Critical Analysis: HallE-Control: Controlling Object Hallucination in Large Multimodal Models

Kai-Yu Lu

1 Methodological Strengths

1.1 Comprehensive Problem Formulation

The paper offers a clear and fine-grained decomposition of hallucination phenomena in Large Multimodal Models (LMMs), distinguishing object existence hallucination from attribute and relationship hallucinations. The analysis focuses specifically on object existence hallucination in detailed captioning, which is a well-scoped and practically important subproblem. This explicit focus enables precise metric design and controlled experiments rather than conflating different hallucination types in a single aggregate score.

1.2 Evaluation Design: From VQA to Detailed Captioning

A major methodological strength lies in the critique of existing VQA-based hallucination benchmarks and the subsequent proposal of a caption-centric evaluation protocol.

1.2.1 Critical Examination of VQA Benchmarks

The paper systematically contrasts POPE and MME with detailed captioning needs. VQA benchmarks are shown to have structural limitations:

- The majority of questions are binary yes/no queries on object existence or coarse attributes. High performance can be achieved by conservative answering strategies, such as lowering the probability of answering "yes", without demonstrating faithful scene understanding.
- The benchmarks do not enforce long-form outputs or dense captioning. Therefore they do not expose hallucination patterns that arise when a model is pressured to enumerate many objects and relations in a single caption.
- The score is insensitive to caption richness and does not penalize models for omitting relevant objects.

This critical analysis is supported by empirical evidence: Shikra7B achieves superior F1 scores and balanced yes ratios in POPE and higher total scores in MME, yet later underperforms on caption-based hallucination metrics under matched output length.

1.2.2 CCEval: A GPT-4 Assisted Caption Evaluation Protocol

The proposed CCEval protocol is methodologically significant. Its design addresses key weaknesses of prior caption-based metrics such as CHAIR:

- **Length and object-count control.** The protocol enforces approximate parity in average caption length (around 100 words) and average number of mentioned objects (around 9) across models. This eliminates the confounding factor where a model can artificially appear less hallucinatory by producing very short captions with few object mentions.
- **Semantic alignment via GPT-4.** Instead of relying on a hard-coded synonym list, CCEval uses GPT-4 to extract objects from generated captions and align them with Visual Genome ground truth objects. This substantially improves robustness to lexical variation and paraphrases in object naming.
- **Joint consideration of hallucination and coverage.** Beyond CHAIRs and CHAIRi, CCEval introduces an object coverage metric, defined as the fraction of ground truth objects that are successfully mentioned in the caption. This ensures that models are not rewarded for low hallucination simply by omitting many true objects.

Empirical results highlight the value of CCEval. For instance, InstructBLIP7B exhibits CHAIRi of 1.7 and CHAIRs of 1.4 under original CHAIR evaluation, but only because it generates extremely short captions with an average of 2.3 words and 0.8 objects per sentence. Under CCEval, when forced to produce long captions similar to other models, its CCEval CHAIRi rises to 22.3 with coverage around 29.8%, revealing nontrivial hallucination when compared fairly.

1.3 Systematic Ablation and Component-wise Analysis

The methodology includes well-structured ablation studies that isolate the impact of different components.

1.3.1 Language Decoder Scaling

The paper evaluates LLaVA with 7B, 13B and 33B language decoders, and InstructBLIP with 7B and 13B:

- On POPE, larger decoders improve F1 scores and reduce hallucination according to VQA-style metrics.
- On CCEval, LLaVA shows modest reduction of CHAIRi from 25.3 (7B) to 23.8 (13B) and 21.8 (33B), with similar coverage around 31–34%.
- InstructBLIP13B improves CCEval CHAIRi to 16.7 from 22.3 for the 7B variant and raises coverage to about 33.6%.

These results are systematically reported and interpreted, supporting the conclusion that language scaling provides some benefit but does not fundamentally resolve hallucination.

1.3.2 Instruction Data Scale and Quality

The paper compares three instruction finetuning regimes for LLaVA7B: 80K, 158K and the 2.4M-scale SVIT dataset. The analysis is nuanced:

- On CCEval, the 80K model attains CHAIRi of 19.7 with coverage 32.7%, while the 158K model degrades to CHAIRi of 25.3 with similar coverage.
- The SVIT model produces extremely long captions (average length 296.6, average object count 18.1), with higher coverage 47.5% but also higher hallucination and extreme yes ratios on POPE (above 89%).

These experiments demonstrate that enlarging instruction data without attention to alignment and quality can increase hallucination. The paper further inspects GPT-4 generated training captions and attributes the issue not to GPT-4 itself, but to MSCOCO ground truth objects that are visually difficult even for humans to ground.

1.3.3 Vision Encoder Resolution and Sliding Window

The effect of input resolution for CLIP-Large is explored for LLaVA:

- With LLaMA2 13B, increasing resolution from 112 to 224 to 336 reduces CCEval CHAIRi from 21.7 to 19.3 to 16.0, while coverage rises slightly from 32.0% to 33.4%.
- For LLaVA7B, using higher resolution and a sliding window technique yields lower CHAIRi (18.7 at 224 with sliding window vs 21.7 at native 224) and higher coverage (36.9% vs 32.0%).

This structured resolution study concretely supports the hypothesis that fine-grained visual grounding is a key determinant of hallucination.

1.4 HallE-Control: Controllable Hallucination Mechanism

The modeling contribution is methodologically clean and lightweight:

- The controller is a single linear layer W added on top of a frozen backbone B and language head H. The controlled embedding is $e'_v = e_v + \varepsilon W(e_v)$, and the model output is $M'(x) = H(B(x) + \varepsilon W(B(x)))$.
- The scalar parameter ε is trained with contrastive supervision: $\varepsilon = -1$ for purely contextual-only data and $\varepsilon = +1$ for parametric joint data with bracketed objects.
- Only W is finetuned, which makes the approach computationally efficient and easily applicable to existing LLaVA backbones.

The experiments show that for LLaVA7B, setting $\varepsilon = -1$ reduces CCEval CHAIRi from 25.3 (baseline) to 20.9 while maintaining similar coverage and even slightly increasing average caption length. For LLaVA13B at 336 resolution, combining indication and $\varepsilon = -1$ reduces CHAIRi from 16.0 to 6.37 and CHAIRs from 64 to 43 without loss of coverage, indicating a strong effect from a very simple control mechanism.

1.5 Transparent Indication Experiments

Another strength is the explicit analysis of parametric indication using bracketed objects. The three evaluation modes (only indicated, without indicated, with indicated) are carefully defined and analyzed:

- Evaluation only on indicated objects yields CHAIRi above 60 and low coverage, confirming that bracketed objects are indeed more hallucination-prone.
- Evaluation without indicated objects reduces CHAIRi from 25.3 to 17.1 for LLaVA7B and from 16.0 to 11.62 for LLaVA13B, demonstrating that indication successfully isolates high-risk objects.
- Evaluation with indicated objects shows a 44.66% reduction in CHAIRi for LLaVA7B and 38.38% for LLaVA13B compared to their baselines, without coverage degradation.

This provides strong methodological support for the interpretability and utility of the indication mechanism.

2 Key Limitations

2.1 Scope of Hallucination Types

The study explicitly focuses on object existence hallucination and does not provide equally deep analysis for attribute or relationship hallucinations. Although the introduction recognizes these other categories, the experimental framework and HallE-Control mechanism are not tested on them. As a result, the conclusions and proposed controls are scoped to existence hallucination in static images, and it remains unclear whether similar mechanisms would work for attribute consistency or relational correctness.

2.2 Dependence on Specific Datasets and Sampling Choices

The evaluation relies heavily on MSCOCO and Visual Genome:

- MSCOCO provides the captioning and object labels used for instruction tuning and controller training.
- Visual Genome is used as the ground truth source for CCEval on 100 randomly sampled images.

This setting raises several concerns:

- The sample size of 100 images in CCEval is relatively small for a highly stochastic generative setting. Variance across runs or prompts is not reported.
- Both datasets contain biases toward everyday scenes and common objects. The conclusions may not transfer directly to specialized domains such as medical imaging or industrial inspection.
- The paper identifies annotation issues in MSCOCO and visually difficult objects, but does not attempt to quantify annotation noise or correct it systematically before training.

2.3 Reliance on Auxiliary Models

The pipeline relies on several powerful external models:

- GPT-4 for generating contextual-only captions and for evaluating hallucination in CCEval through object extraction and alignment.
- RAM as an open vocabulary detector for separating grounded and omitted objects.

This introduces several limitations:

- GPT-4 and RAM are treated as oracles. Possible errors or biases in their predictions are discussed qualitatively but not measured quantitatively.
- The computational and monetary cost of GPT-4 based evaluation and data generation is not analyzed. This limits the practicality of CCEval and the training pipeline for researchers without access to such resources.
- The use of GPT-4 in both data generation and evaluation introduces a form of model dependency that may affect generalization, although the paper argues that the main issue stems from MSCOCO labels rather than GPT-4 itself.

2.4 Global Control Parameter and Lack of Granularity

HallE-Control uses a single scalar control parameter ε applied uniformly to all tokens and time steps. This design yields a smooth global tradeoff between conservative and imaginative behavior, but also introduces limitations:

- The control cannot differentiate between safety-critical objects and benign background objects. For example, references to pedestrians near a road and decorative items on a shelf are modulated together, although they have different safety implications.
- Region-specific or object-specific hallucination is not directly controlled. The mechanism does not exploit spatial information, despite the vision encoder producing region-level features.
- The same ε is used throughout the entire caption, which does not account for later tokens possibly relying more heavily on language-only co-occurrence than early tokens.

2.5 Limited Analysis of Computational Cost and Efficiency

Although the core controller is lightweight, several aspects of efficiency remain unaddressed:

- Training and inference cost for high-resolution CLIP inputs with sliding windows is not quantified. The combination of high resolution and long captions could result in significant computational overhead.
- The additional linear layer W is small, but the cost of generating and using contextual-only and parametric joint datasets is nontrivial due to GPT-4 calls.
- No latency or throughput measurements are reported, which limits understanding of tradeoffs between hallucination reduction and deployment constraints.

2.6 Narrow Evaluation of Downstream Utility

The paper mentions application domains such as robotics and content moderation but does not experimentally assess downstream performance. The impact of hallucination control on end tasks, such as safe robot manipulation, interactive dialogue or image retrieval, is therefore only indirectly inferred from caption-level metrics.

3 Technical Bottlenecks

3.1 Vision–Language Granularity Misalignment

The analysis identifies a fundamental bottleneck: misalignment between the granularity of vision features and the granularity of textual descriptions.

- Training captions often include objects that are small, heavily occluded or visually ambiguous.
- When the vision encoder cannot reliably ground these objects, the model learns object co-occurrences in the language space rather than true image–text grounding.
- During inference, this parametric knowledge leads to hallucinated objects that are plausible but not actually present.

This bottleneck is structural and arises from the interplay between dataset design, annotation practices and vision encoder capacity.

3.2 Information Bottleneck in the Vision Encoder

Even with higher resolution, the vision encoder ultimately compresses the image into a limited number of tokens. The experiments show that:

- Increasing resolution from 112 to 336 improves hallucination metrics but does not eliminate hallucination.
- Sliding window processing provides gains but increases computational cost.

This suggests that information about small or subtle objects is still partially lost in the representation, creating an inherent bottleneck that encourages parametric guessing for fine-grained details.

3.3 Instruction Data Design and Ground Truth Constraints

Another bottleneck lies in the instruction data generation process:

- Captions are forced to include all ground truth objects from MSCOCO, including those that are visually marginal.
- As the number of such objects increases, the mismatch between visual evidence and textual supervision grows.
- This amplifies parametric correlations and encourages hallucination when models are asked to produce detailed descriptions.

The pipeline is constrained by the structure of existing datasets and the decision to treat all annotated objects as mandatory caption content.

3.4 Controller Expressiveness

The linear controller W with scalar ε is expressive enough to capture a useful hallucination direction, but it still represents a low-capacity intervention:

- The transformation is linear in the embedding space and may not capture more complex disentanglement between contextual and parametric knowledge.
- There is no per-object or per-attribute disentanglement. The control affects all lexemes in a similar way, even though some words may be strongly grounded and others almost entirely parametric.
- The mechanism does not incorporate explicit knowledge of which objects are visually grounded at inference time. It only learns correlations present in the training data.

3.5 Evaluation Bottlenecks

CCEval itself introduces technical bottlenecks:

- GPT-4 based object extraction and matching could struggle with highly complex scenes or rare object categories.
- The evaluation is tied to Visual Genome annotations, which suffer from incompleteness and annotation noise. Objects missing from ground truth may be incorrectly treated as hallucinations.
- Because the evaluation is relatively expensive, the number of images and experimental configurations is limited, constraining statistical power.

4 Research Implications

4.1 Reassessment of VQA-based Safety Claims

One important implication is that good performance on VQA-based hallucination benchmarks does not guarantee safe behavior in detailed captioning:

- Shikra7B is an illustrative case. It achieves strong F1 scores and near-balanced yes ratios on POPE and high total scores on MME, yet exhibits higher CCEval CHAIRi and lower coverage than some LLaVA variants under matched caption length.
- This implies that systems validated only on VQA benchmarks may still hallucinate extensively when deployed in tasks that require long-form descriptions, such as robot scene understanding or detailed reporting.

For the broader field, this suggests that safety and reliability claims must be benchmark-specific and cannot be generalized from VQA to open-ended captioning without additional evidence.

4.2 Importance of Vision Capacity and Data Alignment

The findings emphasize that hallucination is not solely a language modeling problem:

- Increases in language model size offer only moderate gains, whereas improvements in vision resolution and object detectability yield more direct reductions in hallucination.
- Misalignment between what the vision encoder can resolve and what captions demand emerges as a dominant factor. This links hallucination to representation capacity and data curation, not only to decoding or prompting strategies.

This perspective connects hallucination control to broader research on visual grounding, multi-scale perception and dataset design.

4.3 Contextual vs Parametric Knowledge as a Conceptual Framework

The decomposition into contextual and parametric knowledge offers a conceptual tool that can be applied beyond this specific study:

- It generalizes ideas from counterfactual question answering and retrieval-augmented generation to the multimodal setting, by distinguishing knowledge grounded in input context from knowledge stored in parameters.
- It suggests that hallucination in LMMs can be viewed as an over-reliance on parametric knowledge in regions where contextual evidence is weak.
- This view can guide design of architectures and training objectives that explicitly separate and control these two knowledge sources in other tasks such as visual question answering, visual dialogue or grounded reasoning.

4.4 Implications for Safety-critical Applications

The analysis of beneficial versus harmful hallucination in the appendix illustrates that not all hallucinations are equally undesirable:

- In robotics, inferring unobserved but highly probable objects (such as chairs around a table) can be helpful for planning.
- In content moderation, inferring entities that are partially occluded or implied can improve detection of policy-violating content.

HallE-Control demonstrates that it is possible to modulate the extent of imagination, which has implications for safety-critical systems. Future deployment may require dynamic control policies that are stricter for safety-critical entities and more permissive elsewhere.

4.5 Methodological Implications for Evaluation

The introduction of CCEval has methodological implications:

- Evaluation frameworks should explicitly control for caption length and object count to avoid rewarding trivially short outputs.
- Metrics should jointly consider hallucination and coverage, since low hallucination without mentioning critical objects is not acceptable in practice.
- The use of strong language models as annotators opens new evaluation paradigms, but also raises questions about cost, reproducibility and dependence on proprietary systems.

5 Potential Research Directions

5.1 Region-aware and Object-aware Control Mechanisms

Future work can extend HallE-Control in several ways:

- Introduce per-object or per-region control signals that modulate hallucination differently for different spatial locations, possibly leveraging region proposals or segmentation masks from the vision encoder.
- Learn multi-dimensional control vectors instead of a single scalar ε, enabling separate control of existence, attributes and relations.
- Integrate uncertainty estimation from the vision encoder, such that control strength is adapted based on local confidence in visual grounding.

5.2 Improved Data Curation and Ground Truth Design

The analysis suggests new directions for dataset design:

- Filter or downweight objects that are below a minimal size or visibility threshold during caption generation, reducing pressure on the model to memorize fine-grained parametric associations.
- Design annotation protocols where annotators explicitly indicate visibility confidence and occlusion, allowing models to differentiate between confidently grounded objects and speculative mentions.
- Construct new datasets in specialized domains where object visibility and safety importance are jointly annotated, enabling more targeted hallucination control.

5.3 Generalizing to Attribute and Relationship Hallucinations

A natural extension is to apply similar ideas to attribute and relational hallucinations:

- Define contextual and parametric knowledge not only at the level of object existence, but also at the level of attributes (such as color and pose) and relations (such as "on", "holding" or "behind").
- Train controllers that can separately modulate the tendency to infer attributes or relations that are not well supported by visual evidence.
- Extend CCEval to evaluate attribute and relation correctness using Visual Genome predicates and region-level annotations.

5.4 Alternative Evaluation Strategies

To address the cost and opacity of GPT-4 based evaluation, future research can explore:

- Replacement of GPT-4 with strong open-source language models fine-tuned for object extraction and matching, reducing dependence on proprietary systems.
- Hybrid evaluation where automated metrics provide coarse measurements and human annotators perform targeted audits on high-risk or ambiguous cases.
- Larger scale CCEval-style evaluations across more datasets and domains, with statistical analysis of variance across seeds, prompts and decoding strategies.

5.5 Integration with Retrieval and External Knowledge

The contextual versus parametric framework suggests integration with retrieval mechanisms:

- Instead of relying solely on parametric co-occurrence, models could retrieve external knowledge or similar images to justify inferred objects.
- Retrieval-based grounding could be combined with HallE-Control to distinguish visually grounded retrieval evidence from purely parametric inference.
- This may link hallucination control in LMMs to retrieval-augmented generation in text-only models, enabling cross-pollination of methods and theory.

5.6 Task-specific Control Policies

Finally, practical deployment requires task-aware control:

- For safety-critical robotics, default settings may favor conservative behavior (ε closer to -1), with strict filters on ungrounded entities.
- For creative generation or storytelling, more permissive settings (ε closer to $+1$) may be acceptable, provided that hallucinations are clearly indicated.
- Learning policies that adapt ε based on user instructions, task description or downstream reward signals is a promising direction.

6 Conclusion

The paper under review provides a detailed and technically grounded investigation of object existence hallucination in Large Multimodal Models, with two main contributions: a more faithful evaluation protocol for detailed captioning and a lightweight yet effective control mechanism for hallucination. The work demonstrates that strong performance on VQA benchmarks does not guarantee reliable behavior in dense captioning and that misalignment between visual granularity and textual supervision is a central cause of hallucination.

At the methodological level, the introduction of CCEval and the separation of contextual and parametric knowledge offer valuable tools for the community. The HallE-Control architecture shows that a simple linear controller and scalar parameter can significantly reduce hallucination without compromising coverage or descriptive richness, especially when combined with higher vision resolution and parametric indication.

The most critical limitations concern the narrow focus on object existence, dependence on specific datasets and auxiliary models, and the use of a global scalar control parameter. These limitations point directly to promising research directions: region-aware control, improved dataset design, extension to attributes and relations, alternative evaluation strategies and integration with retrieval. Overall, the study advances understanding of hallucination in LMMs and provides a concrete foundation for subsequent work on controllable and reliable multimodal generation.