Technical Summary:
HallE-Control: Controlling Object Hallucination in Large Multimodal Models

Kai-Yu Lu

1 Research Problem and Motivation

1.1 Research Problem

The study investigates object existence hallucination in Large Multimodal Models (LMMs) for detailed image captioning. Object existence hallucination refers to cases where a model describes objects that are not present in the underlying image. The work focuses on detailed captioning scenarios in which LMMs are required to produce long, comprehensive descriptions rather than short answers to visual questions.

The central research problem is to understand, evaluate, and control when and how LMMs hallucinate non-existent objects in detailed captions while preserving coverage of truly present objects and maintaining descriptive richness.

1.2 Motivation and Limitations of Prior Work

Existing LMMs, such as LLaVA, InstructBLIP and Shikra, have achieved strong performance on tasks such as visual question answering, visual dialogue and free-form captioning. However, similar to hallucination in Large Language Models, these LMMs exhibit severe hallucination in detailed image descriptions. In particular, three types of object hallucinations are identified in the paper: object existence hallucination, object attribute hallucination and object relationship hallucination, with the primary focus placed on the first type.

Prior evaluations of object hallucination are dominated by VQA-based benchmarks such as POPE and MME. These benchmarks typically pose binary questions about the existence or attributes of objects. They are convenient and low cost, but they exhibit critical limitations for the detailed captioning setting:

- They require only short answers instead of long, dense captions that list many objects and relations.
- Models can appear robust by adopting conservative answering strategies, such as frequently responding “no”, which does not reflect their behavior when required to describe an entire scene.
- They do not explicitly control for average caption length or the number of mentioned objects, which makes comparisons between models with different verbosity levels unreliable.

Caption-based benchmarks such as CHAIR compute hallucination scores based on object concepts in generated captions relative to a ground truth object set. However, these methods still leave substantial gaps:

- They are sensitive to differences in caption length and object count. A model that generates very short captions with few objects can obtain deceptively low hallucination scores.

- They rely on hard-coded synonym lists to align caption objects with ground truth, which introduces brittleness and does not fully exploit modern language models.

These limitations motivate the development of an evaluation protocol that directly targets hallucination in detailed captions, controls for caption length and object count, and uses stronger language understanding to align generated and ground truth objects. At the same time, the analysis aims to identify which components of LMMs (language decoder size, instruction data, vision encoder resolution) drive hallucination, with the goal of enabling controllable reduction of hallucination rather than eliminating all forms of imaginative inference.

2 Related Work

This study is situated at the intersection of large multimodal modeling, hallucination analysis and evaluation methodologies for vision–language systems.

2.1 Large Multimodal Models

Recent LMMs extend Large Language Models with visual encoders to support tasks such as detailed captioning, visual dialogue and VQA. Representative systems include:

- LLaVA, which uses a CLIP-based vision encoder, a linear projector, and a Vicuna or LLaMA style language backbone. The projector and language model are instruction-tuned on image–text pairs.
- InstructBLIP, which adopts the BLIP-2 architecture with a Q-former bridging a frozen vision encoder and a frozen LLM, and is tuned on a large collection of multimodal instruction datasets.
- Shikra, which shares a similar structure with LLaVA but removes the pretraining stage and adds grounding-oriented tasks during instruction finetuning, thereby improving referential dialogue and object-level grounding.

These models provide the backbone architectures for the analysis and for the proposed HallE-Control controller, which is designed as a lightweight add-on over a frozen LMM backbone.

2.2 Hallucination in Vision–Language Models

Hallucination in natural language generation has been extensively studied in tasks such as machine translation, summarization, data-to-text generation and open-domain dialogue. For LMMs, prior work has primarily examined object hallucination in image captioning and VQA. For example, POPE evaluates object existence hallucination by polling models with yes-or-no questions over sampled objects from images, while MME introduces a set of coarse-grained recognition questions about object existence, counting, position and color.

Other work such as CHAIR proposes caption-based hallucination metrics by extracting object nouns from captions and comparing them with ground truth object sets derived from datasets such as MSCOCO. These approaches demonstrate that object hallucination is widespread, but they either focus on VQA-style evaluation or do not adequately control for caption verbosity and object coverage.

The present study extends this line of work by identifying misalignment between what the vision encoder can visually ground and what the instruction data forces the model to mention as a key driver of hallucination. In addition, the study introduces a controllable mechanism that separates contextual knowledge and parametric knowledge, enabling explicit modulation and indication of imagined objects.

2.3 Evaluation of LMMs

Evaluation of LMMs is challenging due to the diversity of tasks and outputs. VQA benchmarks are cheap and scalable but do not reflect richness and coverage in open-ended captioning. N-gram matching metrics such as ROUGE and CIDEr measure textual similarity but are sensitive to ground truth caption length and do not explicitly capture hallucinated objects. CHAIR introduces object-level hallucination metrics, but it still suffers from hard-coded synonym lists and does not provide joint control over coverage, sentence length and object count.

The proposed CCEval framework addresses these limitations by:

- Using GPT-4 to extract objects from captions and align them with detailed ground truth annotations from Visual Genome.
- Measuring hallucination by CHAIR-style metrics together with object coverage, average caption length and average object count.
- Enforcing comparable caption length and object count across models for fair comparison.

3 Dataset Construction

3.1 Overview

The study employs several datasets for evaluation, analysis and controller training. Table 1 summarizes the key datasets, their roles and main characteristics.

3.2 Contextual and Parametric Joint Data

For training the HallE-Control controller, the study constructs two complementary datasets from MSCOCO:

Grouping via RAM The Recognize Anything Model (RAM) is used as an open vocabulary detector. Ground truth object labels from MSCOCO are passed to RAM, which classifies each object into:

- A grounded group, containing objects that RAM can detect with sufficient confidence. These objects are assumed to be visually grounded by a typical vision encoder.
- An omitted group, containing objects that RAM fails to detect. These are likely small, heavily occluded or visually ambiguous, and thus difficult to ground visually.

Contextual-only data For the contextual-only dataset, only grounded objects are used. Ground truth labels, bounding boxes and short regional captions from MSCOCO are provided to GPT-4, which generates detailed captions that mention only grounded objects. These captions represent pure contextual knowledge that is supported by visual evidence.

Parametric joint data For the parametric joint dataset, the original detailed captions from LLaVA are used. Objects belonging to the omitted group that appear in these captions are treated as parametric objects. They are enclosed in special brackets to mark them as inferred or imagined.

Formally, let S denote an original image caption and let $X = \{x_1, \dots, x_n\}$ be the set of omitted objects that occur in S. The processed caption S_{new} is defined by

$$S_{\text{new}} = \text{replace}(S, x_i, [x_i]) \quad \text{for all } x_i \in X. \tag{1}$$

In this expression, replace$(\cdot)$ denotes replacement of each occurrence of the omitted object name x_i in S by the bracketed form $[x_i]$. The symbol S_{new} is the resulting caption that explicitly marks parametric objects. This transformation has two roles: it provides supervision during training that distinguishes contextual and parametric mentions, and it enables explicit indication of inferred objects during inference.

4 Query Protocol and Task Definitions

4.1 Object Existence Hallucination

Object existence hallucination is defined as the phenomenon in which a caption explicitly refers to an object that is not present in the underlying image according to ground truth annotations. This definition is applied both in VQA settings, where answers to existence questions can be judged as hallucinated or not, and in captioning settings, where each object mentioned in a caption is checked against a ground truth object list.

The study distinguishes this phenomenon from attribute hallucination, which mischaracterizes properties such as color or size of an existing object, and relational hallucination, which describes incorrect spatial or interaction relationships.

4.2 VQA-based Evaluation Protocols

POPE POPE evaluates object existence hallucination by posing yes-or-no questions about sampled objects. Three subsets are defined:

- Random subset, where objects are sampled uniformly.
- Popular subset, focusing on frequently appearing objects.
- Adversarial subset, focusing on objects that co-occur frequently in training data and are thus prone to co-occurrence based hallucinations.

For each subset, the study reports standard classification metrics:

- Accuracy, measuring the fraction of correctly answered questions.
- Precision, measuring the fraction of positive answers that are correct.
- Recall, measuring the fraction of truly positive instances that are correctly identified.
- F1 score, the harmonic mean of precision and recall.
- Yes ratio, the percentage of questions that receive a positive answer, reflecting the tendency to over-predict object presence.

MME MME is another VQA-based benchmark that constructs yes-or-no questions for object existence, counting, spatial position and color. Each of 30 images is paired with two questions, one positive and one negative, and the benchmark aggregates scores across categories such as Existence, Count, Position and Color.

4.3 Caption-based Evaluation: CHAIR and CCEval

CHAIR metrics CHAIR evaluates object hallucination in captions by comparing mentioned object nouns with a ground truth object set augmented by hard-coded synonyms. Two metrics are defined. Let the set of hallucinated object mentions be the set of objects that appear in the caption but not in the ground truth set. Let the set of all mentioned objects be the set of all object nouns extracted from the caption. Let the set of hallucinated sentences denote captions that contain at least one hallucinated object. Let the total number of sentences be the number of evaluated captions. Then CHAIRi and CHAIRs are given by

$$\text{CHAIRi} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|} \tag{2}$$

and

$$\text{CHAIRs} = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}. \tag{3}$$

In these formulas, CHAIRi measures the proportion of object mentions that are hallucinated, and CHAIRs measures the proportion of captions that contain at least one hallucinated object. Both are lower when hallucination is less frequent.

Limitations of CHAIR The original CHAIR protocol is sensitive to caption length and the number of mentioned objects. For example, a model that generates very short captions mentioning almost no objects can obtain low CHAIR scores despite providing little useful information. Furthermore, reliance on hard-coded synonym expansions limits robustness and coverage.

CCEval: GPT-4 assisted caption evaluation To evaluate detailed captioning in a more controlled and comprehensive way, the study introduces CCEval, a GPT-4 assisted evaluation protocol with the following steps:

1. Randomly sample 100 images from Visual Genome, which provides dense object annotations per image.
2. Prompt each LMM under evaluation to generate a detailed caption for each image, encouraging long and descriptive outputs.
3. Provide GPT-4 with the ground truth object list from Visual Genome and the generated caption. GPT-4 is used to extract all object mentions from the caption and align them with the ground truth list, resolving synonyms and linguistic variation.
4. Use the aligned objects to compute CHAIRi and CHAIRs as in Equations (2) and (3).
5. Introduce an additional coverage metric, defined as the ratio between the number of ground truth objects that are mentioned in the caption and the total number of ground truth objects for the image. This coverage metric measures how many true objects are actually described.
6. Record and approximately balance average caption length and average number of objects mentioned across models, so that comparisons of hallucination take place under comparable verbosity and object density.

In CCEval, lower CHAIR scores indicate fewer hallucinated objects, while higher coverage indicates more complete descriptions of true objects. Average sentence length and average object count are used to ensure that models are evaluated under similarly demanding captioning conditions.

5 Modeling Approach

5.1 Base Architectures

The analysis and the proposed controller are built on top of existing LMM architectures:

- LLaVA variants that combine CLIP-based vision encoders with Vicuna or LLaMA language backbones of 7B, 13B and 33B parameters. LLaVA uses a linear projector to map visual features into the language model token space and employs a two-stage training process, including caption pretraining followed by instruction tuning.
- InstructBLIP variants that rely on a Q-former to bridge a frozen vision encoder and a frozen language model, and are instruction-tuned across multiple datasets.
- Shikra, which adopts a LLaVA-like structure but introduces additional grounding tasks during fine-tuning and discards the initial caption pretraining stage.

HallE-Control is implemented as a lightweight linear controller on top of LLaVA backbones, leaving both the vision encoder and the language backbone frozen.

5.2 Contextual and Parametric Knowledge

The modeling framework distinguishes two types of knowledge within LMMs:

- Contextual knowledge corresponds to associations between words and visually grounded features that the vision encoder can reliably detect. For example, when the vision encoder strongly activates for a bus region and the caption mentions “bus”, this association is contextual.
- Parametric knowledge corresponds to associations encoded purely in language model parameters, such as word–word co-occurrence learned from textual data. For instance, a model may infer “people” in a street scene even if no people are visible because such co-occurrences are frequent in training data.

The analysis in the paper shows that when training captions mention objects at a finer granularity than what the vision encoder can perceive, the model is forced to rely on parametric knowledge, creating a misalignment between vision and language. During inference, this misalignment manifests as object existence hallucination when the model attempts to “guess” objects that the vision module does not support.

5.3 Control Layer Design

Let M denote a frozen language model, composed of a backbone B that produces contextualized token embeddings and a language model head H that maps embeddings to output token distributions. For an input sequence x, the base model without control is represented as

$$M(x) = H(e_v), \quad e_v = B(x),$$

where e_v is the sequence of output token embeddings from the backbone.

HallE-Control introduces a learnable linear transformation W and a scalar control parameter ε that modulate the word embeddings before they are passed to the LM head. The modified embedding e'_v is given by

$$e'_v = e_v + \varepsilon W(e_v) \tag{4}$$

and the corresponding controlled model M' is defined as

$$M'(x) = H\big(B(x) + \varepsilon W(B(x))\big). \tag{5}$$

In these expressions, $W(\cdot)$ is a linear layer applied to the backbone embeddings and ε is a scalar that controls the influence of W. Intuitively, W learns a direction in embedding space that increases or decreases the tendency to invoke parametric knowledge related to object co-occurrences. When ε is positive, the model is encouraged to activate this direction and thus is more likely to imagine additional objects based on parametric knowledge. When ε is negative, the model is encouraged to suppress this direction and rely more strictly on contextual knowledge grounded in the visual input.

Only the linear layer W is finetuned; both the backbone B and the head H remain frozen. This design ensures that the controller is lightweight and does not require retraining the large base model.

5.4 Training Strategy

Training for HallE-Control uses the contextual-only and parametric joint datasets described earlier:

- For contextual-only samples, ε is set to -1 during training, encouraging the model to generate captions that rely solely on grounded objects and minimize parametric imagination.
- For parametric joint samples, ε is set to $+1$ and the bracketed parametric objects are kept in the target captions. This encourages the model to reproduce both contextual and parametric objects and to encode them into the transformation learned by W.

By training the controller with contrasting values of ε on these two types of data, the model learns to map the scalar control parameter to different regimes of reliance on parametric knowledge.

During inference, ε can take any value in the interval $[-1, 1]$:

- $\varepsilon = -1$ enforces strongly conservative behavior, minimizing object existence hallucination and emphasizing contextual knowledge.
- $\varepsilon = 0$ approximates the behavior of the original model without control.
- $\varepsilon = 1$ maximizes the use of parametric knowledge and encourages imaginative additions, which can be useful for certain applications, while still marking inferred objects when trained with indication.

6 Empirical Results

6.1 Experimental Setup

The empirical study covers:

- Multiple LLaVA variants (7B, 13B, 33B) with CLIP-Large vision encoders at different input resolutions.
- InstructBLIP variants with 7B and 13B language backbones.
- Shikra7B as a strong baseline with grounding-focused instruction data.
- HallE-Control variants built on LLaVA7B and LLaVA13B, with and without explicit parametric indication.

Evaluations are conducted on POPE and MME for VQA-based benchmarks and on CHAIR and CCEval for caption-based evaluation, under configurations that approximately balance average caption length and average number of objects across models.

6.2 VQA-based Benchmarks

On POPE and MME, Shikra7B achieves the strongest performance among the compared models. For instance, on POPE popular:

- LLaVA7B reaches an F1 score of approximately 72.8 with a Yes ratio above 78%.
- LLaVA13B improves to an F1 score around 75.8 with a Yes ratio near 70.6%.
- Shikra7B attains an F1 score of approximately 83.5 with a Yes ratio close to 45.1%, which is much closer to a balanced 50% and suggests a better calibrated tendency to answer "yes".

In MME, Shikra7B also obtains the highest total score, with strong performance across existence, count, position and color categories. LLaVA7B performs worst, while LLaVA13B and InstructBLIP7B lie in between.

These results show that on VQA-based benchmarks, larger models and grounding-oriented models appear to hallucinate less according to object existence questions.

6.3 Caption-based Benchmarks: CHAIR and CCEval

Table 2 summarizes selected results comparing CHAIR and CCEval across models.

In the original CHAIR evaluation, InstructBLIP7B shows extremely low hallucination scores (CHAIRi $= 1.7$ and CHAIRs $= 1.4$), but it also generates extremely short captions averaging 2.3 words and 0.8 objects per sentence. By contrast, LLaVA13B and Shikra7B produce much longer captions with around 90 words and approximately 7.5 objects per sentence, which naturally leads to higher CHAIR scores.

Under CCEval, in which average caption length is controlled to be approximately 100 words and the average number of objects is approximately 9 across models, all models exhibit substantial hallucination:

- LLaVA7B attains CCEval CHAIRi of about 19.7 with coverage around 32.7%.
- LLaVA13B and Shikra7B reach CCEval CHAIRi values above 23.8 and 24.4, respectively, with coverage around 33%.
- InstructBLIP7B, despite its good CHAIR scores for very short captions, exhibits CCEval CHAIRi near 22.3 and slightly lower coverage around 29.8% when forced to produce equally detailed captions.

These results demonstrate that VQA-based and naive caption-based metrics can significantly underestimate hallucination in truly detailed captioning. When caption length and object count are controlled, all evaluated models remain susceptible to object existence hallucination.

6.4 Effect of Language Decoder Size

Scaling the language decoder from 7B to 33B in LLaVA and from 7B to 13B in InstructBLIP produces mixed effects:

- For LLaVA, increasing the decoder size from 7B to 13B and 33B improves POPE F1 scores and reduces CHAIRi on CCEval slightly. For example, CCEval CHAIRi for LLaVA decreases from 25.3 for 7B to 23.8 for 13B and 21.8 for 33B, while coverage remains around 31 to 34%.
- For InstructBLIP, moving from 7B to 13B reduces CCEval CHAIRi from 22.3 to 16.7 and CHAIRs from 72 to 64, with coverage increasing to approximately 33.6%.

Although there are gains from enlarging the language backbone, the improvements in hallucination metrics are not dramatic and are not fully consistent across models, suggesting that language decoder size alone is not the primary factor in controlling hallucination.

6.5 Effect of Instruction Data Scale and Quality

Instruction data scale is examined for LLaVA7B using three datasets: 80K samples, 158K samples and the SVIT dataset with approximately 2.4M samples. Under CCEval:

- The 80K model achieves CCEval CHAIRi of 19.7 and CHAIRs of 72, with coverage around 32.7%.
- The 158K model exhibits increased hallucination with CHAIRi $= 25.3$ and CHAIRs $= 82$, and coverage around 33.6%.
- The SVIT model shows CHAIRi $= 23.3$ and CHAIRs $= 87$, with substantially higher coverage 47.5%, but also excessively long captions averaging nearly 297 words and more than 18 objects per sentence.

On POPE, the SVIT model develops an extreme tendency to answer "yes", with Yes ratios approaching or exceeding 90% in several subsets, leading to degraded accuracy and F1 scores.

Inspection of the instruction data reveals that GPT-4 itself does not introduce additional object existence hallucinations when conditioned on accurate ground truth object lists. Instead, the primary issue arises from MSCOCO ground truth objects that are very hard to ground visually due to low resolution, occlusion or annotation noise. This mismatch forces the model to learn parametric associations for these objects, which increases hallucination during inference.

6.6 Effect of Vision Encoder Resolution

The role of the vision encoder is studied by adjusting the input resolution for CLIP-Large in LLaVA. For example, with a LLaMA2 13B language decoder:

- At an effective resolution of 112 pixels, CCEval CHAIRi is approximately 21.7 with coverage around 32.0%.
- At 224 pixels, CCEval CHAIRi drops to about 19.3 and coverage increases to 32.8%.
- At 336 pixels, CCEval CHAIRi further decreases to 16.0 with coverage around 33.4%.

Additionally, using a sliding window technique to process images at high effective resolutions further reduces hallucination. For LLaVA7B, applying sliding windows at a nominal 224 pixel resolution produces CCEval CHAIRi values below 19 with coverage above 36%.

These results support the conclusion that improving the visual granularity captured by the vision encoder significantly reduces object existence hallucination by enabling more reliable grounding of fine-grained objects.

6.7 Upper Bound Experiments on Indication

Before introducing the control parameter, the study directly finetunes LLaVA on parametric joint data that explicitly marks parametric objects with brackets. This enables evaluation under three settings:

- Evaluation only on indicated objects: CCEval is computed solely on bracketed objects. Here CHAIRi reaches approximately 63.9 for LLaVA7B and 62.3 for LLaVA13B, indicating that bracketed objects are indeed heavily hallucinated, with low coverage.
- Evaluation without indicated objects: bracketed objects are ignored. For LLaVA7B, CHAIRi decreases from 25.3 to 17.1 and CHAIRs from 82 to 57, indicating large reductions in hallucination when parametric objects are filtered out.

- Evaluation with indicated objects: all objects are considered, but indication is maintained. For LLaVA7B, CHAIRi improves from 25.3 to 14.0 under comparable coverage, representing approximately 44.7% reduction in hallucinated object proportion. For LLaVA13B, CHAIRi improves from 16.0 to 9.86, a reduction of about 38.4%.

These experiments show that models can learn to mark parametric objects reliably and that ignoring or downweighting indicated objects can substantially reduce hallucination without sacrificing coverage or length.

6.8 HallE-Control Results

Finally, HallE-Control is evaluated under different control values ε on top of LLaVA7B and LLaVA13B with high resolution vision encoders.

For HallE-Control7B:

- At $\varepsilon = 1$, CCEval CHAIRi is approximately 26.6 with CHAIRs around 89 and coverage near 32.8%.
- At $\varepsilon = 0.5$, CCEval CHAIRi is about 27.9 with CHAIRs around 85.
- At $\varepsilon = -0.5$, CCEval CHAIRi decreases to roughly 24.9 with CHAIRs near 81 and coverage around 35.9%.
- At $\varepsilon = -1$, CCEval CHAIRi further drops to 20.9 with CHAIRs $= 76$ and coverage around 33.9%, while captions remain long (average length around 134 words) with about 8 objects per sentence.

For HallE-Control13B with indication, using a 336 pixel resolution CLIP-L vision encoder:

- The baseline LLaVA13B at this resolution yields CCEval CHAIRi around 16.0 and CHAIRs around 64.
- HallE-Control13B at $\varepsilon = -1$ achieves CCEval CHAIRi approximately 6.37 and CHAIRs around 43, while maintaining coverage above 34% and long captions with nearly 8.8 objects per sentence.

These results show that the control parameter ε provides a smooth tradeoff between hallucination and parametric imagination. Negative values substantially reduce hallucination while preserving object coverage and detailed description, especially when combined with higher vision resolution and indication.

7 Summary

7.1 Key Findings

This study offers a detailed technical analysis of object existence hallucination in LMMs and proposes a controllable framework for mitigating it. The main findings are:

- VQA-based benchmarks such as POPE and MME are insufficient to characterize hallucination in detailed captioning. When evaluated with CCEval under comparable caption length and object count, all examined LMMs exhibit substantial object existence hallucination.
- Scaling up language decoder size yields moderate improvements but does not fundamentally resolve hallucination. In contrast, instruction data scale without quality control can increase hallucination, especially when training captions reference objects that are difficult for the vision encoder to ground.

- Increasing the input resolution of the vision encoder and using techniques such as sliding windows significantly reduce hallucination by improving visual grounding of small and occluded objects.
- Separating contextual and parametric knowledge and training a lightweight linear controller with a scalar control parameter enables explicit modulation of parametric imagination. HallE-Control reduces object existence hallucination by up to approximately 44% relative to the LLaVA7B baseline, while maintaining object coverage and detailed caption length.

7.2 Innovations Compared to Prior Work

Relative to prior work on hallucination and LMM evaluation, the contributions of this study can be summarized as follows:

- Introduction of CCEval, a GPT-4 assisted evaluation method that combines CHAIR-style hallucination metrics with object coverage, average caption length and average object count under controlled conditions for detailed captioning.
- A component-wise analysis that identifies misalignment between vision encoders and instruction data, especially for fine-grained objects that are hard to ground, as a key driver of object existence hallucination.
- Proposal of HallE-Control, a controllable LMM that employs a single linear controller and a continuous scalar parameter to regulate the expression of parametric knowledge, and to highlight inferred objects via explicit markers.

7.3 Limitations and Future Directions

The work focuses primarily on object existence hallucination in image captioning, while attribute and relationship hallucinations are not treated in equal depth. CCEval and the controller training pipelines rely on auxiliary models such as GPT-4 and RAM, which may introduce their own biases and require nontrivial computational resources. In addition, the current control mechanism is global, governed by a single scalar parameter, and does not differentiate between safety-critical objects and less important background details.

Future research directions include extending the control mechanism to attribute and relational hallucinations, designing more fine-grained and region-aware controllers, reducing dependence on large external models for evaluation and data construction, and exploring task-specific control policies that treat safety-relevant objects more strictly while allowing benign imagination in noncritical regions.

Dataset	Scale	Type
MSCOCO	Large scale, 80 object classes	Image captioning with object annotations
Visual Genome	Large scale, densely annotated	Images with dense region, object and relation annotations

Model	CHAIRs ↓	CHAIRi ↓	AvgLen ↑	AvgObj ↑	CCEval CHAIRi ↓	Coverage ↑
LLaVA7B	24.1	9.1	42.5	3.7	19.7	32.7
LLaVA13B	60.6	18.4	90.2	7.6	23.8	33.6
Shikra7B	59.1	16.6	91.2	7.5	24.4	33.3
InstructBLIP7B	1.4	1.7	2.3	0.8	22.3	29.8

Table 2: Selected caption-based results. CHAIR values are from the original CHAIR protocol, while CCEval values are computed under balanced average caption length and average object count across models.