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

Journal Title
XX(X):1–23
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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

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 neuro-symbolic 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 G**e**sta**l**t **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.

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 color distributions, and combinations across different principles. This enables the construction of thousands 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 introduce the Gestalt Vision Benchmark (ELVIS)*, a large scale dataset that systematically covers object level and group level properties, as well as their combinations across multiple Gestalt principles.

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

2. We design the number of tasks per principle up-to thousands, offering broader coverage and improved statistical reliability for evaluation.
3. We evaluate and analyze multiple baseline models on the dataset, highlighting both the progress and limitations of current approaches in capturing perceptual grouping and reasoning.
4. We release the dataset and code to serve as a comprehensive resource for advancing research in neuro-symbolic research community.

To this end, we proceed as follows. We start off with reviewing related work and then introduce our **G**es**al**t **Vis**ion (ELVIS) benchmark. Before concluding, we will present the results of our evaluation using ELVIS.

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. 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.

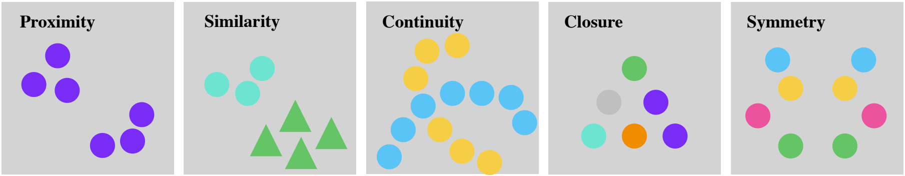


Figure 1. Gestalt Principles Supported by ELVIS. From left to right: **Proximity**: Objects that are spatially close to each other are perceived as a group. **Similarity**: Objects with common attributes, such as shape or color, are grouped together. **Continuity**: Objects with continue positions are 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 side determines a group.

More recently, several benchmarks have sought to test higher-level reasoning and abstraction. 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

G**e**s**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

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—an important challenge for neuro-symbolic 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.

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 neuro-symbolic 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—an ability central to

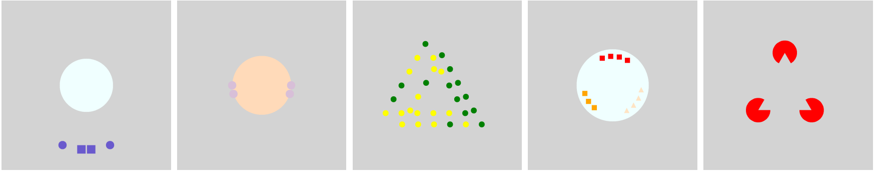


Figure 2. Geometric Feature Scenarios. Example patterns illustrating different geometric feature scenarios. **From left to right**: individual objects, object overlap, group overlap, nested shapes, and incomplete forms, designed to assess model perception under varied spatial configurations.

neuro-symbolic 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.

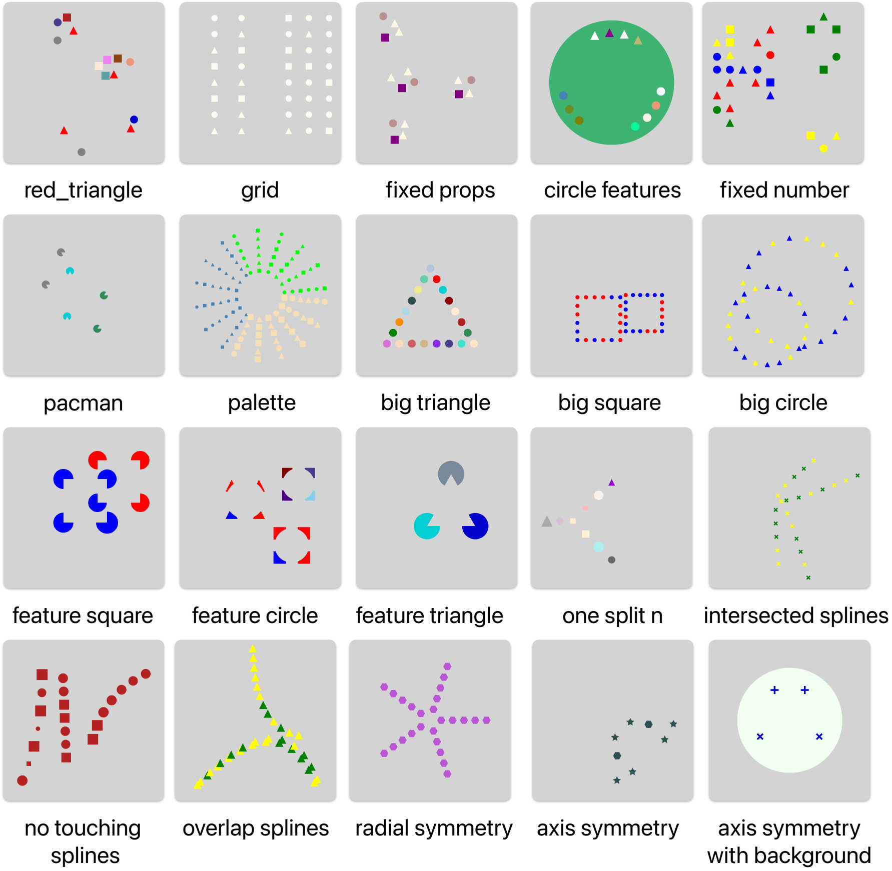
Varied Complexity. The number of objects in a scene ranges from a handful in simple settings to several hundred in complex ones. Regardless of density, each scene is designed to clearly embody a target Gestalt principle. Objects that participate in the same principle may still differ in shape, color, 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.



red_triangle

grid

fixed props

circle features

fixed number

pacman

palette

big triangle

big square

big circle

feature square

feature circle

feature triangle

one split n

intersected splines

no touching splines

overlap splines

radial symmetry

axis symmetry

axis symmetry with background

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.

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 6 in Appendix presents the detailed information for each category.

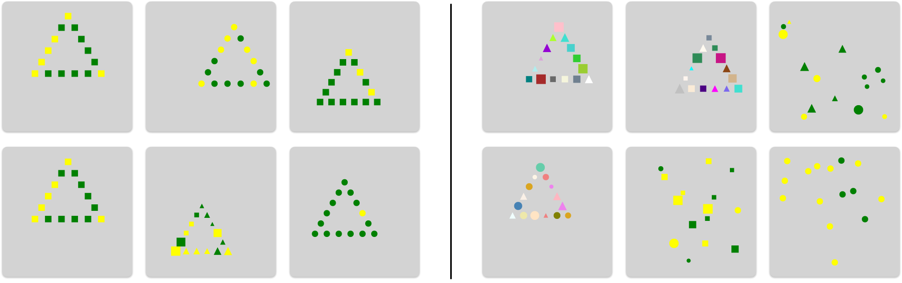


Figure 4. Task Example of ELVIS. **Left:** Positive patterns illustrating the Gestalt principle of closure, where objects collectively form a yellow-green triangle. **Right:** Negative patterns that partially adhere to the rule but violate key constraints, either by not matching the required color or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

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. Although some negative images may share superficial similarities with positive ones, they are guaranteed to break at least one essential constraint, making the task nontrivial and requiring more than low-level visual matching.

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.

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.

Table 2. 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

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.

Models were trained and evaluated independently for each task. Specifically, for a given task, a model was trained on its corresponding labeled examples and evaluated on its own held-out test set. This process was repeated separately for every task in the benchmark. Performance on each task is evaluated using two metrics: accuracy and F1 score. Accuracy is defined as the proportion of correct predictions among all predictions, providing a direct measure of how often the model outputs the right label. The F1 score complements this by combining precision and recall, offering a more balanced assessment that is especially important when the distribution of positive and negative examples is uneven. To obtain an overall measure of model capability, we report the mean and standard deviation of these metrics across all tasks in the benchmark.

Baseline Models

We evaluated four representative baselines, encompassing neural and VLM approaches. Table 2 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 across Gestalt Principles

Table 3 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 on symmetry, where both accuracy and F1 lag behind its other results, suggesting that certain spatial-relational cues remain challenging.

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

Table 3. 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	**0.62** ± 0.18	0.65 ± 0.18
	GPT-5	**0.69** ± 0.19	**0.71** ± 0.23	**0.77** ± 0.19	0.60 ± 0.18	**0.69** ± 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	**0.46** ± 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	**0.65** ± 0.29	**0.63** ± 0.35	**0.67** ± 0.33	0.40 ± 0.35	**0.55** ± 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	**0.51** ± 0.41	0.61 ± 0.39
	GPT-5	**0.65** ± 0.31	**0.66** ± 0.34	**0.76** ± 0.24	0.49 ± 0.42	**0.62** ± 0.38
Rec.	ViT-16-224	0.31 ± 0.35	**0.80** ± 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	**0.55** ± 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	**0.71** ± 0.29	0.66 ± 0.34	**0.65** ± 0.35	0.40 ± 0.38	**0.55** ± 0.40

models show principle-specific weaknesses and lack systematic compositionality across grouping cues.

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.

Effect of Image Resolution

We evaluate InternVL3-78B at two input sizes, 224 × 224 and 448 × 448, to examine whether higher resolution accounts for the observed errors. The choice of these two resolutions is motivated by the pretraining setups of our baseline models: ViT models are conventionally trained at 224 × 224, while InternVL3-78B itself is pretrained with 448 × 448 inputs. Thus, comparing these two settings provides a fair assessment of

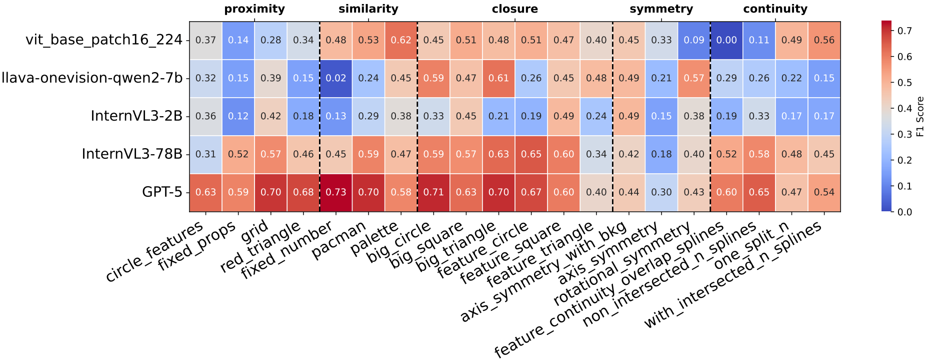


Figure 5. Average F1 score by Categories Over baseline models. The chart compares average F1 scores (y-axis) for proximity, similarity, closure, symmetry, continuity, and related categories (x-axis).

Table 4. 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

whether resolution alone influences performance. As shown in Table 4, accuracy remains nearly identical across both input sizes, while F1 shows a small but consistent gain at 448×448 for most principles. The only exception is symmetry, where both accuracy and F1 decrease at higher resolution, suggesting that increased detail may even amplify the weakness of the model in capturing axis-based structural relations. Overall, these results indicate that limited resolution is not the main bottleneck on this benchmark, and the remaining gap is more plausibly explained by the model's limited capacity for structured perception and reasoning.

Effect of Training Number

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

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 number of training images 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.

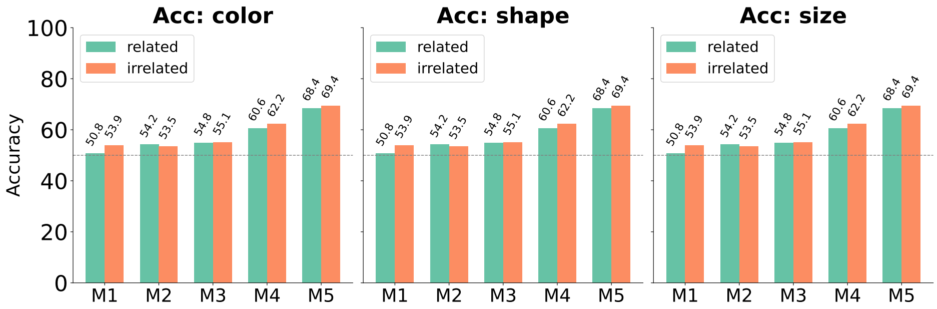


Figure 6. Average accuracy by concept relevance. Average performance of five baseline models across different task types. From M1 to M5: ViT, Llava-OneVision, InternVL3-2B, InternVL3-78B and GPT-5. Each subfigure shows results for one concept dimension (color, shape, size), comparing cases where the concept is related (used in the target rules) or irrelated (not used in the rules and thus irrelevant for distinguishing positives and negatives).

symmetry), reflecting weak but non-trivial generalization. In contrast, the hundred-shot model collapses during training, producing almost constant predictions that result in accuracy near 0.50 and F1 scores close to zero across all principles. This indicates that the model does not benefit from larger training sets but instead over-fits to spurious correlations, underscoring the difficulty of learning generalizable rules in this benchmark with a purely neural baseline.

Concept Level Analysis

Object Level Concepts Analysis Figure 6 shows the average accuracy across color, shape, and size, contrasting cases where the concept is related or irrelated to the task rules. Across all models and principles, the performance on related and irrelated tasks remains highly similar, with only marginal differences. The ViT baseline stays close to chance, and later models such as Llava-OneVision and InternVL3-2B show only minor improvements without developing clear sensitivity to concept relevance. InternVL3-78B achieves higher overall accuracy, while GPT-5 reaches around 0.7, yet both still

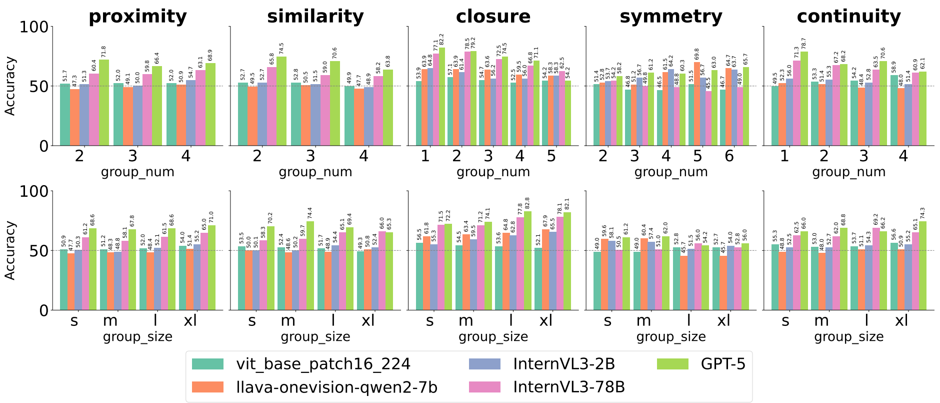


Figure 7. Baseline performance across gestalt groups for five principles. Each panel shows all baseline models for one principle.

display nearly identical trends for related and irrelated conditions. This consistency across settings indicates that none of the models systematically leverage concept-specific cues, underscoring the need for neuro-symbolic methods that enforce explicit grounding of color, shape, and size in reasoning.

Group Level Concepts Analysis Figure 7 compares performance across the group-level concepts *group number* and *group size*. Unlike the object-level concepts analysis, here the focus is on how accuracy changes with varying numbers of groups and with different group sizes.

A clear trend emerges for the closure and continuity principles: accuracy decreases as the group number increases, but improves with larger group sizes. Larger groups include more objects, making group-level features (like a line of objects) easier to identify, while fewer groups simplify the scene and make rules easier to detect. In contrast, tasks with many groups often require objects to align into lines or connected structures to exhibit closure or continuity, which substantially raises the difficulty. This reflects a complexity effect: models perform better on simpler scenes with fewer and larger groups, but their performance declines as visual complexity grows with many small groups.

Limitations and Insights

ELVIS inherently contains biases from synthetic image generation, potentially limiting generalizability to real-world scenarios. Additionally, simplified object shapes and discrete principle-based patterns, while facilitating controlled experimentation, might not fully capture the complexity of natural visual cognition. Models exhibiting high variance in performance across different Gestalt principles suggest opportunities for further optimization and deeper integration of symbolic reasoning with advanced perceptual models.

Conclusion and Future Work

We introduced the Gestalt Vision (ELVIS) benchmark, designed to evaluate neuro-symbolic 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 neuro-symbolic 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.

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* .

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)* .

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.

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* .

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.

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)*.

Table 6. 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 Coverage Analysis

Table 6 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.

Concept-wise Performance per Principle

Figure 8 reports the accuracy of the baseline models across all Gestalt principles, with performance broken down by concept relevance. Each sub-figure corresponds to one principle and separates results by concept dimension (color, shape, size).

The plots highlight that performance differs across principles and concept types. In some principles, such as similarity and continuity, accuracy increases substantially when the related concept is included in the target rules, while in others, like symmetry, models

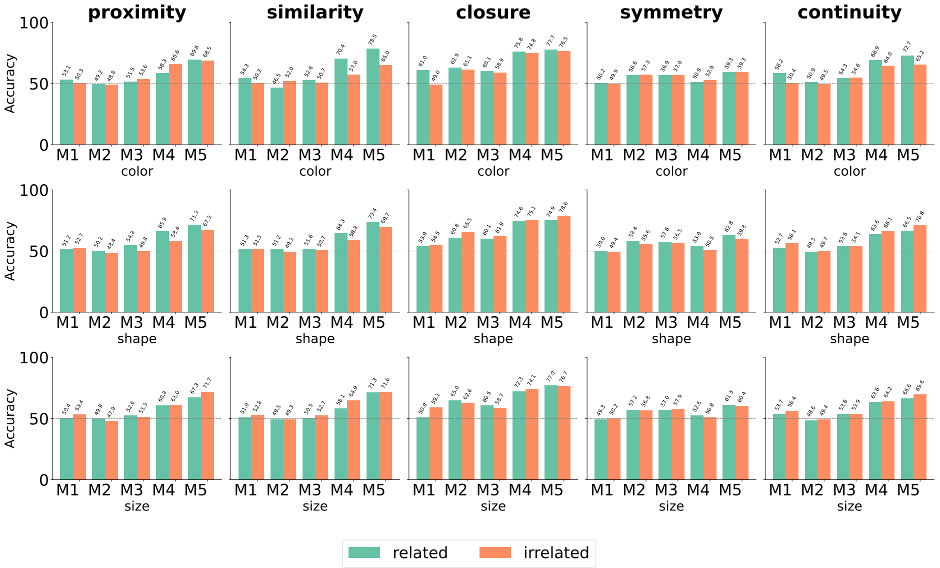


Figure 8. Principle-wise accuracy by concept relevance. Accuracy of five baseline models (ViT, Llava-OneVision-7B, InternVL3-2B, InternVL3-78B, GPT-5) across all Gestalt principles. Each sub-figure shows results for one principle, with performance broken down by concept dimension (color, shape, size) when the concept is related (used in the target rules) or irrelated (not used in the rules and thus irrelevant for distinguishing positives and negatives).

remain closer to chance regardless of concept condition. The comparison also shows how larger multi-modal models (InternVL3-78B, GPT-5) achieve higher scores across more settings, whereas smaller models (ViT-3, Llava-OneVion-7B, InternVL3-2B) tend to hover around baseline levels. Together, these results provide a detailed breakdown of how concept relevance interacts with each Gestalt principle within the benchmark.

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.

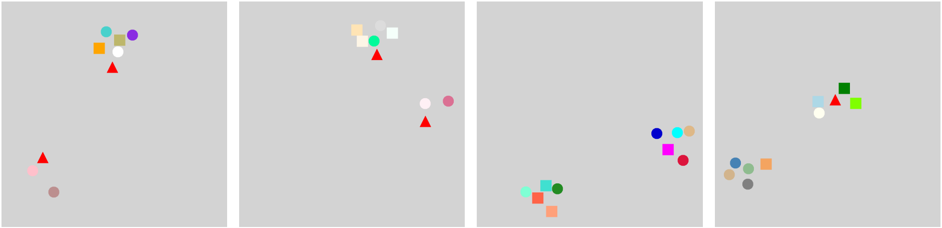
For instance, the category `Red Triangle` is initially designed around the idea that each group contains one red triangle. However, certain variations derived from this category may disregard color in the rule, resulting in tasks where the correct answer is determined solely by the presence of a triangle—regardless 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.

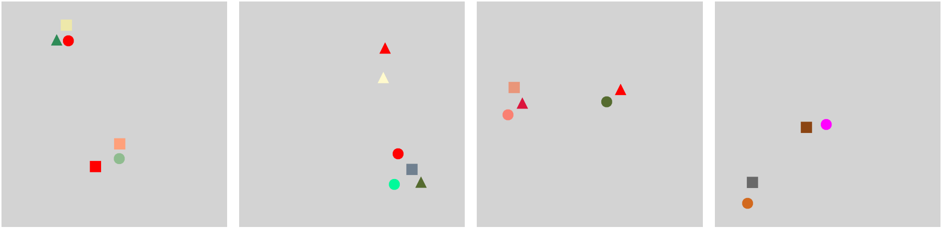
Fig. 9 presents a task where the rule is defined by *color* and *shape*. In the positive pattern, each group contains at least one object with red color and triangle shape, with the rest being random properties.

Fig. 10 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 9. Red Triangle: Each proximity group has at least one red triangle.



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

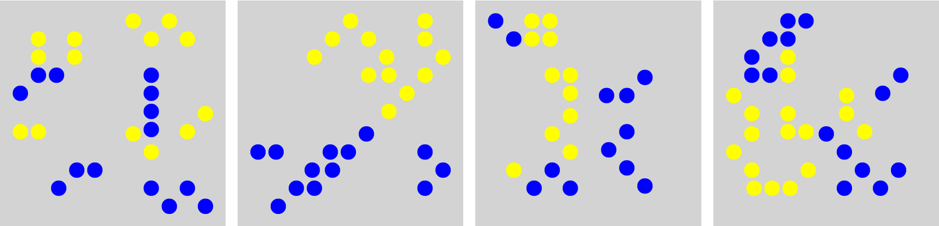
Figure 10. Red Triangle: Each proximity group has at least one red object.

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.

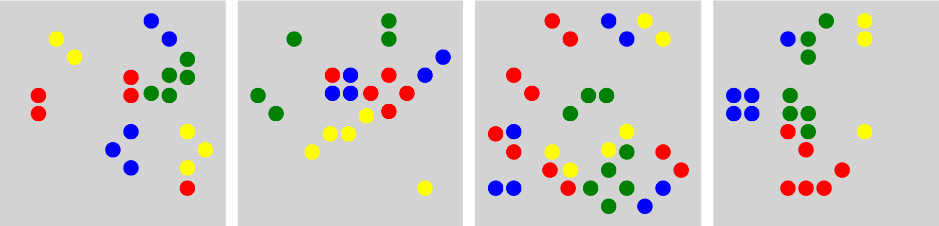
Fig. **??** illustrates a task where the rule involves counting objects of two colors.

Fig. 12 presents a variation where the task requires counting objects among four colors.



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

Figure 11. Fixed Number: Same amount of yellow circles and blue circles.

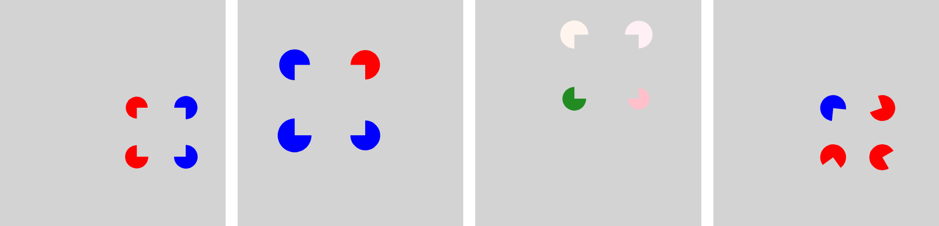


(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 12. Fixed Number: Same amount of red, yellow, blue, and green circles.

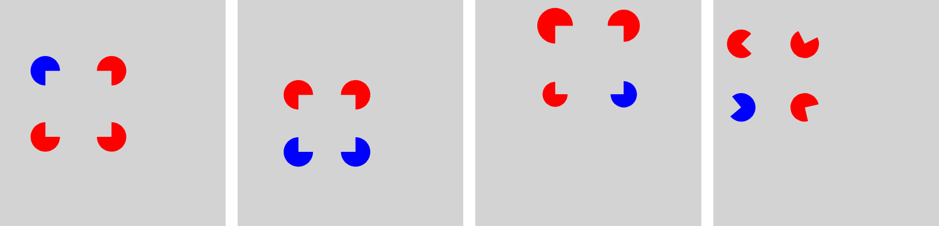
Closure: Feature Square

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. 13 illustrates a task where object colors are limited to red or blue. Fig. 14 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 13. Feature Square: Closure square, obj color is either red or blue



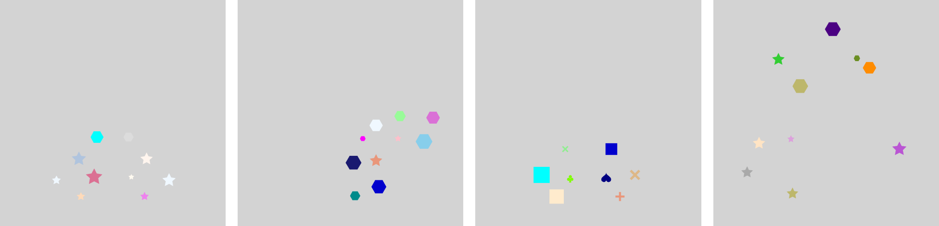
(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 14. Feature Square: Closure square, all objects have same size.

Symmetry: Axis Symmetry

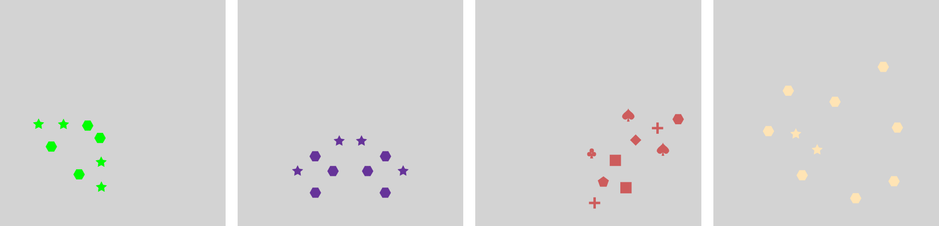
The category `axis sys` is based on the Gestalt principle of symmetry. Its base pattern places a random axis with objects arranged symmetrically around it.

Fig. 15 shows a task where object shapes are symmetric along the axis, with colors and sizes assigned randomly. Fig. 16 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 15. Axis Symmetry: Symmetry shape, random color and size



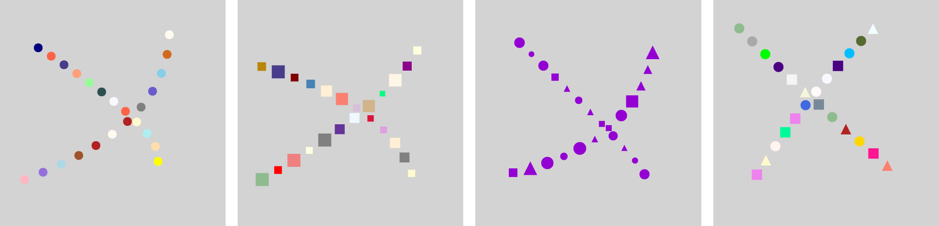
(a) Positive **(b)** Positive **(c)** Negative **(d)** Negative

Figure 16. Axis Symmetry: Symmetry shape, color, and size

Continuity: Intersected Splines

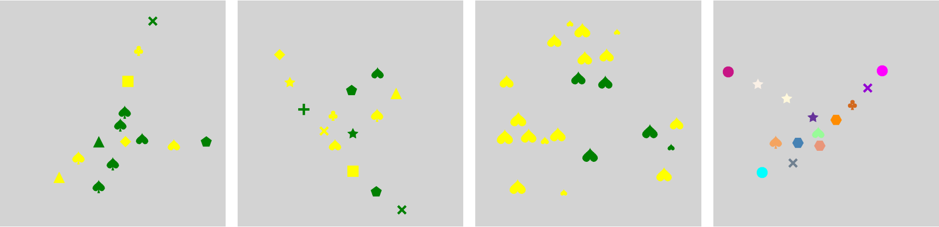
The category `Intersected Splines` follows the Gestalt principle of continuity. Its base pattern consists of n intersecting splines formed by small objects.

Fig. 17 illustrates a task where all objects share the same shape. Fig. 18 presents a variation where both the colors and shapes of the objects are identical.



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

Figure 17. Intersected Splines: Each spline is consists of same shape of objects.



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

Figure 18. Intersected Splines: Two splines of objects. The color of the objects can be either yellow or green.