

3 METHODS

We evaluate the trustworthiness of confidence scores produced by frozen vision and video foundation-model systems across downstream tasks. Rather than proposing a new uncertainty method or recalibrating existing scores, we evaluate how reliably high confidence indicates correct rather than incorrect predictions. This determines whether confidence can be trusted for action selection. Our evaluation applies to any score that indicates prediction certainty. We use maximum softmax probability (MSP) as our primary confidence score. MSP is the largest softmax probability, and correctness indicates whether the predicted class matches the label.

3.1 METRICS

We report conventional failure-detection and calibration metrics alongside LR+, our primary metric for evaluating the trustworthiness of confidence where a system acts.

Failure detection and calibration. AUROC measures how consistently correct predictions receive higher confidence than incorrect predictions across the full confidence ranking. Treating incorrect predictions as the positive class, FPR@95 chooses a threshold that rejects 95% of incorrect predictions. It reports the fraction of correct predictions rejected at that threshold (lower is better). We also report expected calibration error (ECE) using 100 equal-mass bins following Minderer et al. (2021), together with signed ECE (sECE) to distinguish overconfidence from underconfidence.

Evaluating confidence under abstention with LR+. Under abstention, confidence must be evaluated at the fraction of predictions the system is intended to retain. Borrowing from diagnostic testing (Deeks & Altman, 2004), we propose the positive likelihood ratio (LR+) to measure how strongly the confidence filter retains correct predictions over incorrect predictions.

We define the retention fraction q, where $0 < q \leq 1$, as the proportion of predictions a model keeps for evaluation or deployment, while deferring the remaining $1 - q$ lowest-confidence predictions. Predictions are ranked by confidence, and we denote the retained set by $R(q)$. For example, $q = 0.20$ means retaining the 20% most-confident predictions and deferring the remaining 80%. Let N_C and N_E denote the total numbers of correct and incorrect predictions, and let $N_{C,R(q)}$ and $N_{E,R(q)}$ denote the numbers of correct and incorrect predictions in $R(q)$. We compute

$$\mathrm{LR}^{+}(q) = \frac{N_{C,R(q)}/N_C}{N_{E,R(q)}/N_E}.$$

LR+ is the ratio of the fraction of correct predictions retained to the fraction of incorrect predictions retained. An LR+ of 10 means that confidence filtering retains correct predictions at ten times the rate it retains incorrect predictions. This directly measures how well confidence separates predictions worth acting on from those that should be deferred, without being driven by overall accuracy. We report LR+ curves as progressively more low-confidence predictions are filtered out. Point comparisons require at least 20 observations in each group: retained-correct, retained-incorrect, rejected-correct, and rejected-incorrect. The top 20% satisfies this criterion for every model on the primary tasks and is our prespecified comparison point. Diving48 instead requires 50% retention because of its smaller evaluation set.

3.2 MODELS AND THE FROZEN-PROBE PROTOCOL

We evaluate six foundation-model families spanning three pretraining paradigms: video self-supervision (V-JEPA 2.1 (Mur-Labadia et al., 2026), VideoMAE v2 (Wang et al., 2023), InternVideo2 (Wang et al., 2024)), image self-supervision (DINOv3 (Siméoni et al., 2025)), and vision–language contrastive pretraining (SigLIP2 (Tschannen et al., 2025), MetaCLIP2 (Chuang et al.,

2025)). Each family is evaluated at ViT-B, ViT-L, and the largest available model compatible with our video protocol, yielding 18 backbones. All pretrained weights remain frozen across tasks. For classification and action anticipation, we train an attentive probe (Bardes et al., 2024) using the V-JEPA 2.1 evaluation implementation (Mur-Labadia et al., 2026). For semantic segmentation, we follow its dense-evaluation protocol and train a linear decoder. Within each task, all backbones receive the same learning-rate and weight-decay search. We select probe configurations using validation performance; uncertainty metrics are not used for model selection. Complete model, probe, and optimization details are provided in Appendix A.

3.3 Tasks and datasets

We evaluate every backbone on five in-distribution datasets spanning four task domains (Table 4): ImageNet-1K image classification (Deng et al., 2009; Russakovsky et al., 2015), Something-Something-v2 (SSv2) and Diving48 video classification (Goyal et al., 2017; Li et al., 2018), EPIC-KITCHENS-100 (EK100) action anticipation (Damen et al., 2022), and ADE20K semantic segmentation (Zhou et al., 2019). Image backbones encode video frames independently before attentive aggregation, while video backbones receive images in their expected temporal format. For multi-view video evaluation, we average softmax probabilities before computing the final prediction and confidence. EK100 predicts an upcoming verb and noun, which we evaluate separately. We omit joint verb–noun prediction because more than 3,000 classes are too sparse for meaningful confidence comparison. For ADE20K, confidence and correctness are evaluated for pixels assigned to one of its 150 semantic classes; all other pixels are excluded. Further details are provided in Appendix B.

3.4 MC-dropout uncertainty scores

We additionally evaluate MC dropout (Gal & Ghahramani, 2016) as an alternative source of confidence, using uncertainty scores derived from its predictive distribution to rank predictions by certainty on ImageNet-1K, with a dropout rate of 0.10 consistently across backbones. Accuracy and calibration are computed from the predictive distribution averaged over stochastic forward passes, while failure detection and LR+ use uncertainty measures such as predictive entropy, mutual information, and variance. Predictive entropy is our primary uncertainty score, following standard practice (Ovadia et al., 2019). Sampling details are provided in Appendix E.4.

3.5 Statistical protocol

We use a single training seed for each selected probe configuration since our benchmark spans 18 backbones and the full downstream task suite. For each reported metric, we generate 1,000 bootstrap samples by resampling the evaluation examples with replacement and recomputing the metric on each sample. We then exclude the lowest 2.5% and highest 2.5% of the resulting estimates and report the range of the remaining estimates as the 95% confidence interval. For ADE20K, we resample whole images rather than individual pixels. This keeps pixels from the same image together rather than treating each pixel as a separate, independent observation. The 95% confidence intervals do not capture variability from probe initialization or training stochasticity.

For the main ImageNet-1K comparison, we define models as accuracy-matched when their top-1 accuracies differ by less than 0.5 percentage points. We use a paired bootstrap for head-to-head comparisons, resampling the same examples for both backbones and recomputing the differences in accuracy and LR+. We report the magnitude and 95% confidence interval of the accuracy difference alongside the LR+ comparison.

4 Results

4.1 ImageNet-1K: Accuracy vs. Trustworthy Confidence

On ImageNet-1K, nearly identical accuracy can mask large differences in trustworthy confidence. DINOv3 ViT-H+ achieves 88.4% top-1 accuracy compared with 88.1% for SigLIP2 g-opt. At 20% retention, however, their LR+ values are 24.0 and 49.7 respectively, a more than twofold difference. Paired-bootstrap analyses confirm both the small accuracy difference and the large LR+ difference

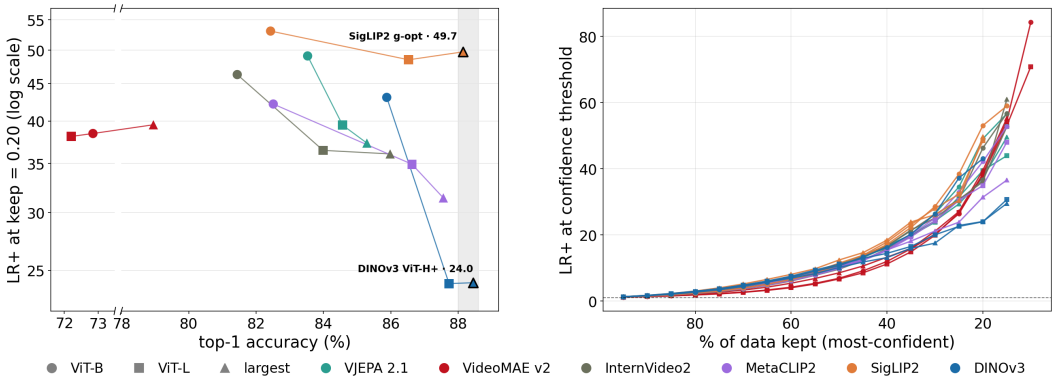


Figure 1: **Accuracy alone does not predict trustworthy confidence on ImageNet-1K, and model differences in LR+ persist across retained fractions.** In both panels, colour denotes model family, while circles denote ViT-B, squares ViT-L, and triangles the largest model from each family evaluated under our video protocol. **Left:** Top-1 accuracy against LR+ at 20% retention. Lines connect the three model scales within each family to show how accuracy and LR+ change with scale. The shaded grey vertical band marks the accuracy-matched region containing DINOv3 ViT-H+ and SigLIP2 g-opt (outlined markers), which have nearly identical accuracy but differ by more than twofold in LR+ (24.0 and 49.7). **Right:** LR+ against the fraction of predictions retained after confidence filtering; moving right, fewer predictions are kept and only the most-confident remain. All models approach LR+= 1 at full retention, while DINOv3 ViT-L and ViT-H+ remain among the lowest-LR+ models across the high-confidence region and SigLIP2 models among the highest.

(Appendix D.2.1). This gap is not simply explained by the models failing on different examples: 4,063 ImageNet-1K examples are misclassified by both, roughly 70% of each model's errors. The LR+ gap therefore reflects how the two models rank their errors by confidence.

The difference between DINOv3 and SigLIP2 is not specific to the 20% retained fraction. Figure 1 (right) shows LR+ as progressively more low-confidence predictions are filtered out. When most predictions are retained, LR+ values are similar across models. As more are filtered out and only higher-confidence predictions remain (moving from left to right), the curves separate. Once 40% or fewer of the most-confident predictions remain, DINOv3 ViT-L and ViT-H+ are among the lowest-LR+ models and SigLIP2 models among the highest. The difference at 20% therefore reflects a consistent separation across the high-confidence region rather than a result at a single filtering level.

The DINOv3 and SigLIP2 comparison reflects a broader pattern on ImageNet-1K: across all 18 backbones, higher accuracy is not consistently associated with higher LR+ (Figure 1, left). SigLIP2 ViT-B and MetaCLIP2 ViT-B achieve similar accuracy, but SigLIP2 has a substantially higher LR+. More broadly, lower-accuracy models can even outperform higher-accuracy ones: SigLIP2 ViT-B is less accurate than DINOv3 ViT-H+ yet has a much higher LR+. Thus, better ImageNet-1K accuracy does not imply better trustworthy confidence.

4.2 Standard uncertainty metrics do not reflect differences in trustworthy confidence

We next examine whether standard uncertainty metrics (AUROC, FPR@95, ECE) reveal the same differences in confidence between models as LR+ (Figure 2). On ImageNet-1K, AUROC and FPR@95 make the 18 backbones look much more similar than LR+ does. AUROC ranges only from 0.864 to 0.885 and FPR@95 from 0.42 to 0.46. In contrast, when only the top 20% most-confident predictions are considered, LR+ ranges from 24.0 to 53.1. The two metrics also disagree with LR+ about which backbones are strongest. DINOv3 ViT-L ranks among the highest models by AUROC (0.879), near the middle by FPR@95 (0.442), and last by LR+ (24.0). VideoMAE v2 ViT-L shows the reverse trend, reaching a higher LR+ (38.1) than DINOv3 ViT-L despite a worse FPR@95 (0.458). Thus, models that appear very similar under the standard metrics can differ substantially in trustworthy confidence, and can be ordered differently by LR+.

Expected calibration error (ECE) also does not reveal the same differences in trustworthy confidence as LR+, as it is computed over the full prediction set rather than among the predictions a model is most confident about. DINOv3 ViT-L and SigLIP2 ViT-L have nearly identical ECE values (0.040 and 0.041), yet their LR+ values differ twofold (24.0 and 48.5). This pattern is also seen across models of different size, where V-JEPA 2.1 ViT-B and MetaCLIP2 ViT-bigG have nearly identical ECE (0.049 and 0.050) but LR+ differing by more than 1.5× (49.1 and 31.4). In other words, two models can look similarly calibrated while differing substantially in how well their most-confident predictions separate correct predictions from errors.

4.3 The relationship between accuracy and trustworthy confidence is task-dependent

The relationship between accuracy and trustworthy confidence changes across downstream tasks. Figure 2 shows accuracy and uncertainty metrics for ImageNet-1K and SSv2, while Diving48, EK100 (verb and noun), and ADE20K are shown in Figure 3, with all LR+ retention curves in Figure 4 (Appendix E). On some tasks, the most accurate models also achieve the highest LR+, while on others, models with similar or lower accuracy can have substantially better LR+.

This variation is clear even across the two video classification datasets. On SSv2, V-JEPA 2.1 achieves both the highest accuracies and the highest LR+ values, showing that higher accuracy and higher trustworthy confidence can coincide. However, this does not hold for every comparison: within DINOv3, ViT-B has the lowest accuracy but the highest LR+ of the three model sizes. Diving48 shows a different result (Figure 3). MetaCLIP2 ViT-L has substantially lower accuracy than DINOv3 ViT-L, yet achieves the highest LR+ of all 18 models. Diving48 contains only 1,970 validation examples, so LR+ confidence intervals are wide and small differences should not be over-interpreted. Together, these two datasets show that even within video classification, the relationship between accuracy and trustworthy confidence depends on the dataset.

Action anticipation and semantic segmentation provide further examples of task-dependent behavior. On EK100, V-JEPA 2.1 and DINOv3 are among the stronger models in both accuracy and LR+, although LR+ values are relatively similar across all models. On ADE20K, DINOv3 ViT-H+ achieves both the highest accuracy and the highest LR+. However, models with similar accuracy can still differ substantially in trustworthy confidence: V-JEPA 2.1, SigLIP2, and MetaCLIP2 occupy a similar accuracy range but show much larger differences in LR+.

Taken together, these results show that trustworthy confidence is not a fixed property of a model. A model that has high LR+ on one task may have much lower LR+ on another, and accuracy does not reliably indicate which models will separate their correct predictions from their incorrect predictions most effectively. Consequently, uncertainty conclusions drawn from a single benchmark cannot be assumed to generalize to other downstream tasks. Trustworthy confidence must therefore be evaluated on the task where the model will actually be used.

4.4 Scaling improves accuracy more consistently than trustworthy confidence

On ImageNet-1K, increasing model scale consistently improves accuracy but usually reduces trustworthy confidence. Accuracy increases from ViT-B to the largest evaluated model in all six families, whereas LR+ decreases in five of six. The largest drops are observed for DINOv3 (-44%), MetaCLIP2 (-26%), V-JEPA 2.1 (-24%), and InternVideo2 (-22%). SigLIP2 changes little (-6%), while VideoMAE v2 increases slightly ($+3\%$). Thus, gains in ImageNet-1K accuracy from scaling do not generally translate into better trustworthy confidence.

This scaling behavior also depends on the downstream task. The consistent ImageNet-1K decrease in LR+ does not appear across the other tasks, where larger models more often improve both accuracy and LR+. DINOv3 and SigLIP2, the two families compared in Section 4.1, also respond very differently to scaling on ImageNet-1K: DINOv3 loses 44% of its LR+ from ViT-B to its largest model, while SigLIP2 largely preserves it. Scaling therefore changes accuracy and trustworthy confidence differently depending on both the model family and task.

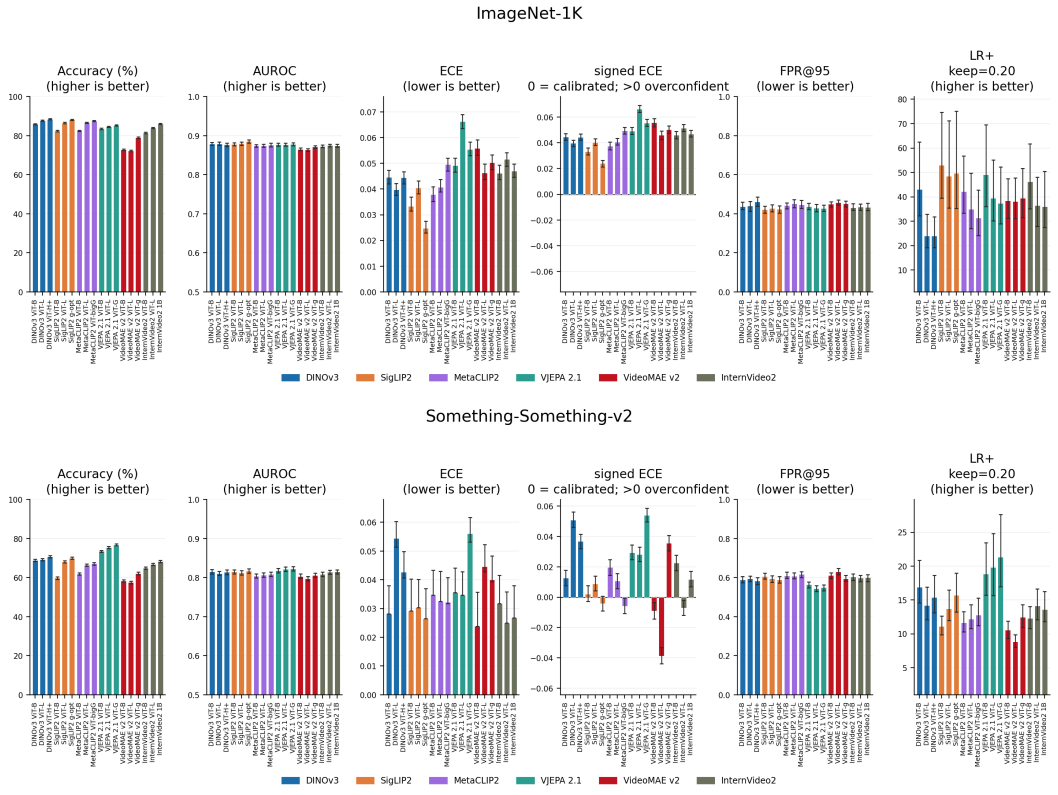


Figure 2: **Accuracy and uncertainty metrics on ImageNet-1K and SSv2.** Rows correspond to ImageNet-1K image classification and Something-Something-v2 (SSv2) video classification. Columns report accuracy, AUROC, ECE, signed ECE, FPR@95, and LR+, respectively. In each panel, the x-axis shows the 18 evaluated models and the y-axis shows the corresponding metric value; error bars indicate 95% bootstrap confidence intervals. LR+ is evaluated among the top 20% most-confident predictions. Accuracy and trustworthy confidence do not vary together consistently across tasks. On ImageNet-1K, AUROC and FPR@95 show little variation across the 18 models (0.864-0.885 and 0.42-0.46, respectively) despite large differences in LR+ (24.0-53.1), while the relationship between accuracy and LR+ changes across the other downstream tasks (the remaining four evaluations are shown in Figure 3, Appendix E).

4.5 Differences in trustworthy confidence persist across uncertainty measures

Our main analyses use maximum softmax probability (MSP) from a deterministic forward pass. We next test whether our conclusions hold when uncertainty is instead estimated using MC dropout in the classifier head. Averaging predictions over stochastic passes reduces ECE for 17 of the 18 backbones on ImageNet-1K, with the median falling from 0.046 to 0.037. Using predictive entropy instead of MSP, median LR+ at 20% retention changes little, from 39.0 to 37.8. We also evaluate mutual information and predictive variance, which achieve lower AUROC than predictive entropy on every backbone; median LR+ remains near 39 for all four confidence sources. Thus, the distinction between calibration and trustworthy confidence persists across uncertainty measures. MC dropout improves average calibration but leaves median LR+ largely unchanged. Full results are in Appendix E.4.

5 Discussion and Limitations

To our knowledge, this is the first large-scale evaluation of model trustworthiness based on its confidence scores across frozen vision and video foundation models, spanning 18 backbones, six families,

three model scales, five datasets, and four task domains. A confidence score tells us how certain a model is, but not whether that certainty is justified: a model can be highly confident and still be wrong. Confidence must therefore be evaluated to determine whether high-confidence predictions are actually more likely to be correct. LR+ provides one way to do this. At a chosen fraction of predictions selected for action, LR+ compares the fraction of correct predictions selected with the fraction of incorrect predictions selected. This directly supports a risk-averse selection policy, where the system acts on its most-confident predictions and defers the rest. LR+ can be applied to MSP, predictive entropy, or any other score used to rank predictions by certainty, and provides an interpretable value at any selected fraction. LR+ is one example rather than the only metric for evaluating how confidence relates to correctness under selective prediction.

Our results show why confidence trustworthiness must be evaluated in addition to accuracy. On ImageNet-1K, two models with nearly identical accuracy differ more than twofold in LR+. A practitioner choosing a backbone for an abstaining system therefore cannot infer the reliability of its most-confident predictions from accuracy alone. AUROC, FPR@95, and ECE vary less across models and do not consistently agree with LR+, showing that conclusions about confidence trustworthiness depend on how confidence is evaluated. DINOv3 and SigLIP2 misclassify many of the same images but assign very different confidence to those errors, showing that trustworthiness depends on how errors are ranked by confidence. These findings also persist when MSP is replaced with MC-dropout predictive entropy: average calibration improves, but median LR+ remains largely unchanged. Across downstream tasks, both the relationship between accuracy and LR+ and the highest-ranking models change, so ImageNet-1K results should not be assumed to transfer.

Our results suggest three practical steps for evaluating an abstaining system. First, one should evaluate LR+ across multiple selected fractions to see how confidence trustworthiness changes as more predictions are selected, or report it at the fraction at which the system is expected to act. Second, one should report LR+ alongside overall accuracy because they evaluate distinct properties: accuracy measures how often the model is correct overall, whereas LR+ compares the fraction of all correct predictions kept with the fraction of all incorrect predictions kept. A model may have more correct high-confidence predictions simply because it is more accurate overall. A higher LR+ therefore indicates a more trustworthy confidence signal, even if another model is more accurate. Both are needed to determine how often the model will be correct when it acts. LR+ should therefore be interpreted as a measure of the confidence signal, not as a replacement for accuracy or a standalone criterion for choosing a model. Finally, one should repeat this evaluation for each intended application because confidence trustworthiness and model rankings may not transfer across tasks.

This study has several limitations. Our conclusions apply only to the frozen-probe setting studied here. Each backbone–task evaluation uses a single probe seed, so bootstrap intervals capture evaluation-set variability but not variation from probe training. Models of different sizes may also differ in patch size, pretraining data, and other design choices, so observed scaling trends cannot be attributed to model size alone. Diving48 contains only 1,970 validation examples, so its wider bootstrap intervals make LR+ differences harder to distinguish. LR+ can also be sensitive when very few incorrect predictions remain in the selected set, because a single mislabeled example can noticeably change the estimate. Finally, we evaluate uncalibrated MSP only on in-distribution data. Future work should examine post-hoc recalibration, distribution shift, and semantic novelty detection.

6 Conclusion

We provide the first large-scale evaluation of confidence trustworthiness across 18 frozen vision and video backbones spanning four task domains and propose LR+ for evaluating confidence under abstention. Models with similar accuracy can differ substantially in LR+, and rankings change across metrics and downstream tasks. On ImageNet-1K, replacing MSP with MC-dropout predictive entropy leaves median LR+ largely unchanged even as average calibration improves. As foundation models move into safety-critical settings, accuracy alone is insufficient. We must also determine whether confidence reliably signals when predictions should be trusted. Our evaluation is a step toward risk-averse systems that act on their most-confident predictions and defer the rest.

AI USE STATEMENT

Generative AI tools were used in two capacities in this work. All AI-assisted output was reviewed and verified by the authors.

Code. AI coding assistants were used to help write and modify the evaluation and analysis code, including the prediction-extraction scripts, the metric implementations, the bootstrap procedures, and the figure-generation code. No AI-generated code was accepted without independent verification. Specifically, we verified that all 18 backbone checkpoints load byte-identically to their released weights; that each model's input normalization matches its published preprocessing; that the label vectors are byte-identical across models within a task; and that LR+ returns approximately 1 under a confidence-shuffling control and degenerates as expected when confidence is set equal to correctness. The ImageNet-1K pipeline was additionally reviewed end to end.

Text. AI writing assistants were used to help draft and revise prose, captions, and appendix material.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the evaluated foundation models, probe architectures, optimization settings, task-specific evaluation protocols, uncertainty metrics, and statistical procedures in the appendix. We also report the hyperparameter search spaces, model-selection criteria, input configurations, bootstrap procedures, and pseudocode for the LR+ computation. These details are intended to support reproduction of the reported results.

REFERENCES

Coen Adler, Yuxin Chang, Samar Abdi, Felix Draxler, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated? In *International Conference on Learning Representations*, 2026.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Luís Felipe P. Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 547–584, 2024.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James R. Glass, Lifei Huang, Jason E. Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen-tau Yih, Shang-Wen Li, and Hu Xu. Meta clip 2: A worldwide scaling recipe. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2902–2913, 2019.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Jonathan J. Deeks and Douglas G. Altman. Diagnostic tests 4: likelihood ratios. *BMJ*, 329(7458): 168–169, 2004.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *International Conference on Learning Representations*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4885–4894, 2017.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5843–5851, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Achim Hekler, Lukas Kuhn, and Florian Buettner. Beyond overconfidence: Foundation models redefine calibration in deep neural networks. *arXiv preprint arXiv:2506.09593*, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision*, pp. 513–528, 2018.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Lorenzo Mur-Labadia, Matthew J. Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, 2020.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–41, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14003–14014, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Aishwarya Venkataramanan, Paul Bodesheim, and Joachim Denzler. Probabilistic embeddings for frozen vision-language models: Uncertainty quantification with gaussian process latent variable models. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4309–4328, 2025.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXV*, pp. 396–416, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73012-2.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

A Model and Probe Details

A.1 Foundation-model backbones

We evaluate 18 frozen backbones from six foundation-model families: V-JEPA 2.1 (Mur-Labadia et al., 2026), VideoMAE v2 (Wang et al., 2023), InternVideo2 (Wang et al., 2024), DINOv3 (Siméoni et al., 2025), SigLIP2 (Tschannen et al., 2025), and MetaCLIP2 (Chuang et al., 2025). For each family, we evaluate ViT-B, ViT-L, and the largest publicly available model that can be evaluated under our video protocol. All pretrained backbone parameters remain frozen throughout downstream probe training.

Table 1 reports the embedding dimension, input resolution, and pretraining paradigm of each evaluated backbone.

A.2 Attentive probe

For image classification, video classification, fine-grained video classification, and action anticipation, we use an attentive probe following Mur-Labadia et al. (2026). The probe applies three self-attention blocks to the frozen backbone token sequence, followed by a cross-attention block in which a learnable query token pools the result, with a residual connection adding the query back. Its embedding width is matched to the output dimension of the corresponding backbone. All other settings are shared across backbones, with feedforward layers four times the embedding width and 16 attention heads.

Matching probe width to backbone width increases the number of trainable probe parameters for larger backbones, from 28.5M parameters for 768-dimensional ViT-B representations to 131.9M for the 1664-dimensional MetaCLIP2 ViT-bigG representation. The DINOv3 ViT-H+ and SigLIP2

Table 1: Frozen backbones. Embedding dimension is read from the encoder modules; input resolution is the classification-eval resolution. ADE20K segmentation instead runs at 512 for patch-16 backbones and 518 for patch-14. All backbones use the publicly released weights of their respective families (Siméoni et al., 2025; Tschannen et al., 2025; Chuang et al., 2025; Wang et al., 2023; 2024; Mur-Labadia et al., 2026).

Family	Scale	Embed. dim	Input res.	Pretraining paradigm
DINOv3	ViT-B	768	384	Self-supervised self-distillation (image)
DINOv3	ViT-L	1024	384	Self-supervised self-distillation (image)
DINOv3	ViT-H+	1280	384	Self-supervised self-distillation (image)
SigLIP2	ViT-B	768	384	Image–text contrastive (sigmoid)
SigLIP2	ViT-L	1024	384	Image–text contrastive (sigmoid)
SigLIP2	g-opt	1536	384	Image–text contrastive (sigmoid)
MetaCLIP2	ViT-B	768	384	Image–text contrastive (CLIP)
MetaCLIP2	ViT-L	1024	392	Image–text contrastive (CLIP)
MetaCLIP2	ViT-bigG	1664	392	Image–text contrastive (CLIP)
VideoMAE v2	ViT-B	768	384	Masked video autoencoding
VideoMAE v2	ViT-L	1024	384	Masked video autoencoding
VideoMAE v2	ViT-g	1408	392	Masked video autoencoding
InternVideo2	ViT-B	768	392	Multimodal video (masked + contrastive)
InternVideo2	ViT-L	1024	392	Multimodal video (masked + contrastive)
InternVideo2	1B	1408	392	Multimodal video (masked + contrastive)
V-JEPA 2.1	ViT-B	768	384	Self-supervised joint-embedding prediction (video)
V-JEPA 2.1	ViT-L	1024	384	Self-supervised joint-embedding prediction (video)
V-JEPA 2.1	ViT-G	1408	384	Self-supervised joint-embedding prediction (video)

Table 2: Backbone blocks read by the ADE20K linear decoder. Four blocks are read per backbone at approximately even intervals, shared across backbones of the same depth. The backbone's final normalization layer is applied to each feature map before the linear head.

Family	Scale	Depth	Backbone layers read (0-indexed)
DINOv3	ViT-B/L/H+	12 / 24 / 32	[2,5,8,11] / [5,11,17,23] / [7,15,23,31]
SigLIP2	ViT-B/L/g-opt	12 / 24 / 40	[2,5,8,11] / [5,11,17,23] / [9,19,29,39]
MetaCLIP2	ViT-B/L/bigG	12 / 24 / 48	[2,5,8,11] / [5,11,17,23] / [11,23,37,47]
VideoMAE v2	ViT-B/L/g	12 / 24 / 40	[2,5,8,11] / [5,11,17,23] / [9,19,29,39]
InternVideo2	ViT-B/L/1B	12 / 24 / 40	[2,5,8,11] / [5,11,17,23] / [9,19,29,39]
V-JEPA 2.1	ViT-B/L/G	12 / 24 / 48	[2,5,8,11] / [5,11,17,23] / [11,23,37,47]

g-opt probes contain 78.4M and 112.5M parameters, respectively. We therefore interpret within-family scaling trends as reflecting the combined backbone–probe system rather than backbone scale alone.

For ADE20K semantic segmentation, we replace the attentive probe with a linear decoder, following the segmentation protocol of V-JEPA 2.1 (Mur-Labadia et al., 2026) and DINOv3 (Siméoni et al., 2025), except that we read four evenly spaced backbone blocks rather than the last block only. Table 2 lists the blocks used for each architecture.

A.3 Probe optimization

Learning-rate and weight-decay grids for each task are given in Table 3. For ImageNet-1K we additionally include a no-warmup baseline, giving 21 configurations in total. Probe configurations are selected by validation performance only: final-epoch accuracy for classification and action anticipation, and best-checkpoint mIoU for segmentation. No uncertainty metric is used for hyperparameter selection.

Probes are trained for 20 epochs on ImageNet-1K, SSv2, and EK100 and for 50 epochs on Diving48; ADE20K is trained for 40,000 iterations. The exact search grids, batch sizes, and optimization settings are reported in Table 3. The same search space and selection procedure are applied to every backbone within a task.

Table 3: Probe optimization. Batch is the effective global batch. The ADE20K learning rates and weight decays are four specific pairings rather than all combinations: (2e-3, 0.05), (5e-4, 0.05), (8e-5, 0.05), and (5e-4, 0.002).

Dataset	Schedule	Learning Rates	Weight Decays	Batch	Optimizer
IN1K	20 epochs	{5e-4, 1e-3, 1.5e-3, 2e-3}	{5e-4, 1e-3, 2e-3, 4e-3, 8e-3}	1024	AdamW
SSv2	20 epochs	{1e-4, 3e-4, 1e-3, 3e-3, 5e-3}	{1e-2, 1e-1, 4e-1, 8e-1}	256	AdamW
Diving48	50 epochs	{1e-4, 3e-4, 1e-3}	{1e-1, 4e-1, 8e-1}	256	AdamW
EK100	20 epochs	{1e-4, 3e-4, 1e-3, 3e-3, 5e-3}	{1e-4, 1e-3, 1e-2, 1e-1}	128	AdamW
ADE20K	40,000 iters	{2e-3, 5e-4, 8e-5, 5e-4}	{0.05, 0.05, 0.05, 0.002}	16	AdamW

Table 4: In-distribution evaluation suite.

Dataset	Task	Size	Classes	Prediction unit
ImageNet-1K	Image classification	50,000	1,000	Image
SSv2	Video classification	24,777	174	Video
Diving48	Fine-grained video classification	1,970	48 (47 present)	Video
EK100	Action anticipation	9,248	97 verb / 289 noun	Video
ADE20K	Semantic segmentation	2,000	150	Labeled class pixel

B Task and Input Details

B.1 In-distribution datasets

Table 4 summarizes the evaluation datasets. ImageNet-1K (Deng et al., 2009; Russakovsky et al., 2015) provides image classification, Something-Something-v2 (Goyal et al., 2017) video classification, Diving48 (Li et al., 2018) fine-grained video classification, EPIC-KITCHENS-100 (Damen et al., 2022) action anticipation, and ADE20K (Zhou et al., 2019) semantic segmentation.

The Diving48 validation split contains examples from 47 of the 48 nominal classes: one class is absent and seven contain at most four examples. Our confidence metrics pool predictions across all examples, so this sparsity affects only class-wise and macro-averaged analyses, which we do not report.

For EK100, verb and noun anticipation are evaluated as separate softmax prediction tasks over the same input clips. We therefore treat them as related rather than independent tasks. We omit the joint verb–noun action from our uncertainty analysis, as more than 3,000 possible action classes are too sparsely represented for a meaningful joint softmax-confidence comparison. The probe has three output heads, for verb, noun, and joint action, each trained with its own cross-entropy loss. Probe selection uses only the verb and noun heads: we choose the configuration with the highest fraction of validation clips for which both the verb and the noun are predicted correctly, with ties broken by the lowest validation loss. The joint action head is therefore used neither for selection nor in our uncertainty analysis. The standard EK100 anticipation protocol instead selects on top-5 mean-class recall over the joint action classes, which is unstable when many classes have few validation examples, so our EK100 accuracies are not directly comparable to results reported under that protocol.

For ADE20K, we select the decoder by validation mIoU, the standard segmentation metric. We report per-pixel accuracy because our confidence metrics require a correct or incorrect label for each prediction, which on this task is defined per pixel. Confidence and correctness are evaluated for pixels assigned to one of the dataset's 150 semantic classes; all other pixels are excluded. Statistical resampling is performed at the image level rather than the pixel level (Appendix D).

B.2 Applying image backbones to video

For image foundation models, each sampled video frame is encoded independently by the frozen image backbone. Patch tokens from all frames are concatenated into a single space–time sequence, and temporal positional embeddings are added before the attentive probe aggregates the sequence (Siméoni et al., 2025).

Table 5: Evaluation views. ImageNet-1K and ADE20K each use a single source image, with temporal formatting for video backbones described below. SSv2 and Diving48 use multi-view sampling (frames × temporal clips × spatial crops). EK100 uses a single 32-frame input clip, matching the V-JEPA 2.1 evaluation input length (Mur-Labadia et al., 2026); our anticipation head and model-selection procedure are described above. Classification resolution is 384 for patch-16 backbones and 392 for patch-14; ADE20K uses 512 and 518 respectively.

Dataset	Frames	Temporal views	Spatial crops	Resolution	Total views
IN1K	1	1	1	384/392	1
SSv2	16	2	3	384/392	6
Diving48	32	4	3	384/392	12
EK100	32	1	1	384/392	1
ADE20K	1	1	1	512/518	1

For multi-view video evaluation, we average the softmax probability vectors across views, then compute the final prediction and MSP from the averaged probabilities. Table 5 gives the frame count, temporal sampling, spatial crops, input resolution, and total number of evaluation views for each task.

B.3 Applying video backbones to images

For image tasks, V-JEPA 2.1 and InternVideo2 process an image as a length-one temporal sequence. VideoMAE v2 instead receives 16 replicated copies of the image to match its pretrained temporal input structure. With its tubelet size of two, this produces eight identical temporal tokens before subsequent transformer processing. Temporal positional embeddings are applied following the corresponding pretrained model.

C Metric Details

C.1 Maximum softmax probability

Given the final probability vector $\mathbf{p}_i$ for example i, maximum softmax probability is

$$s_i = \max_c p_{i,c}.$$

For classification and anticipation, correctness is

$$y_i = \begin{cases} 1, & \text{if } \arg\max_c p_{i,c} = y_i^\star, \\ 0, & \text{otherwise.} \end{cases}$$

where $y_i^\star$ is the ground-truth class. For ADE20K, the same quantities are computed independently for every pixel assigned to one of the dataset's 150 semantic classes.

We evaluate MSP directly without post-hoc recalibration. This isolates the confidence emitted by the frozen-probe system itself.

C.2 Positive likelihood ratio

For a retained fraction q, we compute the corresponding confidence threshold and retain predictions above that threshold. We use the following notation:

- N_C: total number of correct predictions;
- N_E: total number of incorrect predictions;
- $N_{C,R}$: number of correct predictions in the retained high-confidence set;
- $N_{E,R}$: number of incorrect predictions in the retained high-confidence set;

- $N_{C,U} = N_C - N_{C,R}$: number of correct predictions rejected as low-confidence;
- $N_{E,U} = N_E - N_{E,R}$: number of errors rejected as low-confidence.

The resulting 2×2 table is

	Retained	Rejected
Correct	$N_{C,R}$	$N_{C,U}$
Error	$N_{E,R}$	$N_{E,U}$

The positive likelihood ratio is

$$\mathrm{LR}^{+}(q) = \frac{N_{C,R}/N_C}{N_{E,R}/N_E}.$$

The numerator is the fraction of all correct predictions retained and the denominator is the fraction of all errors retained. Thus, $\mathrm{LR}^{+} = 10$ means that the confidence filter retains correct predictions at ten times the rate at which it retains errors.

At fixed q, every model is evaluated under the same retained-fraction constraint, although the numerical MSP threshold required to retain that fraction can differ between models.

C.3 Minimum-count criterion and reported retained fractions

LR+ can be highly variable when one of the four retained/rejected $\times$ correct/error groups contains very few examples. For point comparisons, we therefore require at least 20 observations in each of the four cells

$$N_{C,R}, \quad N_{E,R}, \quad N_{C,U}, \quad N_{E,U}.$$

A retained fraction of $q = 0.20$ satisfies this requirement across the models in our primary task comparisons. Diving48 requires $q = 0.50$ because of its smaller evaluation set. We report full LR+ curves in addition to these point comparisons.

C.4 Calibration metrics

We compute ECE using 100 equal-mass bins following Minderer et al. (2021). Let B_b denote bin b, N the total number of predictions, $\mathrm{acc}(B_b)$ its empirical accuracy, and $\mathrm{conf}(B_b)$ its mean confidence. Then

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\right|.$$

Signed ECE removes the absolute value,

$$\mathrm{sECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left(\mathrm{conf}(B_b) - \mathrm{acc}(B_b)\right),$$

where positive values indicate overconfidence and negative values indicate underconfidence.

Because ECE depends on the binning scheme (Minderer et al., 2021; Nixon et al., 2019), we report ECE at a single fixed bin count throughout and do not interpret small differences in ECE as meaningful.

Algorithm 1 LR+ at fixed retained fraction

Require: Confidence scores $s_1, \ldots, s_N$
Require: Correctness indicators $y_1, \ldots, y_N \in \{0, 1\}$
Require: Retained fraction $q \in (0, 1)$
Ensure: $\mathrm{LR}^+(q)$ and contingency-cell counts

Sort examples by confidence s_i from highest to lowest
$\tau \leftarrow \mathrm{quantile}(s, 1 - q)$
Retain examples with $s_i > \tau$
$N_C \leftarrow \sum_{i=1}^{N} y_i$
$N_E \leftarrow N - N_C$
$N_{C,R} \leftarrow$ number of retained examples with $y_i = 1$
$N_{E,R} \leftarrow$ number of retained examples with $y_i = 0$
$N_{C,U} \leftarrow N_C - N_{C,R}$
$N_{E,U} \leftarrow N_E - N_{E,R}$
if $\min(N_{C,R}, N_{E,R}, N_{C,U}, N_{E,U}) < 20$ **then**
    Flag this retained fraction as below the point-comparison stability criterion
end if
$r_C \leftarrow N_{C,R}/N_C$ ▷ fraction of correct predictions retained
$r_E \leftarrow N_{E,R}/N_E$ ▷ fraction of errors retained
$\mathrm{LR}^+(q) \leftarrow r_C/r_E$
return $\mathrm{LR}^+(q)$, $N_{C,R}$, $N_{E,R}$, $N_{C,U}$, $N_{E,U}$

C.5 Pseudocode for LR+

Algorithm 1 gives the procedure used to compute LR+ at a fixed retained fraction.

For LR+ curves, we evaluate target retained fractions from $q = 0.95$ to $q = 0.05$ in steps of 0.05. For each q, we compute the corresponding confidence threshold τ and retain predictions with confidence greater than τ. The main comparison uses $q = 0.20$, while Diving48 uses $q = 0.50$.

Thresholds are computed using linear-interpolated quantiles. Predictions tied at the threshold are all rejected, so the realized retained fraction can be slightly smaller than the target fraction.

D Statistical Details

D.1 Bootstrap confidence intervals

Each selected probe is trained with a single seed. We therefore use bootstrap resampling to quantify uncertainty from the finite evaluation sample rather than variability across probe training runs. For each metric, we generate 1,000 bootstrap samples by sampling evaluation examples with replacement and report the 2.5th and 97.5th percentiles as a 95% confidence interval. For unsigned ECE, bootstrap intervals can be shifted upward relative to the full-sample estimate because resampling increases the expected absolute calibration gap, particularly on smaller evaluation sets.

For ADE20K, the bootstrap unit is the image. We sample 2,000 images with replacement and include all pixels assigned to one of the dataset's classes from each sampled image before recomputing the metric. This keeps pixels from the same image together rather than treating each pixel as a separate, independent observation.

D.2 Paired model comparisons

For direct comparisons between two models evaluated on the same examples, we use a paired bootstrap. A bootstrap sample of evaluation indices is drawn once and applied to both models, after which the difference in the target metric is recomputed. Repeating this procedure 1,000 times yields the bootstrap distribution of the paired difference.

For the primary ImageNet-1K comparison, we report paired differences in both top-1 accuracy and LR+. Models are considered practically accuracy-matched when $|\Delta\mathrm{Acc}| < 0.5$ percentage points.

Table 6: Differences between SigLIP2 g-opt and DINOv3 ViT-H+ on ImageNet-1K, computed as SigLIP2 minus DINOv3 on the full evaluation set. The 95% confidence intervals are obtained from 1,000 paired bootstrap resamples of the shared evaluation set. Higher is better for accuracy, LR+ and AUROC; lower is better for FPR@95 and ECE. The accuracy difference favors DINOv3, whereas all four uncertainty metrics favor SigLIP2.

Quantity	SigLIP2 − DINOv3	95% CI
Top-1 accuracy (pp)	-0.31	$[-0.56,\ -0.07]$
LR+ at 20% retention	$+25.7$	$[+11.3,\ +49.6]$
AUROC	$+0.0084$	$[+0.0044,\ +0.0129]$
FPR@95	-0.038	$[-0.066,\ -0.016]$
ECE	-0.0197	$[-0.022,\ -0.017]$

D.2.1 Headline ImageNet-1K comparison

Table 6 reports the paired-bootstrap differences between SigLIP2 g-opt and DINOv3 ViT-H+ on ImageNet-1K for the primary metrics in our ImageNet-1K comparison. Differences are computed as SigLIP2 minus DINOv3 on the full evaluation set, with 95% confidence intervals obtained from 1,000 paired bootstrap resamples of the shared evaluation set.

Two observations are important. First, the accuracy difference is small at 0.31 percentage points and therefore satisfies our practical accuracy-matching criterion of $|\Delta\mathrm{Acc}| < 0.5$ percentage points. We describe the pair as closely matched on this basis, even though the 95% confidence interval excludes zero. Second, despite DINOv3's slightly higher accuracy, all four uncertainty metrics favor SigLIP2. AUROC differs by 0.0084 and FPR@95 by 0.038, whereas LR+ differs by 25.7 on a scale spanning 24.0 to 53.1 across the 18 backbones. The conventional summaries therefore favor the same model but show much smaller differences than LR+.

D.3 Single-seed scope

Bootstrap intervals quantify uncertainty from the finite evaluation set but do not include variability from probe initialization or stochastic training. Multi-seed probe training across the complete 18-backbone benchmark was computationally prohibitive, and this source of variation should therefore be considered when interpreting small between-model differences.

E Supplementary Analyses

E.1 Metric panels for the remaining evaluation heads

The full accuracy and uncertainty metrics for all tasks and models are shown in Figure 3.

E.2 LR+ retention curves for all evaluation heads

The full LR+ curves for all tasks and models are shown in Figure 4.

E.3 Error-set overlap for DINOv3 and SigLIP2

To test whether the ImageNet-1K LR+ difference is explained by the models failing on fundamentally different examples, we compare the error sets of DINOv3 ViT-H+ and SigLIP2 g-opt. Let

$$E_D = \{i \mid \text{DINOv3 is incorrect on example } i\}$$

and

$$E_S = \{i \mid \text{SigLIP2 is incorrect on example } i\}.$$

We report the fraction of each model's errors shared by the other,

$$\frac{|E_D \cap E_S|}{|E_D|} \quad \text{and} \quad \frac{|E_D \cap E_S|}{|E_S|}.$$

Table 7: MC-dropout error detection with the dropout rate fixed at 0.10. LR+ is evaluated at 20% retention; AUROC summarises the full confidence ranking. Raw values for the deterministic MSP baseline and the three MC uncertainty scores. AUROC remains similar across confidence scores, while LR+ varies more across individual models. However, no MC-dropout score consistently improves median LR+ over deterministic MSP. LR+ $\to 1$ and AUROC $\to 0.5$ indicate no discriminative signal.

	Det. MSP		MC pred. entropy		MC mutual info.		MC variance	
Backbone	LR+	AUROC	LR+	AUROC	LR+	AUROC	LR+	AUROC
ImageNet-1K								
DINOv3 ViT-B	43.1	0.879	42.3	0.880	39.2	0.867	41.2	0.870
DINOv3 ViT-L	24.0	0.879	25.8	0.876	26.3	0.867	26.3	0.870
DINOv3 ViT-H+	24.0	0.877	22.9	0.877	28.4	0.873	27.2	0.875
SigLIP2 ViT-B	53.1	0.879	47.2	0.872	51.9	0.855	47.2	0.845
SigLIP2 ViT-L	48.5	0.880	55.0	0.877	55.0	0.866	61.6	0.868
SigLIP2 g-opt	49.7	0.885	44.9	0.882	48.1	0.869	43.4	0.871
MetaCLIP2 ViT-B	42.2	0.874	37.3	0.868	40.2	0.848	40.2	0.841
MetaCLIP2 ViT-L	34.9	0.875	40.7	0.877	39.7	0.865	39.7	0.866
MetaCLIP2 ViT-bigG	31.4	0.876	28.5	0.878	30.4	0.871	29.7	0.874
VideoMAE v2 ViT-B	38.5	0.865	36.3	0.860	34.0	0.845	35.0	0.803
VideoMAE v2 ViT-L	38.1	0.864	32.4	0.856	32.2	0.841	33.3	0.793
VideoMAE v2 ViT-g	39.5	0.871	35.2	0.865	35.2	0.853	34.8	0.833
InternVideo2 ViT-B	46.3	0.873	48.6	0.865	42.9	0.848	43.8	0.838
InternVideo2 ViT-L	36.5	0.875	42.9	0.874	41.9	0.858	42.9	0.857
InternVideo2 1B	36.1	0.874	38.2	0.876	42.2	0.861	40.1	0.864
V-JEPA 2.1 ViT-B	49.1	0.877	43.6	0.873	41.7	0.856	42.6	0.853
V-JEPA 2.1 ViT-L	39.5	0.877	31.8	0.880	35.0	0.867	33.0	0.869
V-JEPA 2.1 ViT-G	37.3	0.878	36.1	0.880	39.4	0.865	36.1	0.868

On the ImageNet-1K validation set, $|E_D| = 5{,}775$, $|E_S| = 5{,}931$, and $|E_D \cap E_S| = 4{,}063$, giving shared fractions of 70.4% and 68.5% respectively. This analysis distinguishes differences in the difficulty of the residual error set from differences in the confidence assigned to similar failures.

E.4 MC Dropout

Dropout is applied to the attentive probe only: the frozen backbone remains in evaluation mode and its features are computed once per input, while the probe (the cross-attention pooling layer, its attention, and the classifier MLP) is placed in training mode so that dropout remains active at inference. The same probability is used for the classifier, cross-attention, and attention dropout. The probe contains no batch normalization and no stochastic depth, so dropout is the only source of stochasticity. We use $p = 0.10$, fixed for all backbones and matched to the rate at which the probe was trained. We draw $T = 20$ stochastic forward passes, yielding softmax vectors $p_1, \ldots, p_T$, and average them into the predictive distribution $\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p_t$, with prediction $\arg\max_c \bar{p}_c$. Accuracy and calibration are computed from $\bar{p}$. For failure detection and LR+ we use three scores derived from the passes: the predictive entropy $H(\bar{p})$, the mutual information $H(\bar{p}) - \frac{1}{T}\sum_t H(p_t)$, and the variance across passes of the predicted class's probability. For these uncertainty scores, lower values indicate greater certainty when ranking predictions. We report predictive entropy as the primary score. The deterministic baseline is a single forward pass with dropout disabled, using its MSP as the confidence.

Per-backbone values for all four confidence sources on ImageNet-1K are given in Table 7.

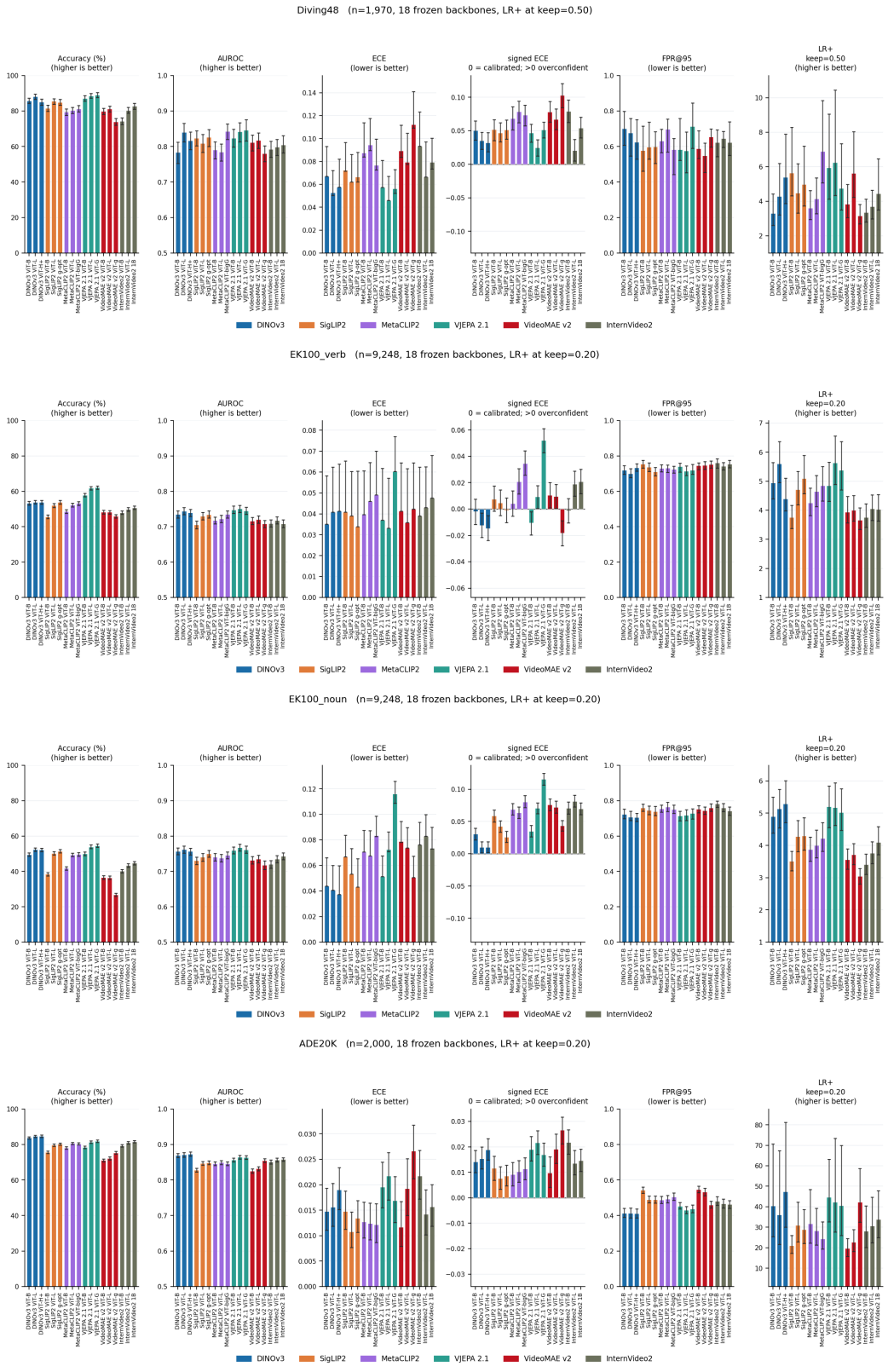


Figure 3: **Accuracy and uncertainty metrics on the remaining evaluation heads.** Rows correspond to Diving48 fine-grained video classification, EPIC-KITCHENS-100 verb anticipation, EPIC-KITCHENS-100 noun anticipation, and ADE20K semantic segmentation. Columns report accuracy, AUROC, ECE, signed ECE, FPR@95, and LR+, respectively. In each panel, the x-axis shows the 18 evaluated models and the y-axis shows the corresponding metric value; error bars indicate 95% bootstrap confidence intervals. LR+ is evaluated among the top 20% most-confident predictions except on Diving48, where the top 50% is used because of the smaller evaluation set.

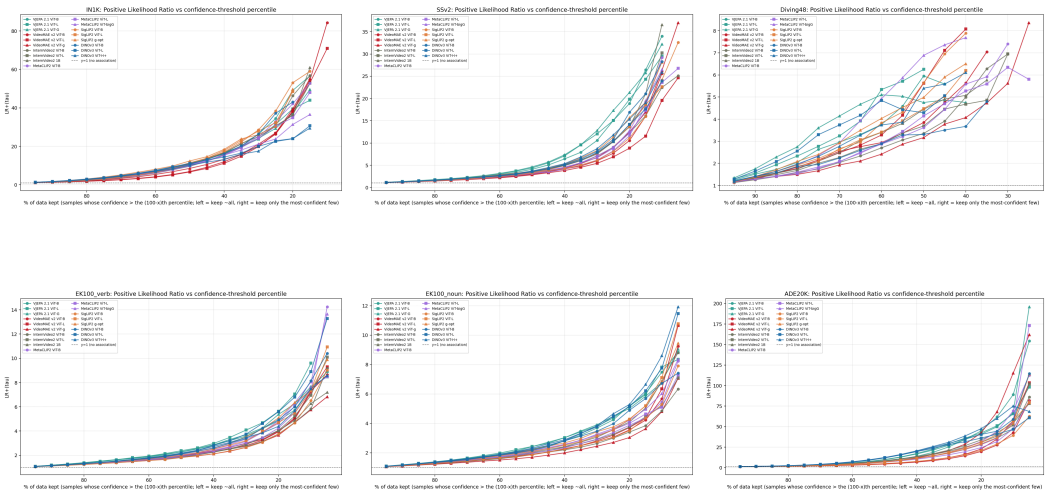


Figure 4: **LR+ varies substantially across retained fractions and downstream tasks.** Each panel shows one task, with separate panels for EPIC-KITCHENS-100 verb and noun anticipation. The x-axis shows the fraction of predictions retained after confidence filtering and the y-axis shows LR+. Colour denotes model family and marker shape denotes scale. The model with the highest LR+ changes across both retained fractions and tasks, showing that confidence trustworthiness depends on both the operating point and the application. All models approach LR+$= 1$ at full retention.