Anonymous User

(Thanks for the review! The corresponding update in the paper is shown in **brown** text.)

Q1: imitation narration of elvis that is only for neurosymbolic

A1: Thanks for the correction. We agree that the ELVIS is not only limited to the neuro-symbolic models, but also for VLMs and purely neuro models. The neuro-symbolic is only the motivation for this work. We have revised all relevant statements in the introduction

Q2: Citation of the bongard problem:

A2: We will add the citation of the bongard problem. Unlike the Bongard problem, ELVIS wants to test the ability of the models that can grasp the important relationships from multiple object scenes effectively. The tasks themselves are mostly easy for humans but hard for the AI models so far. Bongard problem is not easy even for humans and the problems are not dedicated to the grouping of the objects.

Q3: Task is limited to binary classification. Other tasks such as missing part completion for the pattern.

A3: We agree that the conceptual richness of Gestalt reasoning extends beyond simple binary labels. Our design intentionally mirrors ILP problems and extended based on CLEVR, kandinsky patterns, where simple outputs mask substantial underlying structure. The binary interface ensures consistent evaluation across neural networks, symbolic systems, and foundation models, while isolating the core challenge: inferring the latent Gestalt relation that separates the objects. As shown in our experiments, even extremely large VLMs struggle.

To clarify this rationale, we have rewritten the Task Types section to explain that (i) binary labels serve as a minimal but controlled interface, (ii) solving the tasks requires nontrivial relational and grouping reasoning, and (iii) ELVIS naturally supports richer formats such as pattern completion and generative reasoning, which we highlight as promising extensions for future work.

Q4: How was the LLM-based model trained?

A4: We add the training detail provided in the revised version.
For ViT model, it is trained independently for each task using the training images. And tested on the test images.
For the VLMs, each task is evaluated independently using a small supervised demonstration protocol, the model sees 3 positive and 3 negative examples as well as their labels. Along with the images, the model also receives a prompt, including the related gestalt principle of the task, the logic properties that should be focused on. Then the VLMs are asked to reason the logic patterns based on the training examples. In the test session, the VLMs have to label 6 test images which follow the same logic patterns as the training examples.

Remark 1: We moved the resolution experiment to the appendix.

Lia Morra

(Thanks for the review! The corresponding update in the paper is shown in **orange** text.)

Q1: The contribution other than the task number and more baseline models

A1: This version extends the conference version by including more domain specific analysis. We systematic analysis the factors that can impact the AI models performance. Comparison between different group size, group number, object properties and visualize the results.
We also add two more baseline models for the evaluation, i.e. internVL3-78B and GPT-5.

Q2: Less Categories in the extended version?

A2: There was an update for the category in the extended version. To balance the number of categories in each principle. The a_splines and u_splines are essentially the intersection of two splines, thus we merge it together as intersected splines. The big_and_small limit the variation of the size changing, which is not included in the extension version. Besides, we add two new categories, radial symmetry and no touching splines for a wider cover of the task scenarios.

Q3: Training details of the LLM models

A3: We add more evaluation details of the baseline models in the revised version. We also add the LLM prompts into the appendix.
For each task, LLMs receive a prompt at first, which includes 3 positive and 3 negative images and their labels, the prompt also tells which gestalt principle is the main motivation for generating such patterns. Then the LLMs are asked to reason the logic rules that can perfectly distinguish the positive and negative patterns.
In the evaluation session, the LLMs receives 6 images, and label them as positive and negative. The F1 score and accuracy are evaluated based on it.

Q4: Synthetic dataset is too simple?

A4: The dataset is based on synthetic data, but it is not the weakness. Firstly, having synthetic data has a better control for the ground-truth generation. For the grouping labels, it is hard to acquire in the real-world image. Secondly, even the synthetic data is still hard for the state-of-the-art AI models to solve. And synthetic data is easy to scale. For example we can keep the changing of the group numbers but keep other parameters the same.

Q5: Why ViT performance drop when training size increases?

A6: The description in the submitted version was inaccurate. The hundred-shot ViT model does not “over-fit”; rather, it collapses to nearly constant predictions, yielding accuracy around 0.5 and F1 values close to zero. We will correct the text to reflect this behaviour more accurately.

Q6: Comparison with other dataset

A6:
See section Comparison with Existing Datasets (Page 8) in revised version.

Q7: Computation cost for all the models

A7:
See Computation Cost and Hardware Requirement (Page 17) in revised version.

Alessandro Oltramari

(Thanks for the review! The corresponding update in the paper is shown in **blue** text.)

Q1: what factors account for inconsistent performance across different gestalt principles?

A1: The inconsistent performance across different Gestalt principles arises from the varying perceptual and combinatorial demands imposed by each principle. For GPT-5, symmetry-based tasks are particularly challenging. Unlike principles such as proximity or closure, symmetry does not rely on a fixed reference frame: the symmetry axis can be horizontal, vertical, or diagonal in either direction, substantially increasing the hypothesis space. Moreover, symmetry constraints may apply selectively across attributes—e.g., position and color may be symmetric while shape is not, or position and shape may align while color does not. Correctly solving such tasks therefore requires the model to first identify candidate symmetric object pairs and then verify multi-attribute consistency under abstract geometric transformations, a process that is highly sensitive to localization and pairing errors.

In contrast, closure-based tasks are relatively easier for most models. Closure relies on the emergence of a global, stable shape formed by multiple objects, which provides a strong and consistent perceptual cue. For example, in the “big circle” category, although the absolute position, scale, or orientation of the configuration may vary across tasks, the resulting global shape remains a circle. This shape-level regularity substantially constrains the solution space and allows models to rely on holistic pattern recognition rather than exhaustive relational comparisons. Compared to principles such as similarity—where group positions can be arbitrary and where grouping depends on subtle attribute comparisons—closure offers a more deterministic and invariant structure, leading to more robust performance across architectures.

Q2: Are these failures systematic across architectures or model-specific?

A2:
The failures are largely systematic across architectures and reflect intrinsic perceptual ambiguity of certain Gestalt principles (e.g., symmetry), rather than idiosyncrasies of individual models. See Table 4, Figure 10.

Q3: Differentiate between perceptual vs semantic errors?

A3: Conceptually, Gestalt reasoning can be decomposed into two stages: perceptual grouping and subsequent rule-based reasoning. Errors may arise from either stage. In this work, we focus on end-to-end evaluation; isolating these components via oracle grouping or controlled ablations is an interesting direction for future work.