Journal Title
XX(X):1–34

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Gestalt Vision: A Dataset for Evaluating Gestalt Principles in Visual Perception

Jingyuan Sha[1], Kristian Kersting[1,2,3] and Devendra Singh Dhami[4]

Abstract

Gestalt principles, established in the 1920s, describe how humans perceive individual elements as cohesive wholes. These principles, including proximity, similarity, closure, continuity, and symmetry, etc., play a fundamental role in human perception, enabling structured visual interpretation. Despite their significance, existing AI benchmarks fail to assess models' ability to infer patterns at the group level, where multiple objects following the same Gestalt principle are considered as a group using these principles. To address this gap, we introduce Gestalt Vision, a framework designed to evaluate AI models' ability to not only identify groups within patterns but also reason about the underlying logical rules governing these patterns. Gestalt Vision provides structured visual tasks and baseline evaluations spanning neural and neural-symbolic approaches, uncovering key limitations in current models' ability to perform human-like visual cognition. Our findings emphasize the necessity of incorporating richer perceptual mechanisms into AI reasoning frameworks. By bridging the gap between human perception and computational models, Gestalt Vision offers a crucial step toward developing AI systems with improved perceptual organization and visual reasoning capabilities.

Keywords

Gestalt Principle, Scene Reasoning, Neuro-Symbolic AI

[1]Technische Universität Darmstadt, [2]Hessian Center for Artificial Intelligence (hessian.AI),
[3]German Research Centre for Artificial Intelligence (DFKI), [4]Eindhoven University of Technology

Corresponding author:
Jingyuan Sha
Email: jingyuan.sha@tu-darmstadt.de

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

Gestalt principles such as proximity, similarity, closure, symmetry, and continuity describe the innate ways in which human perception organizes visual information into coherent wholes (Wertheimer 1938; Koffka 1935; Ellis 1999; Palmer 1999). These principles allow humans to instinctively identify salient features and abstract high-level concepts from complex scenes. For example, we instinctively perceive symmetrical arrangements as unified structures and tend to complete incomplete shapes through closure, enabling rapid recognition of objects and their interrelationships (see Fig. 1). This perceptual strategy is particularly relevant in complex visual reasoning tasks, where it is important to move beyond the focus on individual pixels or discrete objects to discern overarching patterns and structures. Incorporating Gestalt principles enables visual reasoning models to better emulate human perception, improving object relationships and high-level reasoning.

Neuro symbolic systems typically combine deep learning models such as Mask R-CNN (He et al. 2017) or Slot Attention (Locatello et al. 2020) to detect objects and assign symbolic labels and bounding boxes (Shindo et al. 2023; Sha et al. 2024; Shindo et al. 2024). These symbolic abstractions then serve as the input to reasoning modules that operate over object-level representations. However, such pipelines often overlook crucial attributes including contours, size, color, and spatial distribution that are essential for context-sensitive inference. As a result, existing reasoning models may fail to capture nuanced information required for complex relational or group-level understanding. Addressing this limitation requires benchmarks that preserve both local and global visual features while testing models under systematic and controlled conditions.

To move toward this goal, we introduce the **Ge**sta**lt** **Vis**ion Benchmark (ELVIS), a synthetic dataset designed to evaluate models on perception and reasoning guided by Gestalt principles. Each task in ELVIS is constructed to emphasize one or more principles, with structured visual scenes and rule-based labels. Unlike conventional visual benchmarks, ELVIS focuses explicitly on group-level regularities in addition to isolated object features.

[This paper is an extended version of our previous conference publication (Sha et al. 2025). In this extended version, we significantly expand both the scale and the scope of the ELVIS benchmark. Specifically, we increase the number of tasks across all five Gestalt principles from approximately 1,000 to over 3,000 tasks, introduce a wider range of variations in the underlying logic patterns, and provide a more systematic and fine-grained experimental analysis. In addition, we include a broader set of baseline models, enabling a more comprehensive evaluation of current visual reasoning systems under principle-conditioned grouping tasks. (REVISED)]

We develop a systematic task generation framework that jointly considers object-level properties such as color, shape, size, and position, group-level properties such as group shape and size, and their combinations across different Gestalt principles. This enables the construction of hundreds of diverse tasks per principle, ensuring statistically robust evaluation while exposing more subtle reasoning challenges. We also evaluate several baseline models on the benchmark, including the recent GPT-5 (OpenAI 2025).

Overall, this work makes the following contributions:

1. [We present an extended version of the Gestalt Vision Benchmark (ELVIS)*, significantly scaling the dataset from approximately 1,000 to over 3,000 tasks across five Gestalt principles, with broader coverage of object-level and group-level properties and their combinations. (REVISED)]
2. We develop a systematic task generation framework with richer variations in underlying logic patterns, enabling more diverse and fine-grained evaluation of principle-conditioned perceptual reasoning.
3. We conduct an expanded empirical study with a broader set of baseline models, including recent large-scale VLMs, and provide detailed analyses at multiple levels, including overall performance, category-level, concept-level, and data-scaling behavior.
4. [We identify key limitations of current VLMs in principle-conditioned perceptual reasoning, including difficulty in achieving high-fidelity task solutions, limited ability to leverage explicit principle information, and unbalanced performance across different Gestalt principles, revealing gaps in generalizing perceptual grouping mechanisms. (REVISED)]
5. We release the dataset and evaluation code to facilitate future research on perceptual grouping, visual reasoning, and neuro-symbolic systems.

To this end, we proceed as follows. We first review related work, then introduce the ELVIS benchmark and its task generation process. We subsequently present experimental results and analyses, and conclude with a discussion of limitations and future directions.

Related Work

We will now review the relevant literature focusing on two major subareas, namely visual perception and neuro-symbolic reasoning.

Gestalt Principles and Computer Vision

Gestalt principles have a long and rich history in psychology, tracing back to seminal works by Wertheimer, Koffka, and Palmer (Wertheimer 1938; Koffka 1935; Palmer 1999; Ellis 1999). In recent decades, these foundational ideas have influenced a variety of computational models in machine learning and computer vision (Lörincz et al. 2017; Hua and Kunda 2020; Kim et al. 2021; Zhang et al. 2024), often aiming to replicate or approximate the human capacity for grouping and structural organization.

However, the majority of prior work has relied on convolutional networks or other purely neural techniques, which primarily capture local feature correlations but struggle with higher-order grouping. Explicit integration of Gestalt principles into computational systems remains relatively rare, and even fewer approaches combine neural perception with symbolic representations to preserve holistic grouping and reasoning capabilities.

*https://github.com/ml-research/ELVIS

This motivates the development of benchmarks such as ELVIS, which provide systematic tasks for evaluating how well models capture Gestalt-based organization beyond object-level detection.

Neuro-symbolic Learning and Reasoning

Neuro-symbolic approaches have emerged as a prominent paradigm that combines the perception strengths of neural networks with the interpretability and systematic generalization of symbolic reasoning. Over the past years, a variety of benchmarks have been introduced to evaluate such hybrid systems. Notable examples include CLEVR (Johnson et al. 2017), CLEVRER (Yi et al. 2020), V-LoL (Helff et al. 2025) and visual question answering frameworks that integrate ConceptNet and other knowledge graphs (Yi et al. 2018; Mao et al. 2019; Amizadeh et al. 2020; Tan and Bansal 2019). These resources have driven progress in compositional reasoning but predominantly focus on object detection, attribute recognition, and relatively simple relational inference.

Meanwhile, several benchmarks have sought to test higher-level reasoning and abstraction of the visual reasoning models by given only a small number of samples. Bongard problems (Ruchkin 1971) access the capacity to identify abstract rules that distinguish two sets of visual patterns, providing a foundational testbed for conceptual and relational reasoning. Abstract Visual Reasoning (AVR) tasks assess how well models generalize concepts in abstract settings, requiring compositional reasoning and transfer (Hu et al. 2021). CLEVRER explicitly introduces causal and physics-based reasoning with interacting objects (Yi et al. 2020). The Kandinsky Patterns benchmark (Müller and Holzinger 2021) and its three-dimensional extension (Sha et al. 2024) provide structured synthetic data to study relational abstraction and perceptual grouping. Additionally, the Alphabet Shape dataset (Sha et al. 2024) explores recognition of alphanumeric shapes constructed from grouped objects, highlighting grouping as a fundamental principle of cognition (Sellars 1912).

Despite these advances, most existing benchmarks do not systematically address grouping phenomena grounded in perceptual psychology. Our work extends this line of research by explicitly incorporating Gestalt principles such as proximity, similarity, closure, symmetry, and continuity in CLEVR to generate a new benchmark. The Gestalt Vision Benchmark (ELVIS) evaluates the ability of neuro-symbolic models to detect and reason over grouping-based structures, moving beyond object-level perception toward more holistic and human-aligned reasoning. In the extended version presented here, we broaden the task generation process to systematically include both object-level and group-level properties, as well as combinations across principles. This creates a richer and more comprehensive testbed for measuring the capabilities and limitations of neuro-symbolic reasoning in visual abstraction.

Gestalt Vision (ELVIS): A Gestalt Reasoning Benchmark

Ges**al**t **Vis**ion (ELVIS) is a curated collection of synthetic visual scenes that emphasize five key Gestalt principles: Proximity, Similarity, Closure, Continuity, and Symmetry, as illustrated in Figure 1. These principles are essential for understanding how discrete

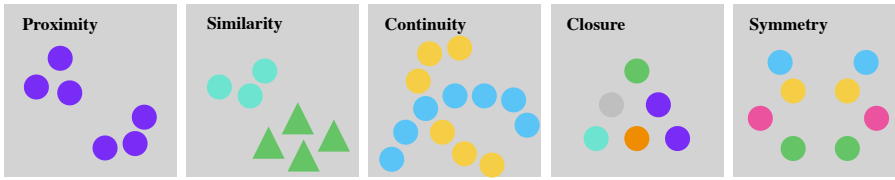


Figure 1. Gestalt Principles Supported by ELVIS. From left to right: **Proximity**: Objects that are spatially close to each other are perceived as a group. **Similarity**: Objects with common attributes, such as shape or color, are grouped together. **Continuity**: Objects with continue positions are grouped together. **Feature Closure**: Objects with aligned visual features create an implicit, complete shape. **Position Closure**: Objects arranged in a manner that suggests a closed contour are grouped. **Symmetry**: Objects mirrored across an axis are perceived as a structure, each symmetric pair determines a group.

Table 1. Benchmark tasks summarization. The table summarizes the benchmark tasks for each Gestalt principle. Columns report the number of categories, number of tasks, object number range, average number of objects, group number range, and object size range. All principles share a common pool of 150 colors and 12 shapes.

Principle	# Cat.	# Task	# Obj. Range	# Avg. Obj.	# G.	Size Range
Proximity	4	588	3-90	22	2-4	$5\% \sim 80\%$
Similarity	3	872	4-196	58	1-4	$2\% \sim 10\%$
Closure	6	596	3-60	19	1-4	$3\% \sim 12\%$
Continuity	4	432	4-68	24	1-4	$3\% \sim 8\%$
Symmetry	3	900	3-49	20	2-7	$5\% \sim 40\%$

visual elements are perceived as cohesive patterns. It is an important challenge for visual reasoning models that integrate learned perceptual features with logical reasoning mechanisms.

Overview of ELVIS

ELVIS provides a systematic collection of synthetic visual tasks that highlight how Gestalt principles shape perceptual grouping and reasoning. Each task is generated from scenes composed of objects defined by core attributes such as color, shape, size, position, and quantity. These object-level properties interact with group-level features including collective arrangements, symmetry axes, shared color distributions, or composite group shapes.

This design moves beyond simple object recognition toward reasoning about high-level organization. Models are expected not only to identify which objects are present but also to infer how these objects form structured patterns under Gestalt constraints. For example, elements that appear close to each other can be grouped by proximity, partially occluded figures can be completed through closure, and objects aligned around a symmetry axis can be perceived as a unified whole.

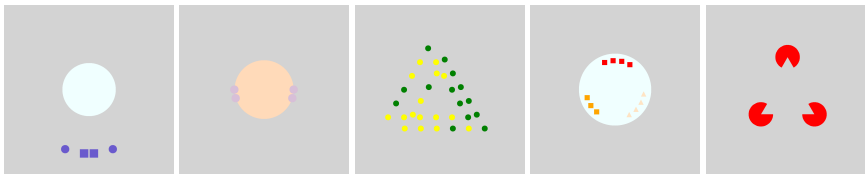


Figure 2. Geometric Feature Scenarios. Example patterns illustrating different geometric feature scenarios. **From left to right**: individual objects, object overlap, group overlap, nested shapes, and incomplete forms, designed to assess model perception under varied spatial configurations.

Table 1 shows the summarization of the tasks in the benchmark. With thousands of tasks spanning diverse principles and property combinations, ELVIS ensures broad coverage and statistically robust evaluation. The benchmark thus provides a challenging yet principled environment for testing visual reasoning models, encouraging them to capture the same perceptual strategies that humans naturally use when organizing visual input into meaningful structures.

Data Generation

The ELVIS benchmark is generated under controlled conditions to systematically capture a wide spectrum of Gestalt principles while ensuring reproducibility and clarity of evaluation. Through these controlled yet diverse design choices, ELVIS challenges computational models to perform context-sensitive reasoning. Rather than limiting evaluation to low-level classification, the benchmark tests whether models can apply logical rules to organize visual elements into coherent wholes, which is an ability central to visual reasoning systems that aim to bridge pixel-level perception with symbolic-level reasoning.

Diverse Objects. As shown in table 1. Scenes contain up to 12 variations of basic shapes, with as many as 150 color variations and a size range spanning approximately 2% to 80% of the image width. This diversity provides rich input for perception and reasoning, reducing the risk of bias and overfitting while improving robustness and generalization across models.

Varied Complexity. The number of objects in a scene ranges from a handful in simple settings to several hundred in complex ones. Regardless of density, each scene is designed to clearly embody a target Gestalt principle. Objects that participate in the same principle may still differ in shape, color, and size, ensuring that task difficulty arises from heterogeneous attributes that models must jointly interpret .

Explicit Groupings. Object arrangements are deliberately constructed to make grouping cues unambiguous. For example, proximity clusters are placed with clear separation from other clusters, and symmetrical arrangements align precisely around defined axes. This design minimizes confounding factors while ensuring that comparisons across models remain reliable.

Features in the patterns

Although the patterns are composed of basic geometric shapes such as triangle, square, etc. Their variations extend beyond simple shape detection. Figure 2 illustrates five distinct scenarios designed to challenge the robustness of perception models: **Individual**, the objects are placed individually without overlapping, which is the most straightforward case; **Object Overlap**, the objects are overlapped with each other, which can cover part of the features of some of the objects in the image; **Group Overlap**, multiple groups are overlapped with each other, whereas the objects are still remaining individual. **Inside**, the objects are completely inside another object; and **Incomplete**, the object is not completely drawn in the image, which sometimes shows the features of other shapes.

These variations test the model's ability to handle occlusion, containment, and missing features, ensuring a deeper understanding of geometric properties.

Category

Although we provide hundreds of tasks for each Gestalt principle, we do not code and design them individually. Instead, we introduce a base pattern called a *category* to efficiently generate multiple tasks. Each category is explicitly designed around a specific Gestalt principle. By modifying key attributes, such as the number of groups, the number of objects within each group, and the color, shape, or size of each object, we can create numerous variations while maintaining the same underlying principle. Figure 3 presents examples of each category used in the ELVIS. Table 7 in Appendix presents the detailed information for each category.

Task Formulation

Each task in ELVIS is defined by a set of rules, which specify a combination of logical conditions that determine the structure of valid visual patterns. These rules are instantiated as constraints on object-level properties (e.g., shape, color, size, count) and group-level configurations (e.g., spatial arrangement, symmetry). For example, a rule might require that each group contains one red triangle, or several objects form a symmetrical structure.

Using these rules, the dataset generation pipeline creates a set of positive images that fully satisfy all constraints and a corresponding set of negative images, each of which violates at least one constraint.

Each image is assigned a binary label: positive sample has label 1 and negative example has label 0. A task is defined as the classification problem of distinguishing these two types of images based on their compliance with the underlying rules. Figure 4 shows an example of a task.

In this setting, the rules capture the complete logical structure of the visual pattern, the constraints represent the atomic predicates that compose the rules, the label indicates whether an image satisfies all constraints, and the task refers to the binary classification challenge associated with rules. [Specifically, 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,

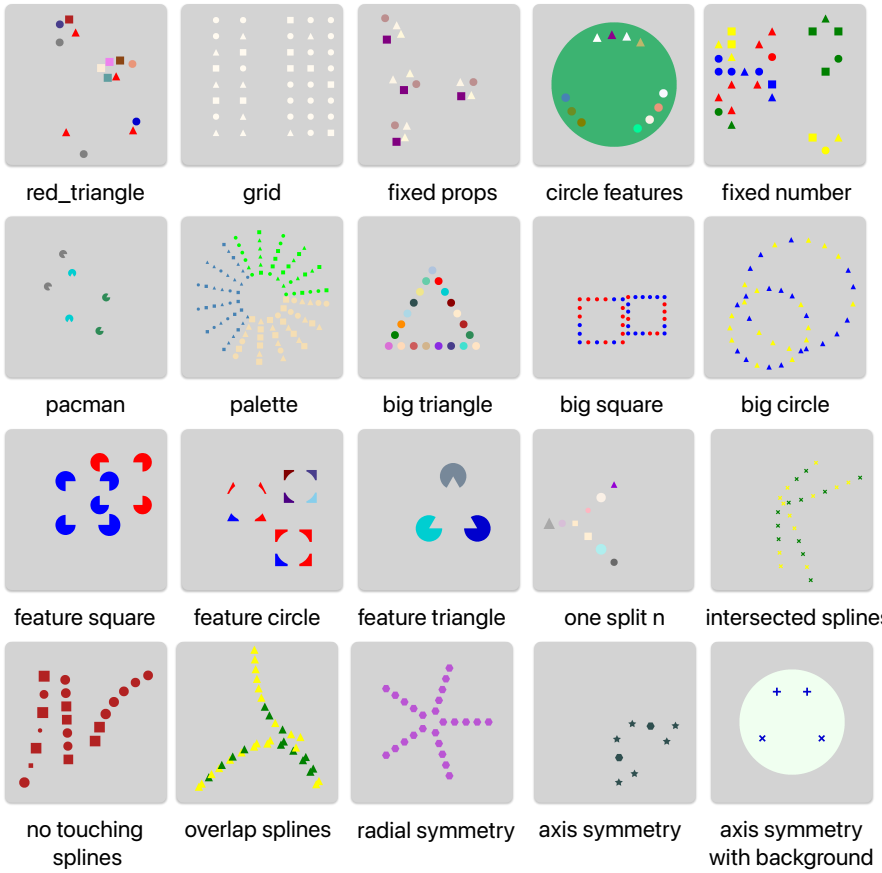


Figure 3. Category Base Patterns of ELVIS. Each category in ELVIS is based on a specific Gestalt principle. The base pattern of each category serves as a foundational structure, which can generate numerous variations by adjusting object properties.

and size, count) 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

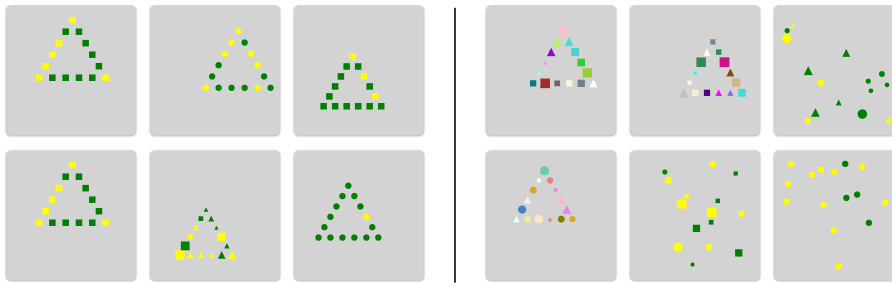


Figure 4. Task Example of ELVIS. **Left:** Positive patterns illustrating the Gestalt principle of closure, where objects collectively form a yellow-green triangle. **Right:** Negative patterns that partially adhere to the rule but violate key constraints, either by not matching the required color or by failing to complete the triangular closure.

among object-level properties may still exist in principle. However, these do not eliminate the need for detecting the Gestalt principle itself. (REVISED)]

This formulation allows models to be evaluated in a focused and interpretable manner, testing their ability to infer meaningful group-level properties from structured visual input.

Comparison with Existing Datasets

To clarify the position of ELVIS among existing visual reasoning benchmarks, Table 2 compares representative datasets along perceptual, structural, and reasoning dimensions. A cross ($\times$) indicates that a capability is *not explicitly evaluated or supervised* by the benchmark, even if it may be implicitly required for solving some tasks.

CLEVR (Johnson et al. 2017) is an object-centric dataset that supports relational and rule-based reasoning over explicitly annotated objects. However, its reasoning operates purely at the object level: perceptual grouping and Gestalt organization are neither required nor evaluated. While CLEVR scenes can contain dozens of objects, this scale remains substantially smaller than ELVIS, which supports scenes with up to hundreds of objects and emphasizes group-level structure. PGM/RPM (Raven and Court 1998) focus on abstract rule induction over fixed structural slots rather than object-centric representations. The spatial layout, primitives, and correspondence structure are predefined by design, making perceptual organization and grouping unnecessary. As a result, it evaluates symbolic pattern abstraction but does not test perceptual grouping or Gestalt principles. Bongard-LOGO (Nie et al. 2020) evaluates high-level concept learning from positive and negative examples and places strong demands on visual perception, such as distinguishing curves, lines, and small geometric primitives. Nevertheless, it does not explicitly model object-centric reasoning or perceptual grouping; the task is framed at the level of global visual concepts rather than structured object groups. ARC(-AGI) (Chollet et al. 2025) requires strong perceptual abilities to

Dataset	Synthetic	Object-Centric	Grouping-Centric	Rule Induction
CLEVR	✓	✓	✗	✓
PGM / RPM	✓	✗	✗	✓
Bongard-LOGO	✓	✗	✗	✓
ARC(-AGI)	✓	✗	✗	✓
ELVIS (ours)	✓	✓	✓	✓

Table 2. Comparison between ELVIS and representative visual reasoning datasets. ELVIS uniquely supports explicit and controllable Gestalt-based perceptual grouping with group-level annotations, enabling systematic analysis of perceptual organization in visual reasoning.

identify relevant visual elements and to infer transformation rules across scenes. Many ARC tasks implicitly rely on grouping. For example, treating multiple tiles as a coherent part or applying a shared rule across subsets of elements. However, ARC does not provide group-level supervision, nor does it explicitly evaluate grouping as a standalone capability. Instead, it targets a broader notion of generalization and abstract reasoning that goes beyond the focused scope of ELVIS.

In contrast, ELVIS is explicitly designed to study visual reasoning grounded in perceptual organization. It provides controllable Gestalt-based grouping mechanisms, group-level annotations, and systematic factor variations, enabling targeted evaluation of how perceptual grouping supports downstream reasoning.

Empirical Evaluation using ELVIS

We now evaluate the ELVIS benchmark with some state-of-the-art neural and neuro-symbolic methods to demonstrate the shortcoming(s) of current machine learning models.

Task Types and Evaluation Metrics

ELVIS comprises a diverse set of tasks designed to evaluate how effectively computational models can identify and reason about Gestalt principles. Table 1 summarizes the task distribution. Each principle is associated with hundreds of tasks that feature considerable variation in visual complexity, such as object count (ranging from a few to several hundred), color diversity (hundreds of different colors), object shapes (12 different shapes), and object sizes (varying between 2% and 80% of the width of the image). These variations ensure that the benchmark tests a wide array of perceptual scenarios.

We evaluate each model in a strictly task-wise manner. Each task is associated with two disjoint splits: a training split and a test split. By default, each split contains three positive and three negative examples. All models are given access to the same number of training and testing examples, ensuring a controlled and fair comparison across model families.

For neural baselines such as ViT, we train a separate binary classifier for each task using the labeled examples in the training split and evaluate it on the corresponding test split. For vision–language and language models (LLaVA-OneVision, InternVL3, GPT-5),

Table 3. Large Models Comparison Five large models were used for benchmark evaluation. The ViT refers to the ViTB16 model pretrained on ImageNet-1K, LlaVA-7B refers to LLaVA-OneVision-Qwen2-7B-SI (a multi-modal model incorporating text-image understanding), InternVL3-2B and InternVL3-78B are models from the InternVL3 series, and GPT-5.

Model	Pretrained Dataset	Image Resolution	Params (M)
ViT	ImageNet-21K	224×224	86
LlaVA-7B	Multi-modal	224×224	7000
InternVL3-2B	Multi-modal	224×224	2100
InternVL3-78B	Multi-modal	224×224	78400
GPT-5	Multi-modal	224×224	635000

tasks are evaluated independently using a lightweight supervised demonstration protocol. Specifically, the model is provided with the training split examples together with a textual description of the Gestalt principle instantiated by the task, and is then asked to predict the labels of the six images in the test split.

[We adopt this principle-aware protocol as the default evaluation setting. The rationale is that the five Gestalt principles correspond to qualitatively different perceptual tasks, each relying on distinct visual cues and grouping mechanisms. Proving the target principle therefore serves as part of the task specification rather than revealing the label rule itself. This design is particularly relevant for future principle-based or modular reasoning systems,in which principle selection and principle execution may be handled by separate components. We treat the principle-blind setting, where the model must additionally infer the relevant grouping cue, as a complementary, harder evaluation task identification rather than the primary benchmark condition. (REVISED)]

This evaluation protocol supplies minimal task-local supervision while explicitly disallowing gradient-based adaptation and task-to-task transfer. Model outputs are generated as free-form text and deterministically parsed into binary labels. Full prompt templates and parsing rules are provided in Appendix . Although the output format is binary, solving an ELVIS task requires identifying the latent Gestalt relation that differentiates two sets of patterns—akin to classical visual ILP problems, where simple labels mask rich underlying structure. We adopt this minimal interface to ensure comparability across neural, symbolic, and foundation models, while isolating the core challenge of forming the correct abstract grouping concept. Performance is measured using accuracy and F1 score, and we report the mean and standard deviation across all tasks in the benchmark.

Baseline Models

We evaluated four representative baselines, encompassing neural and VLM approaches. Table 3 summarizes the characteristics of the baseline models.

Vision Transformer (ViT-B/16) (Wu et al. 2020; Wightman 2019) is a purely neural model pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K, providing strong visual perception capabilities at a resolution of 224×224 pixels. It is a transformer-based

vision model that represents an image as a sequence of patch tokens rather than using convolutional features. Each image is split into 16×16 patches which are embedded into vector tokens.

LLaVA-OneVision (Li et al. 2024), an advanced multimodal Large Language Model (LLM) that extends text-based language modeling to incorporate visual inputs. Built upon the Qwen2 LLM as its language backbone, LLaVA-OneVision is fine-tuned on extensive multimodal instruction data—for example, image-question-answer pairs and vision-language dialogues.

InternVL3 (Chen et al. 2024) represents the latest generation of multimodal foundation models that integrate visual perception and language reasoning in a unified architecture. Unlike earlier models that relied on separate encoders, InternVL3 adopts a shared token-based interface between vision and language, enabling tighter cross-modal alignment. We employ two model variants as baselines: *InternVL3-2B*, a smaller model suited for efficiency and fast inference, and *InternVL3-78B*, a large-scale model designed for state-of-the-art multimodal reasoning. The two scales allow us to assess how model capacity influences performance on our Gestalt reasoning tasks.

GPT-5 (OpenAI 2025) is the latest generation multimodal large language model developed by OpenAI, supporting both image input and text output within a unified architecture. Compared to prior models that layered vision modules on top of text-only LLMs, GPT-5 was jointly trained on large-scale multimodal corpora, enabling integrated reasoning over visual and linguistic information. We use the GPT-5 multimodal variant as a baseline to examine how a state-of-the-art large model performs on our Gestalt reasoning benchmark.

Overall Evaluation

Table 4 reports the comparative performance of all baseline models across five Gestalt principles. The results reveal distinct differences between purely neural models and larger multimodal architectures in their ability to capture structural regularities.

The ViT baseline, despite being trained on large-scale natural images, achieves only moderate accuracy (around 0.5 across principles) and suffers from unstable F1, precision, and recall. This inconsistency indicates that the model fails to form robust grouping representations and often resorts to biased predictions. The high recall score over principle similarity indicates that the ViT is overly biased toward predicting the positive class. LLaVA-Qwen-7B and InternVL3-2B exhibit similar limitations: while they outperform ViT on certain principles such as closure and symmetry, their overall performance remains unstable.

InternVL3-78B demonstrates a notable improvement over the smaller models, with consistently higher scores across all metrics and principles. Its gains are especially visible for similarity, closure, and continuity. This reflects the benefits of scale in capturing higher-order structural relations. GPT-5 achieves the strongest performance overall, with the highest accuracy and precision across nearly every principle, particularly for closure (0.77 accuracy) and similarity (0.71 accuracy). However, GPT-5 shows relative weakness

Table 4. Performance Comparison. The mean and standard deviation over four evaluation metrics: accuracy, F1 score, precision, and recall. ViT-16-224 refers to the ViT-B/16 model, and Llava-Qwen-7B denotes LLaVA-OneVision-Qwen2-7B-SI.

Met.	Model	Proximity	Similarity	Closure	Symmetry	Continuity
Acc.	ViT-16-224	0.52 ± 0.15	0.52 ± 0.12	0.54 ± 0.17	0.50 ± 0.14	0.54 ± 0.14
	Llava-Qwen-7B	0.49 ± 0.15	0.49 ± 0.13	0.63 ± 0.19	0.57 ± 0.18	0.50 ± 0.15
	InternVL3-2B	0.52 ± 0.14	0.51 ± 0.15	0.60 ± 0.17	0.57 ± 0.17	0.54 ± 0.14
	InternVL3-78B	0.61 ± 0.17	0.61 ± 0.21	0.73 ± 0.20	$\mathbf{0.62} \pm 0.18$	0.65 ± 0.18
	GPT-5	$\mathbf{0.69} \pm 0.19$	$\mathbf{0.71} \pm 0.23$	$\mathbf{0.77} \pm 0.19$	0.60 ± 0.18	$\mathbf{0.69} \pm 0.20$
F1	ViT-16-224	0.30 ± 0.30	0.58 ± 0.23	0.48 ± 0.30	0.23 ± 0.30	0.33 ± 0.35
	Llava-Qwen-7B	0.21 ± 0.29	0.33 ± 0.33	0.53 ± 0.33	$\mathbf{0.46} \pm 0.33$	0.22 ± 0.30
	InternVL3-2B	0.23 ± 0.30	0.31 ± 0.31	0.33 ± 0.35	0.36 ± 0.33	0.21 ± 0.30
	InternVL3-78B	0.41 ± 0.35	0.46 ± 0.37	0.59 ± 0.37	0.45 ± 0.36	0.51 ± 0.33
	GPT-5	$\mathbf{0.65} \pm 0.29$	$\mathbf{0.63} \pm 0.35$	$\mathbf{0.67} \pm 0.33$	0.40 ± 0.35	$\mathbf{0.55} \pm 0.37$
Pre.	ViT-16-224	0.37 ± 0.39	0.48 ± 0.22	0.48 ± 0.32	0.25 ± 0.35	0.32 ± 0.34
	Llava-Qwen-7B	0.24 ± 0.34	0.30 ± 0.31	0.55 ± 0.37	0.46 ± 0.34	0.24 ± 0.34
	InternVL3-2B	0.30 ± 0.39	0.36 ± 0.37	0.44 ± 0.45	0.42 ± 0.40	0.28 ± 0.40
	InternVL3-78B	0.49 ± 0.41	0.49 ± 0.40	0.70 ± 0.29	$\mathbf{0.51} \pm 0.41$	0.61 ± 0.39
	GPT-5	$\mathbf{0.65} \pm 0.31$	$\mathbf{0.66} \pm 0.34$	$\mathbf{0.76} \pm 0.24$	0.49 ± 0.42	$\mathbf{0.62} \pm 0.38$
Rec.	ViT-16-224	0.31 ± 0.35	$\mathbf{0.80} \pm 0.19$	0.56 ± 0.40	0.24 ± 0.35	0.40 ± 0.43
	Llava-Qwen-7B	0.22 ± 0.33	0.44 ± 0.46	0.57 ± 0.40	$\mathbf{0.55} \pm 0.41$	0.24 ± 0.34
	InternVL3-2B	0.22 ± 0.31	0.33 ± 0.36	0.29 ± 0.34	0.35 ± 0.36	0.19 ± 0.29
	InternVL3-78B	0.41 ± 0.38	0.50 ± 0.42	0.55 ± 0.39	0.45 ± 0.40	0.50 ± 0.36
	GPT-5	$\mathbf{0.71} \pm 0.29$	0.66 ± 0.34	$\mathbf{0.65} \pm 0.35$	0.40 ± 0.38	$\mathbf{0.55} \pm 0.40$

on symmetry, where both accuracy and F1 lag behind its other results, suggesting that certain spatial-relational cues remain challenging.

However, the heterogeneous behavior observed across different Gestalt principles and task types (reported in the following section) is not unexpected. Each Gestalt principle captures a distinct aspect of perceptual organization and relies on fundamentally different visual cues, making direct performance comparisons across principles inherently limited. For example, similarity-based grouping primarily depends on object attributes such as color, shape, and size, while largely disregarding spatial position, which is instead the dominant cue for proximity-based grouping. In contrast, principles such as closure and continuity crucially depend on the integrity of contour information and the perceptual coherence of group boundaries.

Moreover, spatial position itself constitutes a particularly challenging perceptual feature. The symmetry principle, for instance, requires the model to accurately identify correspondences between object pairs across a latent symmetry axis. Unlike proximity or closure, symmetric groups do not exhibit a fixed or compact spatial configuration; successful grouping, therefore, hinges on correctly inferring global scene structure rather than local feature similarity alone. These intrinsic differences across principles help explain the observed variability in model performance and underscore the need to analyze Gestalt principles as distinct perceptual challenges rather than as interchangeable

grouping tasks. As we discuss further in the category level evaluation, model performance varies even within the same principle across different task categories.

Category Level Evaluation

Figure 5 presents the average F1 scores across task categories. The ViT baseline remains weak overall, with scores rarely exceeding 0.5, and performs particularly poorly on proximity, symmetry, and continuity. InternVL3-2B records the lowest performance across most categories, with only marginal strengths in isolated cases. LLaVA-Qwen-7B shows a more imbalanced profile, performing better on closure and symmetry but worse on the remaining principles. InternVL3-78B achieves a clear performance gain, exceeding 0.6 on several closure-related categories and maintaining more stable results overall. GPT-5 delivers the best performance, surpassing 0.7 on proximity, similarity, and closure, though symmetry continues to present challenges.

In summary, two key trends emerge: (i) purely neural vision models struggle to generalize Gestalt rules despite their strong performance on natural image recognition, (ii) multi-modal integration improves results but only at larger scales. These findings quantitatively support the need for neuro-symbolic mechanisms, as even the strongest models show principle-specific weaknesses and lack systematic compositionality across grouping cues.

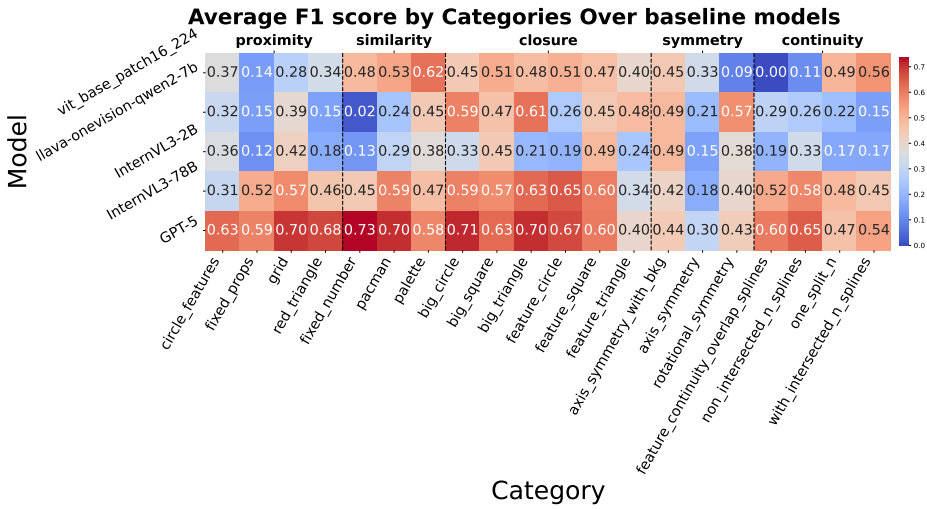


Figure 5. Average F1 score by Categories Over baseline models. The chart compares average F1 scores (y-axis) for proximity, similarity, closure, symmetry, continuity, and related categories (x-axis).

Figure 6 presents a comprehensive factor-level performance analysis across all five Gestalt principles for the best-performing model, GPT-5. Empty cells indicate factors that are not manipulated within specific categories. For example, *group_1* is absent in

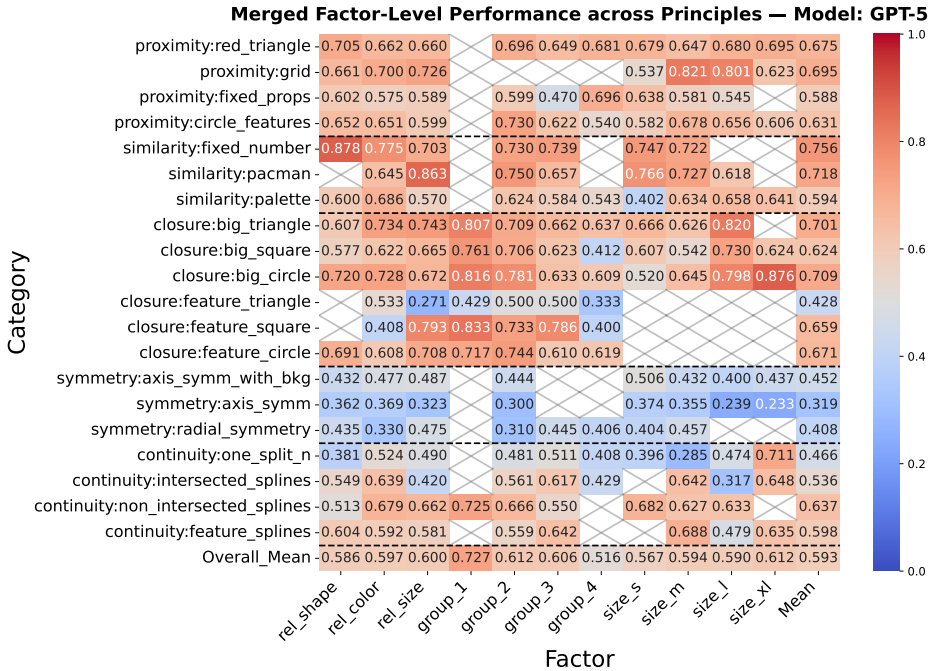


Figure 6. Factor-Level Performance per Category Across Gestalt Principles on GPT-5. Heatmaps show mean F1 scores for different experimental factors across task categories for all evaluated models. Empty cells indicate factors not manipulated in that category. Red indicating better performance.

proximity-based tasks because proximity grouping is only meaningful when at least two groups are present, while group-size factors are fixed for closure categories due to the deterministic number of objects in those stimuli. The figure reveals substantial variation in task difficulty even among categories governed by the same Gestalt principle. For instance, GPT-5 achieves an F1 score of 0.637 on the *non-intersected splines* category, but only 0.466 on the *one-split* category, although both are based on the continuity principle. This intra-principle variability leads to markedly different aggregate performance across Gestalt principles. Overall, symmetry-based tasks are consistently more challenging than tasks based on other Gestalt principles, indicating that symmetry perception remains a particularly difficult capability for current models.

Another notable observation is that none of the baseline models fully solve any task category. Even GPT-5, which achieves its best performance on the *fixed-number* category (F1 = 0.878), fails to reach near-perfect accuracy. This suggests that while large vision–language models such as GPT-5 are effective at describing object attributes and pairwise relationships, they struggle to induce globally consistent logical rules that simultaneously account for all objects in a scene. A similar limitation has been reported

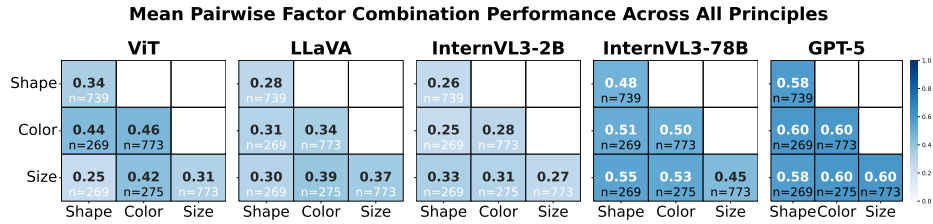


Figure 7. Mean F1 scores for pairwise factor combinations across all Gestalt principles. Each subplot shows a different model's performance when two visual factors (Shape, Color, Size) are present as relevant cues. Values are averaged across all five Gestalt principles and task categories. Cell annotations show mean F1 scores; small text shows task counts ("n=X"). Color scale: red (good performance) to blue (poor performance).

in the context of Bongard-style visual reasoning problems Wüst et al. (2025), where models often identify plausible local patterns but fail to capture the underlying abstract rule governing all examples.

Concept Level Evaluation

Object-Level Concept Analysis Figure 7 summarizes the mean performance for pairwise combinations of object-level factors across all Gestalt principles and baseline models. ViT exhibits a pronounced bias toward color-related patterns, achieving its strongest performance on color-only and color–shape combinations (up to 0.46 F1), while performing substantially worse on size-related tasks. In contrast, vision–language models show a more balanced performance profile across object-level concepts, with relatively smaller gaps between shape, color, and size factors. Notably, tasks involving a single factor yield performance comparable to those involving two combined factors, suggesting limited compositional gains at the object level for these models. Despite its overall strong performance, InternVL3-78B shows a clear F1 drop on size-related patterns.

Group-Level Concept Analysis Figure 8 compares model performance across group-level concepts, specifically *group number* and *group size*. In contrast to the object-level analysis, group-level concept analysis examines how performance varies as the number of groups in a scene increases and as the number of objects within each group changes. Here, *group size* denotes the number of objects belonging to a single group.

In figure 8, we can see that tasks involving a single group are consistently easier than those involving multiple groups across all VLMs. Moreover, performance improves monotonically with increasing group size, indicating that larger groups provide stronger and more redundant perceptual cues for reliable grouping. This trend suggests that sparse or highly fragmented group structures pose a greater challenge for current models, whereas denser groups facilitate more stable group-level representations.

Figure 9 illustrates examples of increasing group size (small, medium, large, and extra-large) from the *big_circle* category under the *closure* principle. Across these scenes,

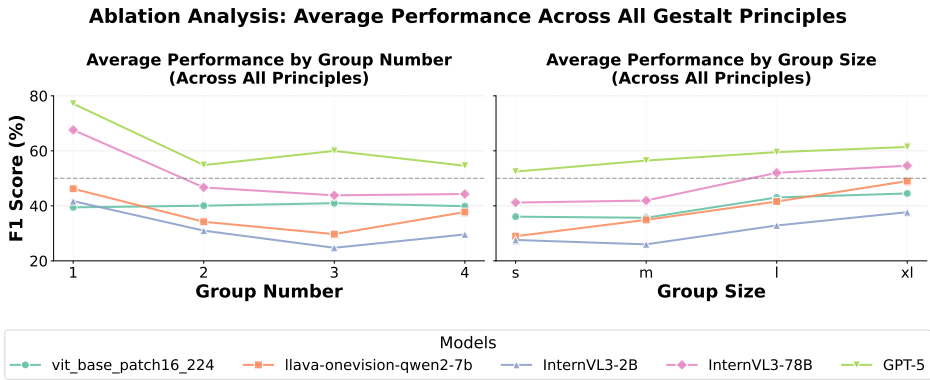
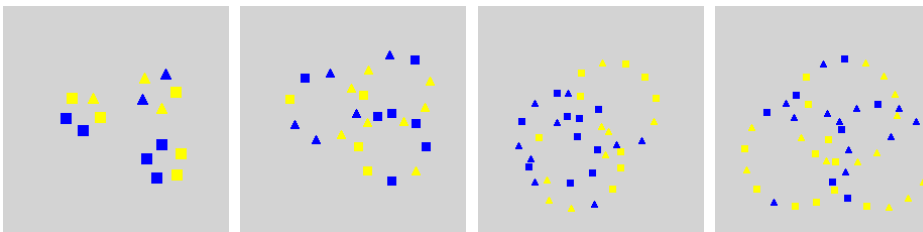


Figure 8. Baseline performance across gestalt groups for five principles. Left figure shows the F1 score over different group numbers. Right figure shows the F1 score over different group sizes.



(a) Group Size: *s* **(b)** Group Size: *m* **(c)** Group Size: *l* **(d)** Group Size: *xl*

Figure 9. Examples of different group sizes. Larger group sizes correspond to a greater number of objects within each group.

objects form three circular groups, with the number of objects per group increasing from left to right. Correspondingly, GPT-5 achieves F1 scores of 0.00, 0.86, 0.80, and 1.00 on the tasks involving four examples. This performance gap highlights the strong dependence of VLMs on the density of visual evidence for reliable grouping. In small-group settings, the reduced number of objects leads to increased ambiguity, as objects can be plausibly partitioned into multiple competing group configurations. At the same time, the proportion of task-relevant relational structure decreases, requiring models to rely more heavily on abstract reasoning to extract meaningful cues from sparse observations. As a result, small-group tasks demand stronger analytical capabilities, since the signal-to-noise ratio is lower and the distinction between meaningful structure and spurious patterns becomes increasingly subtle.

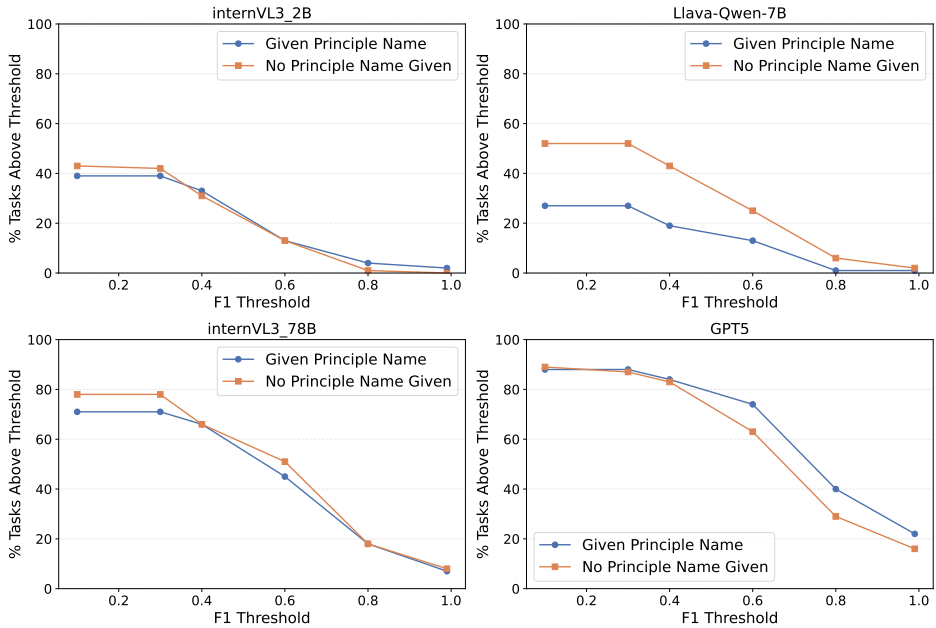


Figure 10. [Comparison between principle-aware and principle-blind prompting on 100 proximity-based tasks across four VLMs. Each curve shows the percentage of tasks whose F1 score exceeds a given threshold. In the high-threshold region, principle-aware prompting yields a modest gain for GPT-5 but little 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. (REVISED)]

Principle-Aware Evaluation Discussion[(REVISED)]

We use a principle-blind setting as a complementary evaluation. To assess whether a VLM can execute the appropriate perceptual grouping mechanism once the task family is specified, we adopt the principle-aware setting as the main protocol and compare it with a principle-blind variant in which the model is not given the Gestalt principle name.

Figure 10 shows this comparison on 100 proximity-based tasks across four VLMs. The effect of providing the principle name is not uniform across models or thresholds. Instead of producing a consistent shift over the whole curve, principle-aware and principle-blind prompting differ mainly in how they affect low- and high-threshold performance.

In the high-threshold region, which better reflects whether a model fully solves a task rather than only achieves partial correctness, providing the principle name does not substantially change the performance of most VLM baselines, including InternVL3-2B, LLaVA-Qwen-7B, and InternVL3-78B. In contrast, GPT-5 shows a modest but visible improvement in the principle-aware setting at higher thresholds. This indicates that

explicit principle information is useful for GPT5, while the smaller VLMs appear largely unable to convert such information into consistently correct task execution.

A second observation concerns the low-threshold region. For several non-GPT-5 models, the principle-blind setting yields a slightly higher proportion of tasks above small F1 thresholds. However, this should be interpreted with caution. Performance in this regime only indicates that the model predicts some examples correctly; it does not establish that the underlying grouping rule has been identified. Such partial success may arise from weak heuristics, superficial feature correlations, or confounding shortcuts that improve predictions without capturing the true task logic. Therefore, low-threshold improvements in the principle-blind setting are not sufficient evidence of better rule-based reasoning, and should be distinguished from gains in the high-threshold regime, where correct task solving is more plausibly associated with successful rule induction.

These findings help clarify the role of the two protocols. The principle-blind setting reflects a harder and more entangled problem, since the model must both infer the relevant grouping cue and apply it correctly. The principle-aware setting removes this first source of ambiguity and therefore provides a cleaner estimate of principle-conditioned perceptual reasoning. We therefore use it as the main evaluation protocol in ELVIS, while retaining the principle-blind condition as a complementary stress test of open-ended principle identification.

Effect of Training Set Size

We further test the impact of training image number using ViT-B/16 with two settings: ViT-16-224/3 trained with three images per class (positive and negative), and ViT-16-224/100 trained with one hundred images per class (positive and negative). As shown in Table 5, the three-shot model achieves slightly above-chance accuracy (around 0.5) and F1 scores that vary across principles (e.g., 0.58 on similarity but only 0.23 on symmetry), reflecting weak but non-trivial generalization.

Principle	Accuracy		F1 Score	
	ViT-16-224/3	ViT-16-224/100	ViT-16-224/3	ViT-16-224/100
Proximity	0.52 ± 0.15	0.50 ± 0.00	0.30 ± 0.30	0.00 ± 0.05
Similarity	0.52 ± 0.12	0.50 ± 0.08	0.58 ± 0.23	0.00 ± 0.04
Closure	0.54 ± 0.17	0.50 ± 0.02	0.48 ± 0.30	0.01 ± 0.09
Symmetry	0.50 ± 0.14	0.50 ± 0.00	0.23 ± 0.30	0.00 ± 0.02
Continuity	0.54 ± 0.14	0.50 ± 0.33	0.33 ± 0.35	0.01 ± 0.08

Table 5. Effect of Training Set Size over ViT-B/16. ViT-B/16 trained with three images retains weak generalization, while training with one hundred images collapses to constant predictions, yielding random-level accuracy and near-zero F1 score.

In contrast, the hundred-shot model collapses during training, producing almost constant predictions that yield accuracy near 0.50 and F1 scores close to zero across all principles. This indicates that enlarging the training set does not help the model

discover the underlying Gestalt relation; instead, it amplifies the model's tendency to rely on non-generalizable cues, leading to uniformly poor performance in the higher-shot setting.(Kim and Kim 2023)

Computation Cost and Hardware Requirement

Table 6 reports the average task-solving time (in seconds) across the five Gestalt principles. The measurement covers the complete task-solving pipeline, starting from the input of the training images and ending with the predicted labels for the test images, and reports wall-clock time.

GPT-5 is accessed through the API, while InternVL-78B is evaluated using three NVIDIA A100-SXM4-80GB GPUs. All remaining models are evaluated on a single NVIDIA A100-SXM4-80GB GPU.

Vision-only models such as ViT and smaller VLMs such as InternVL-2B exhibit substantially lower computational cost than large VLMs. InternVL-2B and ViT show stable runtimes in the range of 5 to 12 seconds across all principles . Despite their efficiency, these models achieve relatively lower F1 scores compared to larger models.

Large VLMs are significantly more expensive. LLaVA (7B) requires approximately 23–40 seconds per task on average, while InternVL-78B achieves runtimes around 10–15 seconds but relies on three 80GB GPUs. GPT-5 is the most computationally expensive model, exceeding 100 seconds across all principles and reaching over 150 seconds on similarity tasks. These large models achieve moderately higher F1 scores than the other baselines, illustrating a clear trade-off between computational efficiency and performance.

Model	Proximity	Similarity	Closure	Symmetry	Continuity
ViT	12.49	10.07	10.48	8.43	9.83
LLaVA	26.84	23.56	39.89	27.26	30.71
InternVL-2B	5.06	5.30	6.26	5.60	6.13
InternVL-78B	14.11	13.34	15.58	12.30	14.89
GPT-5	109.45	94.34	105.46	157.56	82.59

Table 6. Average inference time (seconds) across the five Gestalt principles. Large vision–language models incur substantially higher computational cost.

Limitations and Insights

ELVIS inherits inherent biases from its synthetic image generation process, which may limit direct generalization to real-world visual scenes. While the use of simplified object shapes and discretized, principle-specific patterns enables controlled and interpretable experimentation, it does not fully capture the richness and ambiguity of natural visual cognition. The substantial variance in model performance across different Gestalt principles further suggests that current approaches rely on uneven perceptual and reasoning capabilities, highlighting opportunities for deeper integration of symbolic reasoning with more robust perceptual models. A key limitation of ELVIS is that all

scenes are constructed in a 2D spatial layout, where object positions are explicitly defined and unambiguous. In contrast, real-world images represent 2D projections of a 3D environment, where spatial relationships cannot be reliably inferred from image-plane coordinates alone. Depth information plays a critical role in many Gestalt principles, such as proximity and symmetry, yet is largely absent from standard RGB imagery. As a result, grouping based solely on 2D cues can be fundamentally ambiguous in real-world scenes. To meaningfully extend Gestalt-based reasoning to natural images, incorporating depth-aware representations—either through RGB-D data, multi-view geometry, or learned depth estimation—becomes essential.

Finally, ELVIS provides explicit ground-truth grouping annotations, enabling supervised training of grouping models aligned with specific Gestalt principles. In real-world scenes, however, group boundaries are often subjective and context-dependent, making consistent annotation significantly more challenging. Perspective distortions further complicate the interpretation of object size and spatial relationships, and the absence of reliable depth cues exacerbates annotation ambiguity. These challenges underscore the gap between controlled synthetic benchmarks and real-world visual reasoning, and point toward future work on weakly supervised, probabilistic, or human-in-the-loop grouping frameworks for natural images.

Conclusion and Future Work

We introduced the Gestalt Vision (ELVIS) benchmark, designed to evaluate visual reasoning systems on five core Gestalt principles: Proximity, Similarity, Closure, Continuity, and Symmetry. ELVIS systematically varies object- and group-level properties such as color, shape, size, group number, and group size, requiring models to move beyond object recognition toward structured relational reasoning. Our evaluation shows that purely neural baselines remain close to chance and show little sensitivity to concept relevance, while larger multimodal models such as InternVL3-78B achieve notable gains but still lack principle-specific generalization. GPT-5 achieves the strongest overall performance, reaching around 0.7 accuracy across several settings, yet it continues to struggle on some symmetry tasks.

Future work should focus on advancing visual reasoning frameworks that explicitly encode object- and group-level rules to overcome the reliance on statistical correlations observed in current systems. Extending ELVIS toward more naturalistic scenes and video-based tasks will further bridge the gap to real-world reasoning. Ultimately, ELVIS serves as both a diagnostic tool and a catalyst for building perceptual reasoning systems that integrate accurate perception with structured, concept-grounded inference.

References

Amizadeh S, Palangi H, Polozov A, Huang Y and Koishida K (2020) Neuro-symbolic visual reasoning: Disentangling ”Visual” from ”Reasoning”. In: *Proceedings of the International Conference on Machine Learning (ICML)*.

Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, Zhong M, Zhang Q, Zhu X, Lu L et al. (2024) Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 24185–24198.

Chollet F, Knoop M, Kamradt G, Landers B and Pinkard H (2025) Arc-agi-2: A new challenge for frontier ai reasoning systems. URL https://arxiv.org/abs/2505.11831.

Ellis WD (1999) *A Source Book of Gestalt Psychology*. Routledge.

He K, Gkioxari G, Dollár P and Girshick R (2017) Mask r-cnn. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Helff L, Stammer W, Shindo H, Dhami DS and Kersting K (2025) V-lol: A diagnostic dataset for visual logical learning. *Journal of Data-centric Machine Learning Research* .

Hu S, Ma Y, Liu X, Wei Y and Bai S (2021) Stratified rule-aware network for abstract visual reasoning. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Hua T and Kunda M (2020) Modeling gestalt visual reasoning on raven's progressive matrices using generative image inpainting techniques. In: *Proceedings of the 42th Annual Meeting of the Cognitive Science Society (CogSci)*.

Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Zitnick CL and Girshick R (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim B, Reif E, Wattenberg M, Bengio S and Mozer MC (2021) Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior* .

Kim D and Kim J (2023) Vision transformer compression and architecture exploration with efficient embedding space search. In: Wang L, Gall J, Chin TJ, Sato I and Chellappa R (eds.) *Computer Vision – ACCV 2022*. Cham: Springer Nature Switzerland. ISBN 978-3-031-26313-2, pp. 524–540.

Koffka K (1935) *Principles of Gestalt Psychology*. Harcourt, Brace & World.

Li B, Zhang Y, Guo D, Zhang R, Li F, Zhang H, Zhang K, Li Y, Liu Z and Li C (2024) Llava-onevision: Easy visual task transfer.

Locatello F, Weissenborn D, Unterthiner T, Mahendran A, Heigold G, Uszkoreit J, Kipf T, Dinh L, Dieng AB and Gelly S (2020) Object-centric learning with slot attention. In: *Advances in Neural Information Processing Systems*.

Lörincz A, Fóthi Á, Rahman BO and Varga V (2017) Deep gestalt reasoning model: Interpreting electrophysiological signals related to cognition. In: *IEEE International Conference on Computer Vision Workshops (ICCV Workshop)*.

Mao J, Gan C, Kohli P, Tenenbaum JB and Wu J (2019) The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Müller H and Holzinger A (2021) Kandinsky patterns. *Artificial Intelligence (AIJ)* .

Nie W, Yu Z, Mao L, Patel AB, Zhu Y and Anandkumar A (2020) Bongard-logo: A new benchmark for human-level concept learning and reasoning. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI (2025) Introducing GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-08-29.

Palmer SE (1999) *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Raven JC and Court JH (1998) *Raven's progressive matrices and vocabulary scales*. Oxford: Oxford Psychologists Press.

Ruchkin DS (1971) Pattern recognition. m. bongard , joseph k. hawkins , theodore cheron. *The Quarterly Review of Biology* 46(4): 455–456. DOI:10.1086/407078. URL https://doi.org/10.1086/407078.

Sellars RW (1912) Is there a cognitive relation? *The Journal of Philosophy, Psychology and Scientific Methods* 9(9): 225–232.

Sha J, Shindo H, Kersting K and Dhami DS (2024) Neuro-symbolic predicate invention: Learning relational concepts from visual scenes. *Neurosymbolic Artificial Intelligence* .

Sha J, Shindo H, Kersting K and Dhami DS (2025) Gestalt vision: A dataset for evaluating gestalt principles in visual perception. In: H Gilpin L, Giunchiglia E, Hitzler P and van Krieken E (eds.) *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, *Proceedings of Machine Learning Research*, volume 284. PMLR, pp. 873–890. URL https://proceedings.mlr.press/v284/sha25a.html.

Shindo H, Pfanschilling V, Dhami DS and Kersting K (2023) αilp: thinking visual scenes as differentiable logic programs. *Machine Learning (MLJ)* .

Shindo H, Pfanschilling V, Dhami DS and Kersting K (2024) Learning differentiable logic programs for abstract visual reasoning. *Machine Learning (MLJ)* .

Tan H and Bansal M (2019) LXMERT: learning cross-modality encoder representations from transformers. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wertheimer M (1938) Laws of organization in perceptual forms. In: *A Source Book of Gestalt Psychology*.

Wightman R (2019) Pytorch image models.

Wu B, Xu C, Dai X, Wan A, Zhang P, Yan Z, Tomizuka M, Gonzalez J, Keutzer K and Vajda P (2020) Visual transformers: Token-based image representation and processing for computer vision.

Wüst A, Tobiasch T, Helff L, Ibs I, Stammer W, Dhami DS, Rothkopf CA and Kersting K (2025) Bongard in wonderland: Visual puzzles that still make ai go mad? In: *Forty-second International Conference on Machine Learning*.

Yi K, Gan C, Li Y, Kohli P, Wu J, Torralba A and Tenenbaum JB (2020) Clevrer: Collision events for video representation and reasoning. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yi K, Wu J, Gan C, Torralba A, Kohli P and Tenenbaum J (2018) Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In: *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Zhang Y, Soydaner D, Behrad F, Koßmann L and Wagemans J (2024) Investigating the gestalt principle of closure in deep convolutional neural networks. In: *32nd European Symposium on Artificial Neural Networks (ESANN)*.

Concept Coverage Analysis

Table 7 summarizes how different logical concepts are distributed across task categories in the benchmark. Each category is grounded in one of the Gestalt principles, and the presence or absence of object- and group-level concepts is marked. The covered concepts include fundamental visual attributes (color, shape, size), structural properties (count, background, overlap), and grouping information (principle, group number).

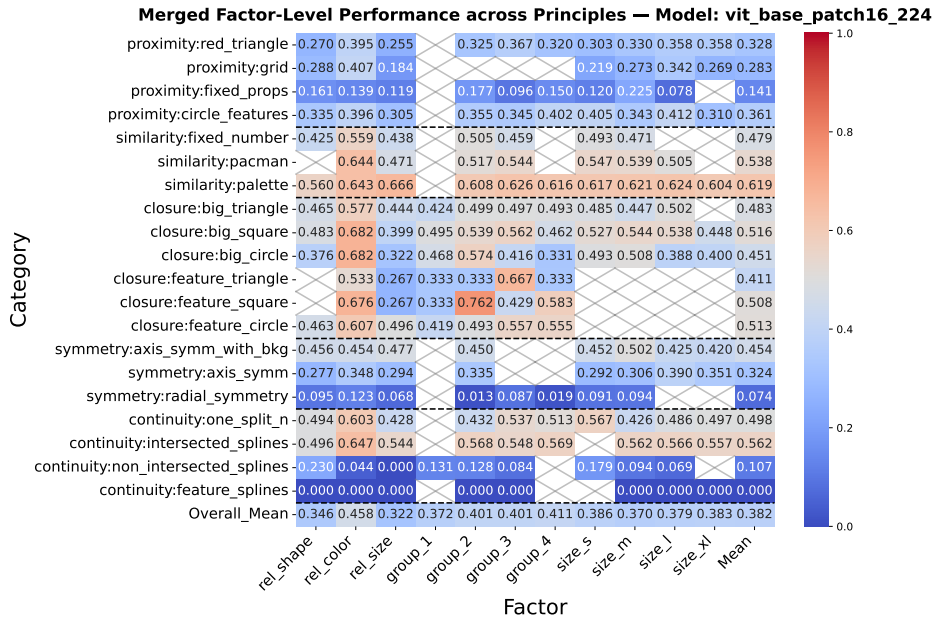
Most categories involve core visual attributes such as color, shape, size, and group number, while certain categories incorporate additional dimensions. For example, some of the similarity and symmetry tasks require reasoning over object count. Some of the proximity and symmetry patterns involve overlap features. This systematic coverage ensures that the benchmark spans both simple attribute-level reasoning and more complex multi-concept integration across Gestalt principles.

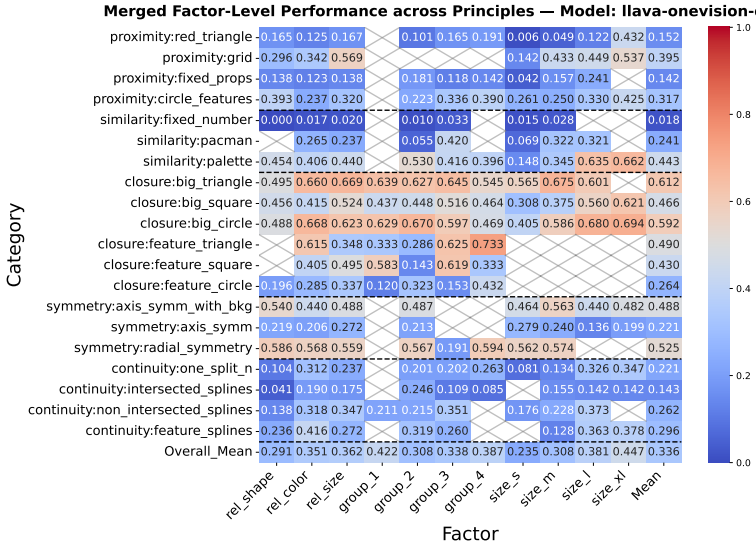
Table 7. Coverage of logical concepts across task categories. Column abbreviations: Col = color, Shp = shape, Cnt = count, Siz = size, Bkg = background, Ovl = overlap, Grp = group number. Fill cells with ✓or ✗.

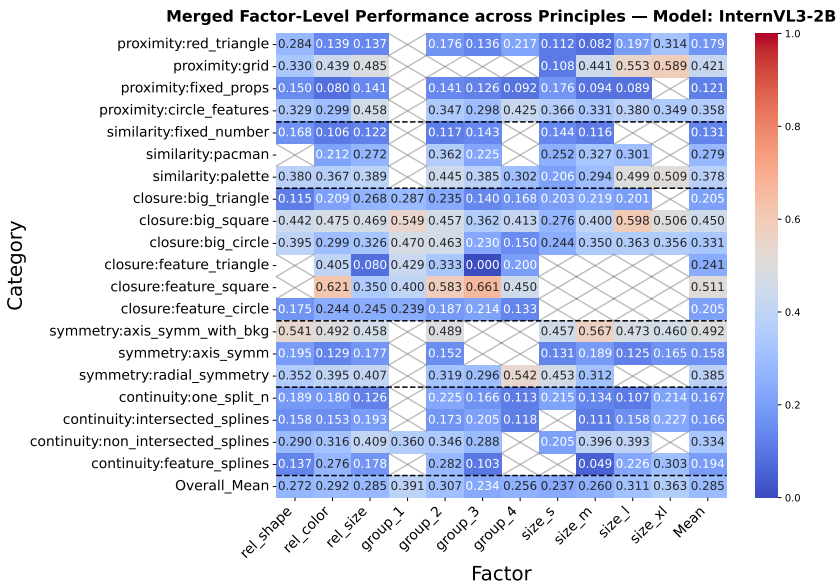
Category	Principle	Col	Shp	Cnt	Siz	Bkg	Ovl	Grp
Red Triangle	Proximity	✓	✓	✗	✓	✗	✗	✓
Grid	Proximity	✓	✓	✗	✓	✗	✗	✓
Fixed Props	Proximity	✓	✓	✗	✓	✗	✗	✓
Circle Features	Proximity	✓	✓	✗	✓	✓	✓	✓
Fixed Number	Similarity	✓	✓	✓	✓	✗	✗	✓
Pacman	Similarity	✓	✗	✓	✓	✗	✗	✓
Palette	Similarity	✓	✓	✗	✓	✗	✗	✓
Big Triangle	Closure	✓	✓	✗	✓	✗	✗	✓
Big Square	Closure	✓	✓	✗	✓	✗	✗	✓
Big Circle	Closure	✓	✓	✗	✓	✗	✗	✓
Feature Square	Closure	✓	✗	✗	✓	✗	✗	✓
Feature Circle	Closure	✓	✓	✗	✓	✗	✗	✓
Feature Triangle	Closure	✓	✗	✗	✓	✗	✗	✓
One Spline N	Continuity	✓	✓	✗	✓	✗	✗	✓
Intersected Splines	Continuity	✓	✓	✗	✓	✗	✗	✓
No Touching Splines	Continuity	✓	✓	✗	✓	✗	✗	✓
Overlap Splines	Continuity	✓	✓	✗	✓	✗	✓	✓
Radial Symmetry	Symmetry	✓	✓	✗	✓	✗	✗	✓
Axis Symmetry	Symmetry	✓	✓	✓	✓	✗	✗	✓
Axis Symmetry with Bkg	Symmetry	✓	✓	✓	✓	✓	✓	✓

Concept-wise Performance per Principle

Figure 11 reports the factor-level performance per category across Gestalt principles under the five baseline models. Each sub-figure corresponds to one model and separates results by concept dimension on both object level (color, shape, size) and group level (group size, group number).




Merged Factor-Level Performance across Principles — Model: llava-onevision-qwen2-7b
Category
proximity:red_triangle 0.165 0.125 0.167 0.101 0.165 0.191 0.006 0.049 0.122 0.432 0.152
proximity:grid 0.296 0.342 0.569 0.142 0.433 0.449 0.537 0.395
proximity:fixed_props 0.138 0.123 0.138 0.181 0.118 0.142 0.042 0.157 0.241 0.142
proximity:circle_features 0.393 0.237 0.320 0.223 0.336 0.390 0.261 0.250 0.330 0.425 0.317
similarity:fixed_number 0.000 0.017 0.020 0.010 0.033 0.015 0.028 0.018
similarity:pacman 0.265 0.237 0.055 0.420 0.069 0.322 0.321 0.241
similarity:palette 0.454 0.406 0.440 0.530 0.416 0.396 0.148 0.345 0.635 0.662 0.443
closure:big_triangle 0.495 0.660 0.669 0.639 0.627 0.645 0.545 0.565 0.675 0.601 0.612
closure:big_square 0.456 0.415 0.524 0.437 0.448 0.516 0.464 0.308 0.375 0.560 0.621 0.466
closure:big_circle 0.488 0.668 0.623 0.629 0.670 0.597 0.469 0.405 0.586 0.680 0.694 0.592
closure:feature_triangle 0.615 0.348 0.333 0.286 0.625 0.733 0.490
closure:feature_square 0.405 0.495 0.583 0.143 0.619 0.333 0.430
closure:feature_circle 0.196 0.285 0.337 0.120 0.323 0.153 0.432 0.264
symmetry:axis_symm_with_bkg 0.540 0.440 0.488 0.487 0.464 0.563 0.440 0.482 0.488
symmetry:axis_symm 0.219 0.206 0.272 0.213 0.279 0.240 0.136 0.199 0.221
symmetry:radial_symmetry 0.586 0.568 0.559 0.567 0.191 0.594 0.562 0.574 0.525
continuity:one_split_n 0.104 0.312 0.237 0.201 0.202 0.263 0.081 0.134 0.326 0.347 0.221
continuity:intersected_splines 0.041 0.190 0.175 0.246 0.109 0.085 0.155 0.142 0.142 0.143
continuity:non_intersected_splines 0.138 0.318 0.347 0.211 0.215 0.351 0.176 0.228 0.373 0.262
continuity:feature_splines 0.236 0.416 0.272 0.319 0.260 0.128 0.363 0.378 0.296
Overall_Mean 0.291 0.351 0.362 0.422 0.308 0.338 0.387 0.235 0.308 0.381 0.447 0.336
rel_shape rel_color rel_size group_1 group_2 group_3 group_4 size_s size_m size_l size_xl Mean
Factor
1.0 0.8 0.6 0.4 0.2 0.0


Merged Factor-Level Performance across Principles — Model: InternVL3-2B
Category
proximity:red_triangle 0.284 0.139 0.137 0.176 0.136 0.217 0.112 0.082 0.197 0.314 0.179
proximity:grid 0.330 0.439 0.485 0.108 0.441 0.553 0.589 0.421
proximity:fixed_props 0.150 0.080 0.141 0.141 0.126 0.092 0.176 0.094 0.089 0.121
proximity:circle_features 0.329 0.299 0.458 0.347 0.298 0.425 0.366 0.331 0.380 0.349 0.358
similarity:fixed_number 0.168 0.106 0.122 0.117 0.143 0.144 0.116 0.131
similarity:pacman 0.212 0.272 0.362 0.225 0.252 0.327 0.301 0.279
similarity:palette 0.380 0.367 0.389 0.445 0.385 0.302 0.206 0.294 0.499 0.509 0.378
closure:big_triangle 0.115 0.209 0.268 0.287 0.235 0.140 0.168 0.203 0.219 0.201 0.205
closure:big_square 0.442 0.475 0.469 0.549 0.457 0.362 0.413 0.276 0.400 0.598 0.506 0.450
closure:big_circle 0.395 0.299 0.326 0.470 0.463 0.230 0.150 0.244 0.350 0.363 0.356 0.331
closure:feature_triangle 0.405 0.080 0.429 0.333 0.000 0.200 0.241
closure:feature_square 0.621 0.350 0.400 0.583 0.661 0.450 0.511
closure:feature_circle 0.175 0.244 0.245 0.239 0.187 0.214 0.133 0.205
symmetry:axis_symm_with_bkg 0.541 0.492 0.458 0.489 0.457 0.567 0.473 0.460 0.492
symmetry:axis_symm 0.195 0.129 0.177 0.152 0.131 0.189 0.125 0.165 0.158
symmetry:radial_symmetry 0.352 0.395 0.407 0.319 0.296 0.542 0.453 0.312 0.385
continuity:one_split_n 0.189 0.180 0.126 0.225 0.166 0.113 0.215 0.134 0.107 0.214 0.167
continuity:intersected_splines 0.158 0.153 0.193 0.173 0.205 0.118 0.111 0.158 0.227 0.166
continuity:non_intersected_splines 0.290 0.316 0.409 0.360 0.346 0.288 0.205 0.396 0.393 0.334
continuity:feature_splines 0.137 0.276 0.178 0.282 0.103 0.049 0.226 0.303 0.194
Overall_Mean 0.272 0.292 0.285 0.391 0.307 0.234 0.256 0.237 0.260 0.311 0.363 0.285
rel_shape rel_color rel_size group_1 group_2 group_3 group_4 size_s size_m size_l size_xl Mean
Factor
1.0 0.8 0.6 0.4 0.2 0.0

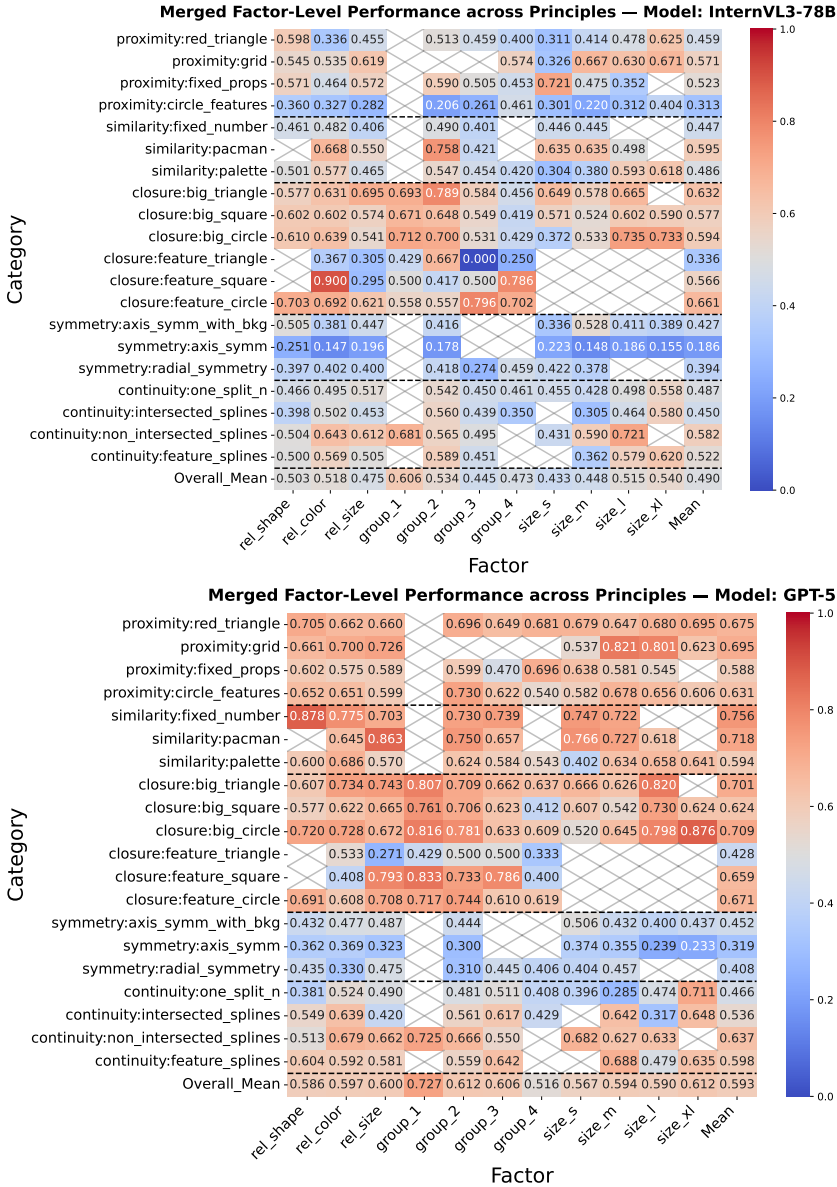


Figure 11. Factor-Level Performance per Category Across Gestalt Principles. Heatmaps show mean F1 scores for different experimental factors across task categories for all evaluated models. Empty cells indicate factors not manipulated in that category. Mean rows summarize category-level or cross-principle performance. Color scale ranges from 0 (blue) to 1 (red), with red indicating better performance.

Effect of Image Resolution

As a control experiment, we examine whether input resolution contributes to the performance limitations observed in large vision–language models. We evaluate InternVL3-78B at two commonly used resolutions, 224×224 and 448×448, corresponding to the pretraining settings of our baseline ViT models and InternVL3-78B itself. This comparison allows us to isolate the role of low-level perceptual fidelity without altering any other component of the pipeline.

As shown in Table 8, accuracy remains effectively unchanged across the two resolutions, and F1 exhibits only small increases at 448×448 for most principles. The sole exception is *symmetry*, where higher resolution slightly degrades performance, suggesting that additional visual detail does not help—and may even exacerbate—difficulties in capturing axis-based structural relations. Overall, these results confirm that resolution is not a primary bottleneck in ELVIS: the remaining gaps stem from challenges in structured perception and grouping rather than from insufficient pixel-level detail.

Table 8. Performance of InternVL3-78B at two resolutions. Accuracy is nearly identical across settings, while F1 at 448×448 is slightly higher, indicating resolution is not the main factor behind the errors.

Principle	Accuracy		F1 Score	
	224×224	448×448	224×224	448×448
Proximity	0.61 ± 0.17	0.61 ± 0.19	0.41 ± 0.35	0.44 ± 0.35
Similarity	0.61 ± 0.21	0.61 ± 0.20	0.46 ± 0.37	0.48 ± 0.36
Closure	0.73 ± 0.20	0.74 ± 0.20	0.59 ± 0.37	0.60 ± 0.36
Symmetry	0.62 ± 0.18	0.52 ± 0.16	0.45 ± 0.36	0.35 ± 0.30
Continuity	0.65 ± 0.18	0.65 ± 0.18	0.51 ± 0.33	0.51 ± 0.33

VLMs Prompts

This appendix provides the full prompting templates used for all VLMs (LLaVA-OneVision, InternVL3, GPT-5) evaluated in this work. As described in the main text, models receive (i) a brief task description, (ii) six labeled demonstration examples, and (iii) six unlabeled test examples for prediction. Each task is evaluated independently, and the prompts follow a unified structure across all models.

All prompts follow the format below:

> **Rule Induction:**
>
> You are an AI reasoning about visual patterns using Gestalt principles. Principle under consideration: *give the task principle*.
>
> The positive examples are image 01, image 02, image 03. The negative examples are image 04, image 05, image 06.
>
> Based on the Positive and Negative examples, infer the logic rules that distinguishes them. Output ONLY the rules.
>
> **Prediction Task:**
>
> Using the following reasoning rules: *the logic rules returned by the rule induction step*. Classify this image as Positive or Negative. Only answer with positive or negative for each image.

A visual depiction of all images (training and test) is passed to the VLMs directly through their native multi-image input interface. GPT-5 receives the images as base64-encoded attachments.

Task Examples

For each Gestalt principle in ELVIS, we present one or two representative task categories to illustrate the underlying design. The category names serve as intuitive references, but they do not always reflect the full range of variations. Due to controlled perturbations, some task variants may differ significantly from their original category name.

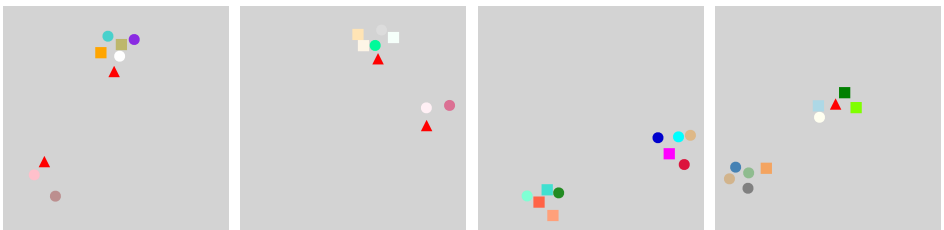
For instance, the category `Red Triangle` is initially designed around the idea that each group contains one red triangle. However, certain variations derived from this category may disregard color in the rule, resulting in tasks where the correct answer is determined solely by the presence of a triangle—regardless of its color. These variants are still formally associated with the `Red Triangle` category, though their governing logic differs. Other categories follow the same behavior.

Proximity: Red Triangle

The pattern `Red Triangle` follows the Gestalt principle of proximity. The base pattern is structured with multiple object groups, where each group consists of at least one red triangle and several smaller ones placed closely together.

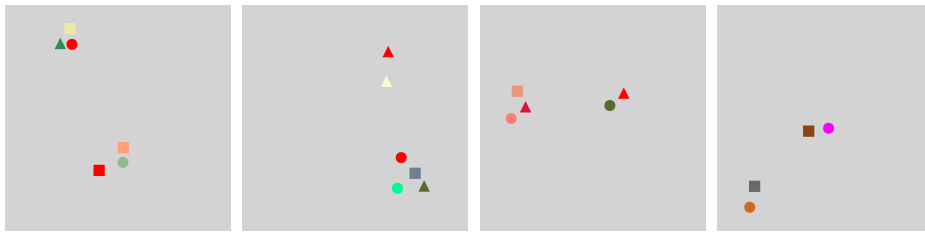
Fig. 12 presents a task where the rule is defined by *color* and *shape*. In the positive pattern, each group contains at least one object with red color and triangle shape, with the rest being random properties.

Fig. 13 illustrates another task variation, incorporating *color* only. In the positive pattern, each group contains at least one red object; the shape of the red object is randomly determined.



(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 12. Red Triangle: Each proximity group has at least one red triangle.



(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

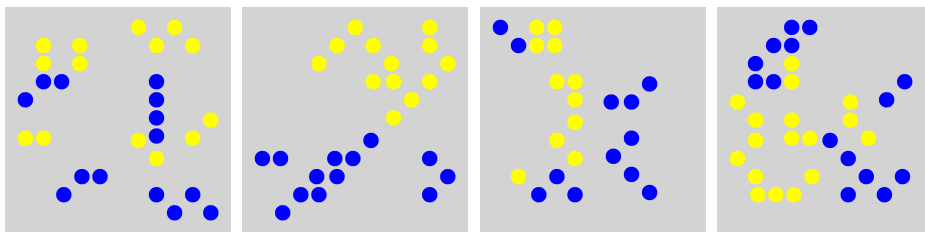
Figure 13. Red Triangle: Each proximity group has at least one red object.

Similarity: Fixed Number

The category `Fixed Number` is based on the Gestalt principle of similarity. The base pattern consists of an equal number of objects in different colors, with up to four color variations. Additionally, object size and shape can vary to introduce further task variations.

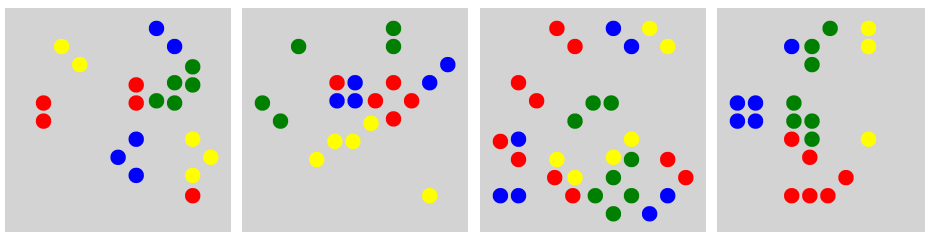
Fig. 14 illustrates a task where the rule involves counting objects of two colors.

Fig. 15 presents a variation where the task requires counting objects among four colors.



(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 14. Fixed Number: Same amount of yellow circles and blue circles.

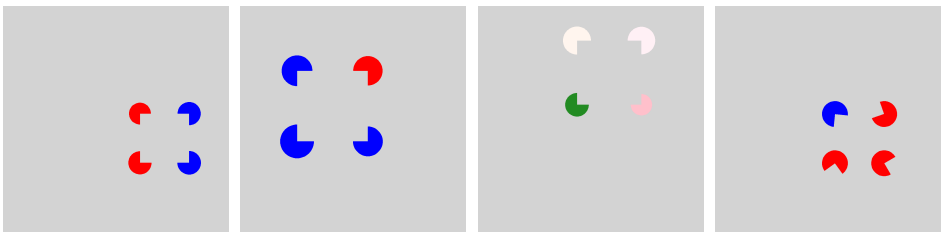


(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 15. Fixed Number: Same amount of red, yellow, blue, and green circles.

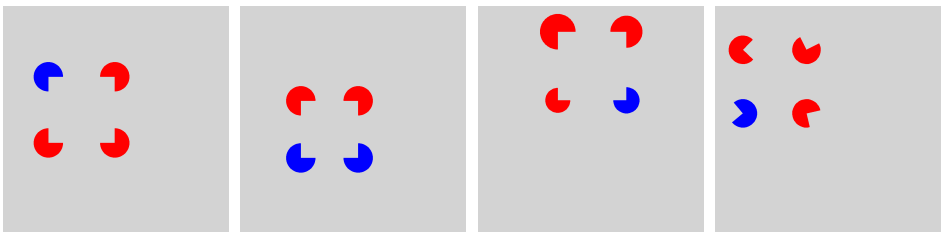
Closure: Feature Square

The category `Feature Square` follows the Gestalt principle of closure. Its base pattern consists of four 3/4 circles arranged to outline a square. Fig. 16 illustrates a task where object colors are limited to red or blue. Fig. 17 presents a variation where all circles are of equal size. Each task includes a counterfactual pattern that disrupts closure while maintaining all other rules.



(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 16. Feature Square: Closure square, obj color is either red or blue



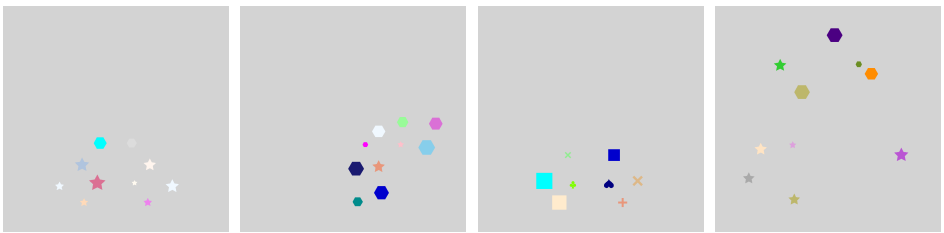
(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 17. Feature Square: Closure square, all objects have same size.

Symmetry: Axis Symmetry

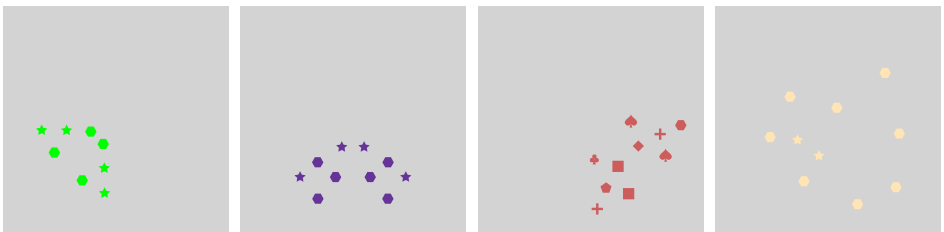
The category `axis sys` is based on the Gestalt principle of symmetry. Its base pattern places a random axis with objects arranged symmetrically around it.

Fig. 18 shows a task where object shapes are symmetric along the axis, with colors and sizes assigned randomly. Fig. 19 shows a variant where shapes remain symmetric but all objects share the same color and size.



(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 18. Axis Symmetry: Symmetry shape, random color and size



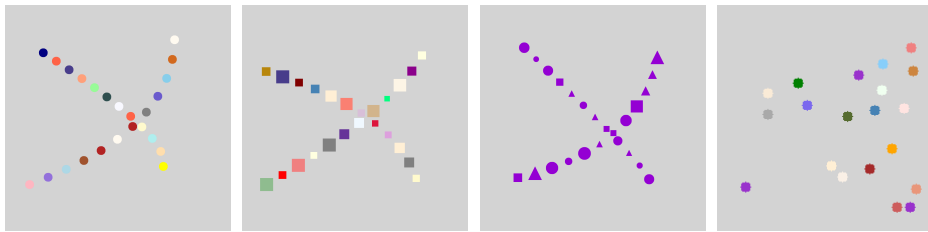
(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 19. Axis Symmetry: Symmetry shape, color, and size

Continuity: Intersected Splines

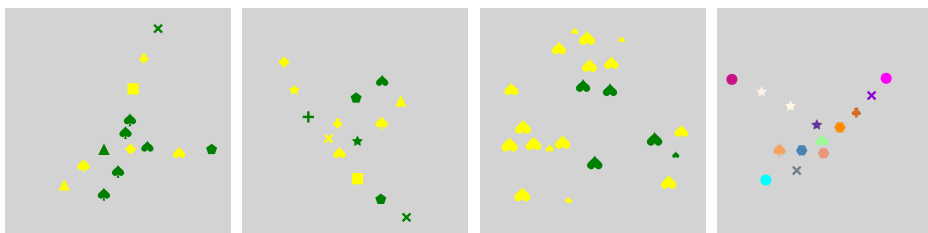
The category `Intersected Splines` follows the Gestalt principle of continuity. Its base pattern consists of n intersecting splines formed by small objects.

Fig. 20 illustrates a task where all objects share the same shape. Fig. 21 presents a variation where both the colors and shapes of the objects are identical.



(a) Positive **(b)** Positive **(c)** [Negative (REVISED)] **(d)** Negative

Figure 20. Splines: Each spline is consists of same shape of objects.



(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 21. Intersected Splines: Two splines of objects. The color of the objects can be either yellow or green.