Critical Analysis: VALOR-EVAL: Holistic Coverage and Faithfulness Evaluation of Large Vision-Language Models

Kai-Yu Lu

1 Methodological Strengths

1.1 Multi-dimensional benchmark design

A central methodological strength lies in the design of VALOR-BENCH as a multi-dimensional hallucination benchmark. The benchmark explicitly decomposes hallucinations into three categories:

- object existence (object–object),
- attributes (object–attribute), further split into color and counting for both objects and people,
- relations (object–relation–object), divided into positional and comparative relations.

This decomposition extends prior work that concentrates almost exclusively on object existence. The explicit separation of color and count for attributes, and positional versus comparative relations, enables fine-grained error localization. For example, a caption can be correct about the presence of a "woman" but incorrect about the color of her jacket, which is evaluated in the attribute subset rather than being conflated with object hallucination. This granularity is methodologically sound because it allows different error modes to be measured rather than aggregated into a single hallucination rate.

The benchmark also combines two complementary data sources:

- GQA is used for images with rich structural annotations for objects and relations.
- Pixel (Pexels) images are used to supplement attribute cases, particularly where GQA attribute annotations are sparse.

This design leverages the dense structural annotations of GQA while compensating for its limitations in attribute coverage through curated web images.

1.2 Associative-bias-based image selection

The image selection strategy based on co-occurrence statistics is another strong methodological component. The authors compute co-occurrence frequencies and conditional probabilities between:

- objects and other objects,
- objects and attributes,
- objects and relations.

Conditional probabilities of the form

$$P(\text{feature} \mid \text{object}) = \frac{\text{Frequency(feature, object)}}{\text{Frequency(object)}}$$

are used to identify objects with skewed co-occurrence distributions. For these objects, the benchmark selects:

- features that co-occur most frequently (set H),
- features that co-occur least frequently but still occur (set I).

Images are then chosen where an object from the high-dependency set O appears with a rarely associated feature from I while excluding strongly associated features from H. This construction intentionally breaks the most common co-occurrence patterns seen during training, thus exposing associative bias. The ablation comparing co-occurrence-based selection with random GQA images shows that the selected images reduce faithfulness scores and contain on average more objects per image, confirming that the benchmark is more challenging and structurally richer than a random slice of GQA.

1.3 Human-centric annotation protocol

The benchmark adopts fully manual annotation for the final labels, in contrast to several prior works that rely on object detectors or GPT-4V for pseudo ground truth. The annotation protocol is carefully defined:

- object existence includes foreground and background entities,
- attributes record both color and count for non-human objects, and clothing colors plus number of people for human-centric scenes,
- relations specify positional layout and size ranking.

The use of human annotators and explicit tuple formats for objects, attributes, and relations improves reliability and interpretability. The definitions are consistently applied across subsets, allowing a unified feature representation (I, F_G, p_G) where I is the image, F_G is the feature set, and p_G is the prompt used to elicit LVLM captions.

1.4 LLM-based two-stage evaluation framework

The VALOR-EVAL framework introduces a structured, two-stage evaluation pipeline:

1. **Feature extraction**: an LVLM generates a free-form caption R given (I, p_G), and an LLM is prompted to extract a set of features $F_R = \{f_1^R, \ldots, f_m^R\}$ from this caption for objects, attributes, and relations.
2. **Feature matching**: another set of LLM prompts aligns F_R with ground-truth features F_G, producing:
    - a dictionary of strict matches D_M (semantic equivalence),
    - a dictionary of broader conceptual matches D_B (hypernym-like matches, such as "clothes" versus "dress").

This design addresses two major weaknesses of previous metrics such as CHAIR:

- It avoids dependence on a fixed vocabulary of 80 MSCOCO objects, enabling open-vocabulary evaluation where LVLM outputs may use synonyms or paraphrases.
- It incorporates semantic similarity through LLM judgments rather than exact string matching or manually curated synonym lists.

1.5 Faithfulness and coverage metrics

The paper introduces two complementary metrics:

- **Faithfulness**: defined as
$$\text{Faithfulness}(R, F_G) = \frac{|D_M \cup \text{set}(D_B)|}{|F_R|},$$
which measures the proportion of extracted features that are supported by the ground truth, including broader conceptual matches. Intuitively, this penalizes hallucinated features but gives credit when the caption describes correct information, even if the wording is more general.

- **Coverage**: defined as
$$\text{Coverage}(R, F_G) = \frac{|\text{set}(D_M)|}{|F_G|},$$
which measures how many ground-truth features are explicitly captured in the caption through strict semantic matches. This metric captures informativeness and completeness of the description.

Separating faithfulness and coverage is methodologically important. Previous metrics often implicitly mix the two, which can reward very short but safe captions. Here, a model that suppresses hallucination by omitting information receives high faithfulness but low coverage, making the tradeoff explicit.

1.6 Systematic experimental design and ablations

The experimental section exhibits several strengths:

- The evaluation covers 10 diverse LVLMs, including instruction-tuned open-source models and commercial systems such as GPT-4V and Emu2, which provides a broad view of current capabilities.

- The results are reported per dimension (object existence, attributes-object, attributes-people, positional relations, comparative relations) with both faithfulness and coverage scores, rather than only global averages.

- A correlation study compares VALOR-EVAL scores with human judgements for InstructBLIP outputs. Pearson correlations above 0.9 for most categories demonstrate that the LLM-based metric aligns well with human evaluation in both faithfulness and coverage.

- An ablation study on data selection (co-occurrence-based versus random selection) shows that the proposed selection method significantly lowers faithfulness scores, especially for LLaVA-1.5, confirming that the benchmark better exposes hallucinations than random GQA images.

- Another ablation compares CHAIR with an LLM-augmented variant ($\text{CHAIR}_{\text{LLM}}$) on a subset of COCO images. Accuracy in hallucination detection increases dramatically, illustrating concretely that LLM-based semantic matching improves over fixed-vocabulary metrics.

1.7 Transparency and limitations discussion

The paper explicitly documents ethical considerations and limitations, including the dependence on GPT-4 as the evaluation LLM and the restricted attribute and relation types. Such transparency about the scope of the contribution and potential biases is a methodological strength, since it helps contextualize the results and guides future extensions.

2 Key Limitations

2.1 Benchmark scale and representativeness

A first major limitation is the small size of VALOR-BENCH. The benchmark contains:

- 50 images for object existence,
- 27 object-attribute images,
- 34 people-attribute images,
- 50 positional relation images,
- 50 comparative relation images.

This yields a total of about two hundred images. While the features per image are dense, the dataset is small compared with other vision-language benchmarks. Such a compact benchmark can be highly diagnostic but raises concerns about robustness, statistical reliability, and potential overfitting by future methods that tune specifically to VALOR-BENCH.

In addition, the domain coverage is limited:

- Object existence and relational scenes are inherited from GQA, which itself mainly reflects COCO-style everyday scenes.
- Attribute cases come from curated stock photos on Pixel, which often contain clean and staged visuals rather than realistic complex environments.

This combination leads to a benchmark that is informative for everyday scenes but does not necessarily generalize to other domains such as medical images, documents, or diagrams.

2.2 Restricted attribute and relation semantics

The attribute and relation spaces are deliberately restricted:

- Attributes focus only on color and count.
- Relations include only spatial position and size ranking.

Other important types of attributes, such as material, texture, pose, or fine-grained category, are not covered. Similarly, richer relations such as ownership, interaction, functionality, or temporal progression are absent. As a result, hallucinations that involve higher-level semantics, such as inferring actions or intentions, are not captured by this benchmark, even though such hallucinations are common in LVLM outputs.

2.3 Dependence on a single evaluation LLM

VALOR-EVAL relies on GPT-4 as the sole evaluation agent for both feature extraction and semantic matching. This leads to several limitations:

- Any biases, blind spots, or hallucinations of GPT-4 will propagate into the evaluation. If GPT-4 misinterprets a caption or fails to recognize a concept present in the ground truth, both faithfulness and coverage may be misestimated.

- The correlation study with human judgements is conducted on one model (InstructBLIP) and one benchmark. The degree to which these correlations hold across other LVLMs, different prompts, or other domains remains uncertain.
- The use of GPT-4 introduces practical constraints: evaluation becomes expensive and requires access to a proprietary API, which reduces reproducibility for groups without such access.

2.4 Limited reporting of annotation quality

The paper describes the use of human annotators and detailed guidelines, but does not provide quantitative measures of annotation quality such as inter-annotator agreement. For a benchmark that aims to diagnose subtle hallucinations at the level of attributes and relations, disagreement rates can be non-trivial, especially for positional or comparative relations. Without explicit statistics on agreement and adjudication procedures, it is difficult to assess how stable the ground-truth labels are.

2.5 Metric design assumptions and edge cases

The faithfulness and coverage metrics depend on the dictionaries D_M and D_B produced by the evaluation LLM. This design introduces several subtle limitations:

- Faithfulness counts broader conceptual matches in D_B as correct, which treats statements such as "red clothes" as faithful to "red dress". This is reasonable in many cases, but it also risks over-crediting vague or underspecified descriptions that avoid committing to a specific object type.
- Coverage considers only strict matches in D_M. A caption that correctly mentions a broader category (for example, "vehicle" instead of "car") may be penalized as missing the feature, even if it demonstrates partial understanding.
- The extraction step assumes that the evaluation LLM identifies all relevant features from the caption. If the extractor misses an attribute that the caption does express, coverage can be underestimated, while faithfulness is computed with respect to an incomplete F_R.

Moreover, the comparison with CHAIR uses a very small sample of 20 COCO images and focuses only on object hallucination. The observed improvements in hallucination detection accuracy are promising but not sufficient to fully characterize behaviour across diverse datasets and LVLMs.

2.6 Prompt and task scope limitations

All evaluations are conducted in an open-vocabulary captioning setting with fixed prompts for each subset (object existence, attributes, positional relations, comparative relations). This leads to two limitations:

- The analysis does not explore how sensitive faithfulness and coverage are to prompt wording. Some LVLMs may respond much better to alternative prompts or more structured query formats, which is not reflected in the reported results.
- The framework is evaluated only on generative captioning tasks, although hallucination also appears in discriminative tasks such as visual question answering or grounded reasoning. Extending VALOR-EVAL to those settings would require additional design work that is outside the current scope.

3 Technical Bottlenecks

3.1 Complex multi-step evaluation pipeline

The evaluation pipeline itself contains several stages:

1. LVLM caption generation from (I, p_G).
2. LLM-based feature extraction from captions to form F_R.
3. LLM-based feature matching between F_R and F_G to obtain D_M and D_B.
4. Metric computation for faithfulness and coverage.

Each stage introduces potential error sources and cumulative uncertainty:

- If the LVLM generates overly long or redundant captions, the feature extractor must normalize and deduplicate features reliably.
- If the extractor fails to identify a feature mentioned in the caption, the metrics underestimate coverage and possibly overestimate faithfulness for the remaining features.
- If the matcher fails to align semantically equivalent descriptions or confuses different objects, both metrics can be distorted.

This multi-step dependency constitutes a technical bottleneck: debugging metric behaviour requires inspecting internal LLM outputs, which are not deterministic and can evolve over time as the evaluation model is updated.

3.2 Open-vocabulary matching and reproducibility

The key methodological innovation is open-vocabulary evaluation, but this also introduces reproducibility challenges:

- The matching function between F_R and F_G is defined implicitly through prompts and LLM behaviour. Small changes in prompts, decoding parameters, or LLM versions can change D_M and D_B, even for the same (I, p_G) and caption.
- The paper does not define a formal specification of acceptable semantic matches beyond illustrative examples, which makes it difficult to reimplement the evaluation logic in a non-LLM way.

As a result, VALOR-EVAL currently behaves more like an evaluation service than a fully deterministic metric, which is a technical barrier for long-term benchmarking.

3.3 Limited control over associative bias

The benchmark design carefully selects images that disrupt common co-occurrence patterns, but the underlying training distributions of LVLMs are unknown and heterogeneous. The conditional probabilities used for selection are derived from GQA annotations, not from the models' actual training data. Consequently, the benchmark approximates the associative biases that training may induce but cannot directly control them. This creates a structural bottleneck: the benchmark can reveal susceptibility to certain biases, but it cannot systematically cover the full space of biases present in proprietary datasets.

3.4 Computational and cost constraints

Using GPT-4 for feature extraction and matching for all images and all models introduces non-trivial computational and financial cost. Evaluating 10 LVLMs across multiple subsets requires a substantial number of API calls:

- one call per LVLM output for feature extraction,
- additional calls for feature matching across objects, attributes, and relations.

This cost structure makes large-scale or repeated evaluation (for example, during model development or hyperparameter tuning) challenging, and it limits the feasibility of using VALOR-EVAL as a routine diagnostic tool in many settings.

4 Research Implications

4.1 Understanding current LVLM behaviour

The empirical results provide a nuanced picture of current LVLM capabilities:

- Emu2 achieves the highest average faithfulness (around 75 percent) but the lowest coverage (around 8 percent), illustrating a model that plays safe by describing fewer features but describing them accurately.
- GPT-4V attains the highest average coverage (around 28 percent) but a more moderate faithfulness (around 55 percent), reflecting a model that attempts to mention many details and therefore incurs more hallucinations.
- Models such as LLaVA-1.5 and CogVLM exhibit intermediate behaviour, with more balanced faithfulness and coverage profiles.

These patterns show that hallucination is not a binary property but reflects a tradeoff between precision and informativeness. Different LVLMs implicitly choose different points along this tradeoff, presumably due to differences in training data, objectives, and instruction-tuning strategies.

4.2 Beyond object-level hallucination

By explicitly measuring hallucinations in attributes and relations, the benchmark demonstrates that focusing solely on object existence underestimates the hallucination problem. For example:

- Some models perform reasonably well on object existence but hallucinate clothing colors or counts of people.
- Positional and comparative relations are more challenging, with lower faithfulness and coverage scores, reflecting difficulties in spatial reasoning and size comparison.

This suggests that future research on hallucination mitigation needs to target not only object detection but also attribute grounding and relational reasoning.

4.3 Implications for evaluation methodology

The strong correlation between VALOR-EVAL and human judgements indicates that LLM-based evaluation is a viable path for complex multimodal benchmarks:

- LLMs can serve as semantic bridges between structured annotations and free-form descriptions, reducing the need for handcrafted vocabularies and synonym lists.
- The improvement of $\text{CHAIR}_{\text{LLM}}$ over original CHAIR in hallucination detection and coverage shows that LLM-based matching can adapt better to rich and varied captions.

However, the dependence on a single evaluation LLM also highlights the need for formalizing and standardizing LLM-as-judge protocols, including versioning, prompt definitions, and calibration against human ratings.

4.4 Connections to broader multimodal challenges

The paper's focus on associative bias connects hallucination evaluation to longstanding issues in vision-language reasoning:

- LVLMs tend to exploit dataset shortcuts and co-occurrence patterns, inferring objects or attributes that usually appear together even when they are absent.
- Similar biases appear in visual question answering and grounding tasks, where models often rely on language priors rather than visual evidence.

By explicitly constructing cases that break such co-occurrences, VALOR-BENCH contributes to a broader agenda of evaluating and mitigating shortcut learning in multimodal systems.

5 Potential Research Directions

5.1 Expanding benchmark coverage

Several concrete directions follow from the benchmark design:

- Increase the number of images in each subset to improve statistical reliability and support more fine-grained analysis across model families and training regimes.
- Extend attribute coverage beyond color and count to include pose, material, texture, fine-grained categories, and action-related properties.
- Enrich the relation space with functional, interactional, and temporal relations (for example, "person holding cup", "dog chasing ball", "before/after" temporal relations).
- Incorporate additional domains, such as medical, document, chart, and scientific imagery, to evaluate hallucinations in specialized contexts.

5.2 Multi-LLM and human-calibrated evaluation

To reduce dependence on a single evaluation LLM, future work can:

- Use ensembles of evaluation LLMs to derive consensus feature matches and quantify uncertainty in D_M and D_B.

- Periodically recalibrate VALOR-EVAL scores against human annotations for a held-out set of images and models, updating prompts and decision rules to maintain alignment.
- Explore lightweight student evaluators distilled from GPT-4 judgements, making evaluation cheaper while preserving semantic sensitivity.

5.3 Training objectives informed by faithfulness and coverage

The explicit separation between faithfulness and coverage suggests new training and decoding strategies:

- Design loss functions or reinforcement learning rewards that penalize hallucinated features while rewarding coverage of ground-truth features, using VALOR-EVAL-like signals as supervision.
- Develop decoding strategies that balance risk and informativeness, for example by re-ranking candidate captions according to predicted faithfulness and coverage estimates.
- Incorporate feature-level confidence estimates, encouraging models to refrain from emitting low-confidence features or to express uncertainty explicitly.

5.4 Mitigating associative bias through data and architecture

Given that VALOR-BENCH is constructed around associative bias, another set of directions targets the causes of such bias:

- Introduce counterfactual or debiased training data that breaks common co-occurrences, similar in spirit to the benchmark but used during training.
- Add architectural components that explicitly ground objects, attributes, and relations in visual evidence, for example through differentiable detectors or scene graphs, combined with a language model that operates on grounded representations.
- Use feature-level regularization that penalizes predictions of features that are weakly supported by the visual input, estimated through attention maps or gradient-based saliency.

5.5 Generalizing the framework to other tasks

Finally, the methodological ideas behind VALOR-EVAL can be extended:

- Apply similar feature extraction and matching to visual question answering, reasoning chains, or multimodal dialogue, defining task-specific notions of faithfulness and coverage.
- Explore cross-lingual evaluation where captions and annotations are in different languages, using multilingual LLMs as evaluation agents.
- Integrate temporal information for video-based LVLMs, evaluating hallucinations about events, actions, and temporal ordering.

6 Conclusion

The VALOR-EVAL paper contributes a carefully designed benchmark and an LLM-based evaluation framework that together provide a more holistic view of hallucinations in large vision-language models. Methodological strengths include the multi-dimensional decomposition into objects, attributes, and relations, the associative-bias-driven data selection, human-annotated ground truth, and a two-stage evaluation pipeline that supports open-vocabulary semantic matching and separates faithfulness from coverage.

At the same time, the work exhibits important limitations: the benchmark is small and domain-limited, attributes and relations are restricted to a narrow set, the evaluation depends heavily on a single proprietary LLM, and the metrics themselves inherit non-trivial uncertainty from multi-step feature extraction and matching. These aspects constrain the generality and reproducibility of the results.

Overall, the most promising future directions include expanding and diversifying the benchmark, reducing dependence on a single evaluation LLM through multi-judge or distilled evaluators, and using faithfulness and coverage as explicit supervision signals for training LVLMs that are both reliable and informative. The paper thus serves as both a useful diagnostic tool for current models and a conceptual template for next-generation hallucination evaluation frameworks.