Responses to reviewer Lia

Unclear demonstration of the Figure 19

Each task in the benchmark contains three positive and three negative examples. In the figures shown in the paper (e.g., Figure 19), only two examples per class are visualized for space reasons. We replaced one of the negative examples with another one (randomly positioned) to clarify the task setup.

In our pattern generation logic, we have considered the mitigation of the shortcut solutions. The negative examples are constructed in two complementary ways.
First, at least one negative example explicitly breaks the target Gestalt principle while preserving the object-level logical properties (e.g., shape, color, and size) observed in the positive examples. In this case, objects follow the same attribute patterns as the positives but are arranged randomly so that the Gestalt relation no longer holds.
Second, other negative examples preserve the Gestalt structure but violate one or more object-level attribute rules (e.g., shape or color consistency).
In other words, the dataset explicitly includes both counterexamples of the form

(Gestalt=false, attributes=true) and
(Gestalt=true, attributes=false).

Through this design, the presence of the Gestalt relation cannot alone determine the label, and attribute-level rules cannot alone determine the label either. Therefore, solving the task requires jointly considering both the Gestalt organization and the object-level properties.

We note that while this construction significantly reduces simple shortcut solutions, some residual shortcuts among object-level properties may still exist in principle. However, these do not eliminate the need for detecting the Gestalt principle itself.

We have updated this explanation in the main text. (see page 7~8)

What would happen if the prompt simply included positive and negative examples.

This is an interesting question.

Firstly, we clarify that providing the Gestalt principle in the prompt is our default evaluation protocol because ELVIS is designed as a principle-conditioned benchmark. Each Gestalt principle corresponds to a qualitatively different perceptual mechanism (e.g. proximity, similarity, continuity), and tasks across principles differ substantially in their underlying structure. Proving the principle therefore serves as part of the task specification., rather than revealing the solution rule. This design, is also motivated by future works, which may adopt principle-based or modular architectures, where principle selection and principle execution are handled by separate components or logic rules.

At the same time, we agree that evaluating models without explicitly providing the principle is important. In response, we conducted an additional experiment in which the prompt includes only positive and negative examples, without specifying the Gestalt principle. See Figure below.

The results show that the effect of removing the principle is model-dependent. In particular,

- For GPT-5, providing the principle yields a modest improvement at high F1 thresholds, indicating that explicit principle information helps the strongest model achieve more fully correct solutions. (i.e. more tasks solved perfectly).
- For smaller models (e.g. InternVL3-2B, Llava-Qwen-7B, InternVL3-78B), removing the principle does not improve the perfect solving task percentage (i.e. the percentage at high F1 threshold).
- Some models show better performance without the principle at low F1 thresholds, suggesting more partial matches, but this does not necessarily reflect successful task solving.

We added the principle-blind results as a complementary analysis in the revised manuscript. (page 11, 18)

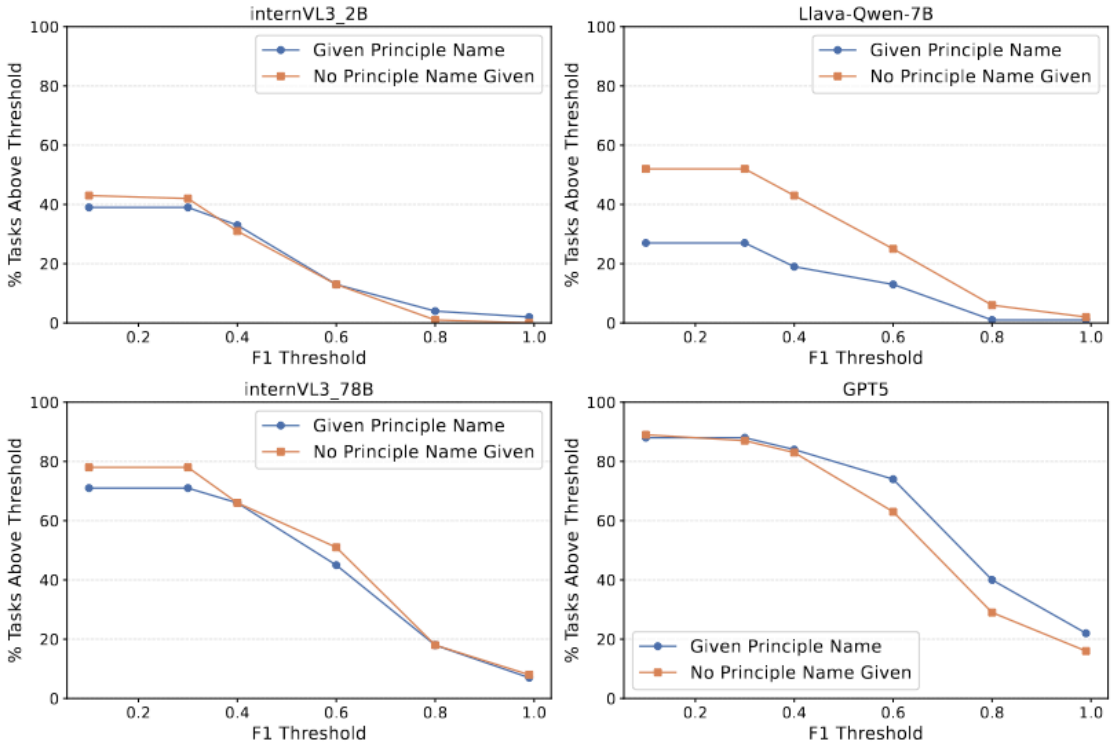

Figure 10. Comparison between principle-aware and principle-blind prompting on 100 proximity-based tasks across four VLMs. Each curve shows the percentage of tasks whose F1 score exceeds a given threshold. In the high-threshold region, principle-aware prompting yields a modest gain for GPT-5 but little change for the other VLMs, suggesting that explicit principle information mainly helps the strongest model achieve more fully solved tasks. In the low-threshold region, non-GPT-5 models show slightly higher task coverage without the principle name, although such gains are ambiguous and may reflect partial correctness or shortcut-based behavior rather than successful rule induction.

-