# Birds of a Feather Reason Together: Gestalt Grouping Meets Neuro-Symbolic Inference

**Jingyuan Sha**[1] **Hikaru Shindo**[1] **Kristian Kersting**[1,2,3] **Devendra Singh Dhami**[4]

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

## Abstract

This paper introduces Gestalt Reasoning Machines (GRMs), a novel neuro-symbolic framework that integrates Gestalt principles to enhance reasoning models with perception capabilities similar to human cognition. Traditional models, which rely on large datasets and complex computations, often overlook the crucial human cognitive function of grouping, resulting in inefficiencies when dealing with abstract concepts. GRMs address this challenge by incorporating a grouping mechanism grounded in Gestalt principles, enabling the system to recognize and reason over complex visual patterns that are otherwise difficult to capture through object-level features alone. This grouping capability allows GRMs to identify higher-order structures and relational configurations that are essential for human-like reasoning. We demonstrate that GRMs achieve competitive or superior performance compared with purely neural baselines by leveraging logic-based reasoning infused with perceptual grouping cues, offering a more interpretable and cognitively aligned approach. Our contributions include the design of GRMs and the empirical validation of their effectiveness in visual reasoning tasks that demand structured perception.

## 1 INTRODUCTION

Human visual perception excels at organizing complex scenes into meaningful structures through perceptual grouping. *Gestalt principles*—such as proximity, similarity, and continuity—explain how individuals organize visual elements into coherent wholes rather than processing them as isolated components. These principles, rooted in psychology research [Koffka, 1935, Wertheimer, 1938, Palmer, 1999, Ellis, 1999], are fundamental to how humans efficiently parse and reason about visual scenes. As illustrated in Fig. 1 (right), humans naturally perceive objects based on their spatial arrangements and shared attributes, identifying patterns that enable structured understanding of complex visual environments.

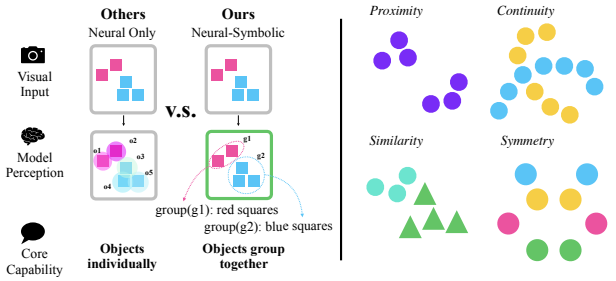



Figure 1: **Scene reasoning with Gestalt grouping. Left:** Comparison of model perception. Neural-only models process objects individually without structural organization, whereas GRM integrates perceptual grouping with symbolic reasoning, producing interpretable group-based representations (e.g., red vs. blue squares). **Right:** Examples of gestalt principles. Proximity, similarity, symmetry, and continuity. Each of these principles guide how objects are grouped into meaningful structures.

Current approaches to visual reasoning typically rely on scaling up neural models with massive datasets and parameters [Kojima et al., 2022, Huang and Chang, 2023, Cheng et al., 2024, Zhang et al., 2025]. However, these data-driven models struggle with abstract reasoning [Huang et al., 2024], particularly in Vision-Language Models (VLMs) where reasoning capabilities remain severely limited [Chen et al., 2025, Wüst et al., 2025]. Their reasoning often lacks grounding and becomes inconsistent when processing complex multi-object scenes [Fu et al., 2024, Majumdar et al., 2024, Zhang et al., 2024a], highlighting limitations in their ability to capture structured relationships.

Neuro-symbolic approaches offer a promising alternative by integrating symbolic reasoning with neural perception.

*Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence* (UAI 2026), PMLR 337:6125–6148.

These models have demonstrated strong performance on complex reasoning tasks requiring relational inference and numerical computation over visual inputs [Yi et al., 2018, Amizadeh et al., 2020, Manhaeve et al., 2021, Marra et al., 2024]. Differentiable rule learners [Evans and Grefenstette, 2018, Shindo et al., 2023, 2024b] have been successfully applied to visual reasoning challenges, discovering explicit rules to explain visual patterns through program induction [Shindo et al., 2024a, Sudhakaran et al., 2025]. However, existing neuro-symbolic approaches face a critical scalability bottleneck: they lack effective *grouping* mechanisms that are fundamental to human visual perception [Han et al., 2002, Xu and Chun, 2007, Thórisson, 2019]. When processing complex scenes, humans naturally organize objects according to perceptual cues like similarity and proximity, enabling efficient reasoning by reducing redundant relational computations. In contrast, most neuro-symbolic models rely on exhaustive pairwise relation generation, enumerating all possible object relationships. This approach becomes computationally prohibitive and brittle as scene complexity increases. This raises the central research question: *How can we endow neuro-symbolic models with human-like grouping mechanisms that enable efficient and robust reasoning over complex visual scenes?*

We propose the *Gestalt Reasoning Machine (GRM)*, a neuro-symbolic framework that explicitly incorporates perceptual grouping into visual reasoning. Unlike conventional object-centric approaches, GRM first organizes scene elements into structured groups based on spatial and attributional patterns (Fig. 1, left). These grouped entities serve as the foundation for symbolic reasoning, allowing the model to operate over scenes in a hierarchically structured and interpretable manner. By treating scenes as compositions of perceptual groups rather than collections of isolated objects, GRM captures higher-level regularities and solves complex visual problems that traditional approaches often fail to address. Given a complex scene with multiple objects, GRMs perform rule learning to identify and group objects that share underlying patterns, simultaneously producing structured perceptual representations and symbolic programs for reasoning. The resulting structured perception is then fed to the reasoning module, which efficiently infers solutions by operating over these coherent groups rather than individual objects.

Our work makes the following key contributions:

- We introduce Gestalt Reasoning Machines (GRMs)[1], the first neuro-symbolic framework that integrates perceptual grouping with symbolic rule learning, offering a robust and interpretable approach to complex visual scene understanding.
- We develop a scalable grouping mechanism that operates directly on visual input and seamlessly integrates with symbolic reasoning, enabling GRMs to scale effectively with increasing numbers of objects by reducing unnecessary relational complexity.
- We demonstrate that GRMs achieve competitive or superior performance on structured visual reasoning tasks while providing interpretable group-based rules, with the largest gains appearing when perceptual grouping reduces redundant object-level relations.

The remainder of this paper presents related work, details the GRM architecture and training procedure, provides a comprehensive experimental evaluation, and concludes with future research in neuro-symbolic visual reasoning.

## 2 RELATED WORK

Visual reasoning is a fundamental problem in machine learning research, leading to the development of various benchmarks [Antol et al., 2015, Johnson et al., 2017, Yi et al., 2020] and subsequent frameworks [Yi et al., 2018, Mao et al., 2019, Amizadeh et al., 2020, Hsu et al., 2023] focused on reasoning through symbolic programs and multi-modal transformers [Tan and Bansal, 2019]. These benchmarks primarily aim to answer queries expressed in natural language in conjunction with visual inputs. Our work evaluates GRMs, particularly emphasizing their ability to perform the grouping function based on Gestalt reasoning principles. Reinforcement learning has been used to enhance the reasoning capability of large Vision-Language Models (VLMs) [Liu et al., 2025, Tan et al., 2025, Zhai et al., 2024]. However, the resulting reasoning traces are not logically grounded in terms of objects [Sarch et al., 2025]. Consequently, these models struggle to perform structured perception with grouping. We aim to develop the foundation of structured perception using a neuro-symbolic approach.

Additionally, Abstract Visual Reasoning (AVR) explores the capability to apply previously acquired knowledge and techniques in completely new contexts, posing unique challenges for deep neural networks (DNNs) [Hu et al., 2021, Malkinski and Mandziuk, 2023, Camposampiero et al., 2023]. AVR methods have been primarily evaluated through simple abstract puzzles like Raven's progressive matrices [Raven and Court, 1998]. The Kandinsky patterns framework [Müller and Holzinger, 2021] provides a unique method for generating patterns with abstract objects, which we have expanded to address Gestalt reasoning tasks. To address these challenges, neuro-symbolic rule learning frameworks have been developed, emphasizing the learning of discrete rule structures via backpropagation [Evans and Grefenstette, 2018, Minervini et al., 2020, Shindo et al., 2023, 2024b, Zimmer et al., 2023, Sha et al., 2024]. These methodologies have predominantly been tested on visual arithmetic tasks or synthetic environments tailored for reasoning [Stammer et al., 2021]. With GRMs, we seek to

[1]We make our code publicly available at `https://github.com/ml-research/gestalt_reasoning_machine#`

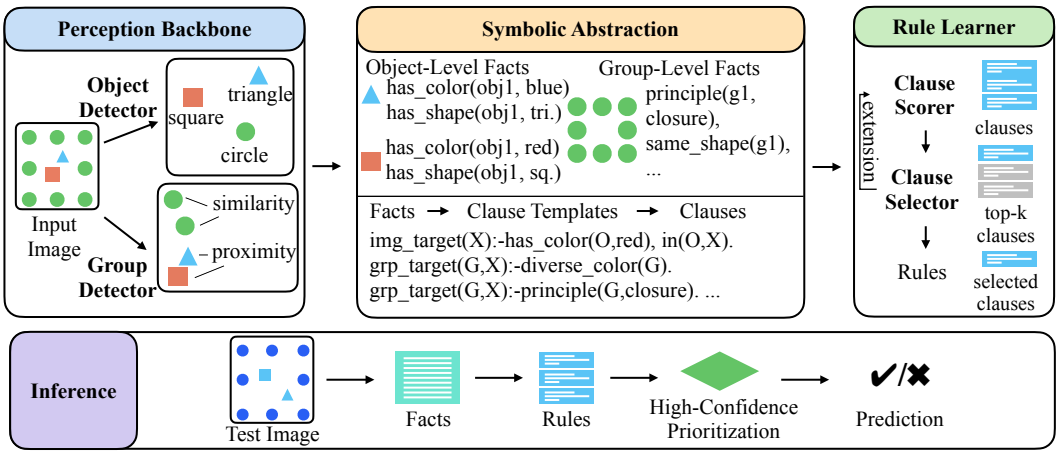



Figure 2: **Overview of the Gestalt Reasoning Machine (GRM) Pipeline.** An input image is first processed by a **Perception Backbone** that detects objects and identifies Gestalt-based group structures. Next, the **Symbolic Abstraction** module converts these perceptual features into a logical fact base and generates candidate clauses. The **Rule Learner** then performs a search over this symbolic space to discover a final set of *interpretable rules*. Finally, the **Inference** engine applies these rules to a new visual input, employing a *high-confidence prioritization strategy*. This ensures that predictions are based on highly certain, transparent rules whenever possible, falling back to a weighted aggregation of evidence from all relevant rules in more ambiguous cases.

bridge the gap between existing neuro-symbolic paradigms and human cognitive processes.

Gestalt reasoning has been extensively studied in psychology [Wertheimer, 1938, Koffka, 1935, Palmer, 1999, Ellis, 1999] and has also intersected significantly with machine learning and deep learning research [Lörincz et al., 2017, Hua and Kunda, 2020, Kim et al., 2021, Zhang et al., 2024b] with a strong emphasis on convolutional neural networks. GRMs represent the first neuro-symbolic framework that explicitly encodes the grouping function in Gestalt principles.

## 3 GESTALT REASONING MACHINES

The Gestalt Reasoning Machine (GRM) is a neuro-symbolic framework that integrates perceptual organization with logical reasoning. Inspired by human visual cognition, it processes raw images to detect objects and perceptual groups, abstracts them into object- and group-level facts, and applies a logic-based rule learner to derive interpretable decision functions (Fig. 2). By combining neural perception with symbolic abstraction and rule learning, GRM achieves flexible yet interpretable reasoning over complex visual patterns. The following subsections detail its components.

### 3.1 PERCEPTION BACKBONE

The perceptual backbone of GRM transforms raw visual input into structured representations suitable for symbolic reasoning. It operates at two levels: *object-level perception* and *group-level perception*. Importantly, the backbone is defined functionally rather than architecturally: any implementation that satisfies the required representational interface can be used, including neural models or symbolic procedures.

**Object-Level Perception.** Given an input image $I$, the system extracts a set of perceptual objects

$$\mathcal{O} = \{o_1, o_2, \ldots, o_n\},$$

where each object $o_i$ corresponds to a visually coherent region with spatial extent $M_i \subset \mathbb{Z}^2$. Each object maintains a dual representation:

$$o_i = (\phi_i^{\text{sym}}, \phi_i^{\text{neu}}),$$

where $\phi_i^{\text{sym}}$ encodes symbolic attributes (e.g., category, color, spatial position), and $\phi_i^{\text{neu}}$ captures continuous geometric or appearance features. The symbolic component supports logical manipulation, while the neural component preserves fine-grained perceptual detail.

The specific extraction mechanism is modular. In synthetic environments, object segmentation and symbolic attributes can be derived from ground-truth rendering metadata. In natural images, neural object detectors and depth estimators may be employed. GRM assumes only that object instances and their attributes can be reliably obtained, without constraining the underlying perception architecture.

**Group-Level Perception.** The goal of group-level perception is to construct a set of perceptual groups

$$\mathcal{G} = G(\mathcal{O}),$$

where each group $g \in \mathcal{G}$ is a subset of objects $g \subseteq \mathcal{O}$ representing a coherent perceptual unit. The grouping operator $G$

thus transforms object-level representations into structured higher-level abstractions.

The operator $G$ is defined abstractly and can be instantiated in different ways depending on the availability of supervision. When group annotations are available, grouping is formulated as a supervised affinity estimation problem. When supervision is absent or unreliable, grouping can be derived from explicit symbolic rules over relational predicates.

In the supervised setting, each object is associated with a ground-truth group label during training. Because the number of objects varies across scenes, the model must operate over variable-sized object sets. We therefore formulate grouping as a pairwise affinity estimation: for each object pair $(o_i, o_j)$, the system computes

$$s(o_i, o_j \mid \mathcal{O}),$$

which represents the confidence that the two objects belong to the same group. $s(\cdot)$ denotes a learned affinity score between object pairs. Accordingly, the affinity function incorporates permutation-invariant contextual information over the full object set $\mathcal{O}$, ensuring robustness to varying object counts while remaining sensitive to global scene structure. The resulting affinity scores are then composed into candidate groups.

In contrast, under the symbolic rule-based instantiation, grouping is derived directly from relational predicates (e.g., proximity, similarity, symmetry, depth consistency) without learned affinity estimation. Implementation details for both variants are provided in App. B.

Regardless of implementation, the output of the grouping module is a set of groups $\mathcal{G}$ together with explicit access to their member objects. Each group thus serves as a first-class symbolic entity over which group-level predicates can be defined. For example, `unique_color(g)` evaluates whether all objects in group $g$ share the same color; `diverse_shape(g)` evaluates whether objects in $g$ exhibit multiple shape types; and cardinality constraints such as `grp_size(g, k)` evaluate whether the size of $g$ equals $k$. (See App. F Table. 8 for all predicates.) These group-level symbolic descriptors form the structured interface to downstream reasoning modules.

### 3.2 SYMBOLIC ABSTRACTION

GRMs use program induction to generate interpretable, Gestalt-based rules from visual input. This is achieved by translating perceptual features into a first-order logic representation suitable for rule learning.

First, GRMs convert observations into atomic formulas (facts) using domain-specific predicates, yielding two distinct fact sets. The first, $\mathcal{F}^{\text{obj}}$, encodes object-level properties such as shape, color, and position (e.g., `shape(obj1, circle)`, `x(obj1, 0.01)`). The second, $\mathcal{F}^{\text{grp}}$, captures group-level structures like Gestalt principles and membership relations (e.g., `principle(g1, proximity)`).

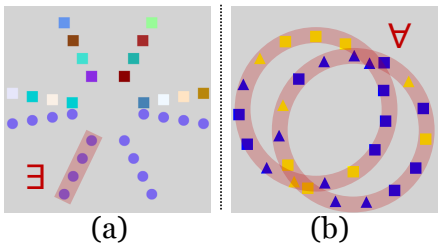



Figure 3: **GRM Scoring with quantifiers.** GRMs use existential (∃) and universal (∀) quantifiers to define rules over groups of objects as shown in the red shade.

From this symbolic representation, GRMs construct an initial candidate clause set, $\mathcal{R}_{\text{init}}$. This set contains simple clauses, each composed of a head and body atom, that serve as a starting point for a subsequent search step where they are expanded into more complex clauses. By bridging perceptual processing with logical reasoning, GRMs enable the discovery of compositional clauses that explain complex visual patterns.

### 3.3 RULE LEARNING WITH STRUCTURED PERCEPTION

GRMs leverage the initial candidate clause set $\mathcal{R}_{\text{init}}$ to learn *target rules*[2] that explain visual scenes according to Gestalt principles. For every candidate clause in $\mathcal{R}_{\text{init}}$, GRM create two variants: an existence clause and a universal clause. They are defined as below:

**Existential Clauses** (∃) are satisfied if **at least one** group in a scene meets a specific condition. For example, the rule for Fig. 3(a) requires that "there exists a group containing objects of the same color."

**Universal Clauses** (∀) are satisfied only if **every** group in a scene meets a specific condition. For instance, the rule for Fig. 3(b) requires that "all groups must contain both a yellow and a blue object."

If no group-based patterns are found, GRMs can treat each object as an individual group. This fallback allows them to handle non-Gestalt patterns, such as detecting the presence of a single red triangle in the image. Learning then proceeds via a top-$k$ beam search over $\mathcal{R}_{\text{init}}$ for at most $T$ expansion steps: in each step, every clause is assigned a soft confidence score determined by its head (see App. C for the scoring equations), clauses are ranked by confidence, and only the top-$k$ clauses are expanded to uncover the scene's underlying group structures. The search terminates either after $T$

[2]We distinguish *rules*, the final explanatory solution, from *clauses* (specifically, definite clauses in first-order logic), the intermediate representations manipulated during search (See App. L).

steps or early when a high-confidence rule has been found, with an overall computational complexity of $O(|C|^T \cdot t_C)$, where $|C|$ is the number of clauses considered and $t_C$ is the cost of evaluating a clause.

### 3.4 INFERENCE

With the learned rules, GRMs finally perform inference on new inputs. At test time, the model performs interpretable inference by applying its learned symbolic rules to a given visual scene. This process unfolds in three stages. First, in the symbolic grounding stage, a perceptual backbone analyzes the input image to detect objects and groups, converting them into a logical fact base ($\mathcal{F}_{\text{test}}$). Next, during rule evaluation, each rule from the final learned set ($\mathcal{R}_{\text{final}}$) is matched against this fact base to compute a soft satisfaction score ($s_j \in [0, 1]$), quantifying its relevance under potential perceptual uncertainty. Finally, the *high-confidence prioritization strategy* determines the output: if any high-confidence rules are activated, the prediction is derived exclusively from their outputs. In all other cases, the system falls back to a robust aggregation method, computing a confidence-weighted average over all partially satisfied rules. For the mathematical descriptions and more details, please check App. D. This two-tiered decision process ensures that transparent, highly reliable rules govern predictions when applicable, while maintaining robust performance in ambiguous scenarios.

## 4 EXPERIMENTS

In this section, we evaluate GRMs on structured visual reasoning tasks that require both perceptual organization and symbolic abstraction. Our experiments are designed to assess GRM's performance, interpretability, and scalability by answering 3 key questions: **(Q1)** How does GRM perform on visual reasoning benchmarks compared to leading neural and neuro-symbolic baselines? **(Q2)** Does the perceptual grouping lead to interpretable and well-structured symbolic rules? **(Q3)** How robust is GRM to architectural ablations and increased task complexity?

Through comprehensive quantitative evaluations, qualitative analysis, and ablation studies, we demonstrate that GRM provides an accurate, interpretable, and scalable approach to neuro-symbolic reasoning, effectively guided by perceptual grouping principles.

### 4.1 EVALUATION PROTOCOL

**Dataset and Task.** We evaluate GRM on the ELVIS benchmark [Sha et al., 2025], a large-scale collection of visual reasoning tasks grounded in Gestalt principles. While existing datasets like CLEVR [Johnson et al., 2017] or RAVEN [Zhang et al., 2019] test object-centric and relational reasoning respectively, they do not require *group-centric* reasoning. ELVIS is specifically designed to fill this gap, with tasks that require models to first aggregate objects into meaningful higher-order entities based on Gestalt principles (e.g., proximity, similarity) before applying logic. In ELVIS, group annotations are available, allowing the grouping operator to be trained in a supervised manner via affinity estimation before downstream reasoning. To our knowledge, ELVIS is the only benchmark that systematically integrates these grouping principles into a neuro-symbolic pipeline, making it uniquely suited for evaluating GRM.

The extensive benchmark features over 100 distinct tasks for each Gestalt principle. Each task is a few-shot learning problem consisting of positive and negative example images. Positive examples adhere to a latent symbolic rule (e.g., "at least one group of similar objects is all red"), while negative examples subtly violate it. From a training set of 3 positive and 3 negative 224×224 RGB images, the objective is to infer the underlying rule and predict binary labels for a held-out test set of the same size, without direct rule supervision. The few-shot setting applies to rule induction; the perception and grouping modules are pretrained and fixed.

**Metrics and Baselines.** GRM is evaluated in a strictly task-wise manner: each task is treated as an independent learning problem with its own train and test splits. For every task, logic rules are learned exclusively from that task's training split and evaluated on its corresponding test split.

We compare GRM against a range of baselines to situate its performance and highlight the benefits of its design. To simulate a few-shot learning scenario, all models are provided with just 3 positive and 3 negative examples for training or in-context learning, and evaluated on a held-out test set of the same size. Our baseline suite begins with *Human Performance*, measured via a web interface[3] to provide a robust reference point. We include a standard deep learning approach, the *Vision Transformer (ViT-16-224)*, trained end-to-end on raw pixels to capture low-level patterns without structured reasoning. To assess the capabilities of state-of-the-art generalist models, we evaluate three *Large Multimodal Models*: LLaVA-7B [Li et al., 2024] (zero-shot), the 78B-parameter variant of InternVL3 [Chen et al., 2024], and GPT-5 [OpenAI, 2025], which are powerful on natural images but are challenged by ELVIS's abstract patterns. Finally, we evaluate *NEUMANN* [Shindo et al., 2024b], which learns object-level rules but lacks perceptual grouping, thereby serving as an ablation to isolate the contribution of GRM's core mechanism. Scallop [Huang et al., 2021], a differentiable Datalog framework, as a second NeSy baseline alongside NEUMANN. Scallop is a widely adopted NeSy reasoner at the same level of abstraction as NEUMANN, and one of the state-of-the-art frameworks on the visual reasoning with logic programs.

[3] https://ml-research.github.io/grb_human_test/

Table 1: **Quantitative Performance on Gestalt Reasoning Tasks.** We compare GRM against baselines and human performance on the ELVIS benchmark. **Bold** indicates the best-performing model (excluding human performance). Cell colors are normalized to visualize relative scores, from high (green) to low (pink). GRM achieves competitive results while providing interpretable group-based reasoning. Model shorthands are as follows: ViT-16-224 is ViT-B/16; Llava-Qwen-7B is LLaVA-OneVision-Qwen2-7B-SI.

| Met. | Model | Proximity | Similarity | Closure | Symmetry | Continuity |
|---|---|---|---|---|---|---|
| Acc. | ViT-16-224 | 0.56 ±0.19 | 0.54 ±0.15 | 0.59 ±0.20 | 0.53 ±0.18 | 0.53 ±0.19 |
| | Llava-Qwen-7B | 0.50 ±0.13 | 0.50 ±0.10 | 0.53 ±0.12 | 0.57 ±0.14 | 0.55 ±0.15 |
| | InternVL3-78B | 0.53 ±0.18 | 0.62 ±0.23 | 0.70 ±0.20 | 0.59 ±0.17 | 0.72 ±0.21 |
| | GPT-5 | **0.72**±0.23 | 0.71 ±0.22 | 0.66 ±0.23 | 0.52 ±0.14 | **0.81**±0.19 |
| | NEUMANN | 0.58 ±0.15 | 0.52 ±0.08 | 0.71 ±0.18 | 0.53 ±0.09 | 0.50 ±0.03 |
| | Scallop | 0.56 ±0.12 | 0.62 ±0.16 | 0.68 ±0.16 | 0.54 ±0.15 | 0.58 ±0.15 |
| | GRM(Ours) | 0.71 ±0.17 | **0.72**±0.21 | **0.78**±0.16 | **0.64**±0.17 | 0.78 ±0.17 |
| | Human | 0.97 ± 0.03 | 0.87 ± 0.13 | 0.92 ± 0.08 | 0.85 ± 0.15 | 0.98 ± 0.02 |
| F1 | ViT-16-224 | 0.50 ±0.29 | 0.45 ±0.30 | 0.56 ±0.28 | 0.48 ±0.29 | 0.55 ±0.25 |
| | Llava-Qwen-7B | 0.22 ±0.30 | 0.27 ±0.32 | 0.53 ±0.28 | 0.38 ±0.34 | 0.24 ±0.33 |
| | InternVL3-78B | 0.32 ±0.33 | 0.45 ±0.39 | 0.59 ±0.34 | 0.36 ±0.35 | 0.52 ±0.41 |
| | GPT-5 | 0.65 ±0.33 | 0.61 ±0.34 | 0.56 ±0.35 | 0.22 ±0.30 | 0.71 ±0.29 |
| | NEUMANN | 0.27 ±0.37 | 0.13 ±0.24 | 0.59 ±0.36 | 0.30 ±0.35 | 0.33 ±0.33 |
| | Scallop | 0.62 ±0.17 | **0.69**±0.14 | 0.74 ±0.13 | **0.59**±0.16 | 0.63 ±0.20 |
| | GRM(Ours) | **0.65**±0.29 | 0.63±0.33 | **0.78**±0.22 | 0.48±0.35 | **0.78**±0.22 |
| | Human | 0.96 ± 0.04 | 0.81 ± 0.19 | 0.90 ± 0.10 | 0.81 ± 0.19 | 0.97 ± 0.03 |

**Pretraining.** GRM's perception backbone uses pretrained object and group detectors that are fixed during reasoning. An object detector identifies shapes, while separate group detectors, one for each Gestalt principle, cluster objects into groups based on learned affinities (MLP based). These detector outputs, along with attributes like color and position derived directly from the object masks, are converted into a symbolic fact base for the rule learner. A complete list of these predicates is provided in App. F. The object and group detector architecture is provided in App. A and B.

**Hardware Requirements** We ran InterVL3-78B on 3 NVIDIA A100-SXM4-80GB, ran GPT-5 via API and ran rest of the models on a single NVIDIA A100-SXM4-80GB. The GRM, NEUMANN and ViT-16-224 models are runnable on a MacBook Pro with M2 Chip.

### 4.2 QUANTITATIVE AND QUALITATIVE EVALUATION

To answer **Q1**, we evaluate GRM on the ELVIS benchmark, spanning five Gestalt principles: *proximity*, *similarity*, *closure*, *symmetry*, and *continuity*. Each task shares a latent structural rule but varies in features like shape, color, and size, requiring both object recognition and group reasoning.

Tab. 1 compares GRM against five baselines (ViT, LLaVA, InternVL3, GPT-5, NEUMANN). GRM achieves the best accuracy among non-human models on similarity, closure, and symmetry, while remaining competitive on proximity and continuity. The main advantage of GRM is that it combines strong task performance with explicit, auditable first-order rules over perceptual groups. Fig. 4 (middle) further analyzes performance by conditioning on object- and group-level properties. GRM maintains balanced accuracy ($\sim$73%) across shape, color, size, group number, and group size, whereas GPT-5 shows imbalances (e.g., strong on shape but weaker on size and color). This highlights the benefit of rule-based evaluation in avoiding confusion from perceptually similar attributes, generalizing to various abstract concepts.

We further evaluate the quality of symbolic fact extraction on 500 test tasks. Object-level properties are extracted with high reliability, including *size* (99%), *position* (95%), and *color* (90%). In contrast, group-level attributes are more challenging due to abstraction and perceptual ambiguity: *group label* reaches 71% accuracy, *object count* 92%, *group number* 76%, and *per-group count* only 44%. These results indicate that while low-level symbolic properties are robustly recovered, higher-order group-related facts remain a bottleneck for reasoning. Full results are provided in App. G.

To answer **Q2**, we examine whether perceptual grouping improves the structural quality and interpretability of the induced symbolic rules. In contrast to black-box visual baselines, GRM represents each task as a set of explicit first-order clauses grounded in detected objects and perceptual groups. Interpretability here refers to two properties: (i) *human readability* of the induced predicates (e.g., shape-, color-, or group-level relations), and (ii) *structural coherence*, meaning that rules operate over semantically meaningful units (groups) rather than isolated objects.

Listing 1 shows representative clauses discovered by GRM on ELVIS. The rules are expressed over group-level predicates such as `in_group/2`, combined with object attributes. For example, GRM induces rules stating that the target group contains triangles or that a group must include

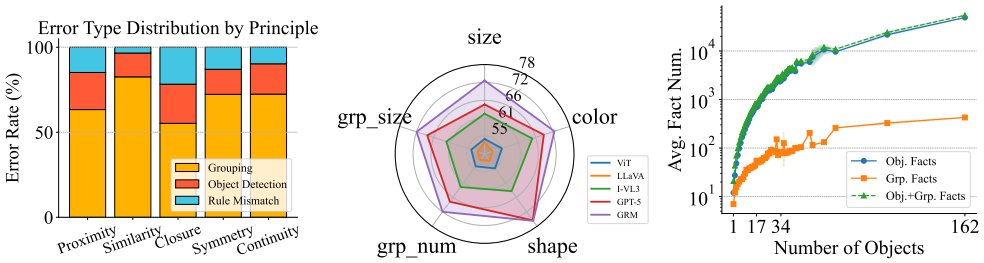



Figure 4: **Combined Results. Left**: Error breakdown by principle, showing grouping errors dominate across all Gestalt principles. **Middle**: Average accuracy (%) over all principles for each property and model. Object-level properties: size, color, shape; Group-level properties: group number and group size. **Right**: Symbolic scalability across object counts, where group-level reasoning adds moderate overhead when combined with object facts but remains lightweight when used alone.

Listing 1: **GRM learns interpretable rules over groups.** Example rules discovered on ELVIS.

```
% Proximity principle
group_target(G,X):-has_shape(O,triangle),
    in_group(O,G).
   [confidence=1.000, scope=universal]
group_target(G,X):-has_color(O,red),
    has_shape(O,square),in_group(O,G).
   [confidence=1.000, scope=existential]
```

a red square. These clauses are compact, human-readable, and directly aligned with Gestalt principles (e.g., proximity-based grouping followed by attribute-based reasoning). Additional task-solving examples are provided in App. H.

Fig. 4 (left) shows the distribution of error sources. Grouping errors dominate, indicating that the object-to-group abstraction remains the main bottleneck. A common failure occurs when distinct structures are mistakenly merged, preventing correct group-level fact construction (see closure examples in App. G). Object detection errors form the second-largest category, typically caused by subtle color variations or incomplete contours. Rule mismatches account for the remaining cases, reflecting limitations in predicate alignment despite correct perception.

The grouping comparison study (App. B) shows that the MLP-based grouper is sufficient for Gestalt principles primarily driven by positional or symbolic attributes, such as proximity, similarity, and continuity, achieving performance comparable to the Transformer-based variant. However, both models exhibit notable performance drops on closure and symmetry tasks. This indicates that the main bottleneck lies in the structural adequacy of object representations rather than model capacity. Currently, grouping relies mainly on symbolic object features (e.g., shape category, color, position, size). While sufficient for simple relational principles, these features do not capture contour geometry, orientation, partial occlusion, or higher-order spatial configurations that are essential for principles such as closure and symmetry. Future improvements therefore require incorporating richer neural features, such as object contour embeddings, geometric descriptors, or mathematically structured representations. Then the grouping module has access to the information necessary to infer group identities.

To further examine whether imperfect grouping directly limits downstream reasoning, we condition GRM's task accuracy on grouping quality in Tab. 2. Although correct grouping generally improves performance, GRM does not collapse when the predicted groups are incomplete or partially incorrect. For example, even when less than 20% of ground-truth groups are recovered, GRM still achieves accuracies between 0.57 and 0.79 across principles. Moreover, the relationship between grouping quality and reasoning accuracy is not strictly monotonic: in some cases, very low grouping quality can be less harmful than partially correct grouping, because partially correct groups may support high-confidence but semantically misleading rules. This suggests that predicted groups act as soft perceptual abstractions rather than hard prerequisites, and that GRM's rule scoring can suppress many inconsistent group-level signals across positive and negative examples.

### 4.3 ROBUSTNESS AND COMPONENT ANALYSIS

To answer **Q3**, we analyze GRM's robustness through ablation studies and scalability tests.

**Object-Only Reasoning as a Degenerate Case.** Although our evaluation focuses on tasks where perceptual grouping is relevant, GRM does not require grouping to be active in all cases. When no meaningful group structure is needed, the grouping module can be disabled, and GRM reduces to an object-level neuro-symbolic rule learner. In this setting, the fact base contains only object-level predicates, such as `has_shape`, `has_color`, `x`, `y`, and pairwise object relations, while group-level predicates, such as `in_group`, `group_size`, and `principle`, are removed. This setting corresponds to the *w/o Group* variant in Tab. 3. Thus, Gestalt grouping should be understood as an optional inductive bias for group-centric reasoning rather than a mandatory assumption for all visual reasoning tasks.

Table 2: **Reasoning accuracy conditioned on grouping quality.** Acc. (Good) denotes reasoning accuracy on samples with correct grouping results, while Acc. (Bad) denotes reasoning accuracy on samples with incorrect or partially correct grouping results. Low ($< 0.2$) and High ($0.8$–$1.0$) denote reasoning accuracy when less than 20% or more than 80% of ground-truth groups are recovered, respectively.

| **Principle** | **Acc. (Good)** | **Acc. (Bad)** | **Low** ($< 0.2$) | **High** ($0.8$–$1.0$) |
|---|---|---|---|---|
| Proximity | 0.75 | 0.59 | 0.69 | 0.74 |
| Closure | 0.82 | 0.81 | 0.79 | 0.82 |
| Symmetry | 0.67 | 0.55 | 0.57 | 0.67 |
| Continuity | 0.82 | 0.75 | 0.71 | 0.82 |
| Similarity | 0.74 | 0.73 | 0.77 | 0.73 |

**Ablation of Perceptual Grouping.** We perform an ablation study to quantify the contribution of the grouping mechanism and to instantiate the object-only setting described above. As shown in Tab. 3, we compare the full GRM model (*w/ Group*), which uses both object- and group-level facts, against a variant that reasons only over object-level features (*w/o Group*). Adding group-level information improves accuracy across all five principles, with the largest relative gains on *continuity* (+61%) and *similarity* (+36%). This confirms that group-level abstraction provides a useful inductive bias for structured reasoning, while the object-only variant remains a valid fallback when group structure is unnecessary. [4]

**Grouping Benefit Across Rule Learners.** To verify that the benefit of grouping is not specific to GRM's beam-search rule learner, we further evaluate Scallop with object-only facts and with object-plus-group facts in Tab. 4. Adding group-level facts improves Scallop on four of five principles, with particularly large gains on *continuity* ($0.42 \rightarrow 0.58$) and *similarity* ($0.51 \rightarrow 0.62$). This result indicates that perceptual grouping is a useful inductive bias beyond GRM's specific search procedure: even when the downstream differentiable logic engine is changed, exposing group-level abstractions improves structured visual reasoning. GRM still achieves higher accuracy than both Scallop variants across all principles, suggesting that its rule search and high-confidence prioritization strategy make more effective use of group-based symbolic abstractions.

**Scalability with Scene Complexity.** We assessed scalability by measuring the growth of the symbolic fact base as the number of objects in a scene increases (Fig. 4, right). While purely object-based representation grows near-quadratically, adding group-level facts introduces only a modest representational overhead. This small cost yields a substantial performance benefit (see Tab. 3). Perceptual grouping, therefore, acts as a lightweight yet highly effective enhancement to the symbolic pipeline. It enriches the model's representational capacity and provides a natural mechanism for abstracting away redundant object-level details.

[4] A group-only variant was not evaluated, as the reasoning tasks require object-level predicates (e.g., shape and color), and the groups themselves are derived from detected objects.

Table 3: **Grouping enhances the reasoning performance.** Accuracy (%) comparisons of GRM without grouping vs. with grouping. The green values are the relative improvement of the w/Group relative to the w/o Group

| **Principle** | **w/o Group** | **w/ Group** |
|---|---|---|
| Proximity | $58.0_{\pm 15.0}$ (0%) | **70.0**$_{\pm 17.0}$ (+19%) |
| Similarity | $52.0_{\pm 8.0}$ (0%) | **70.0**$_{\pm 22.0}$ (+36%) |
| Symmetry | $53.0_{\pm 9.0}$ (0%) | **62.0**$_{\pm 18.0}$ (+17%) |
| Closure | $71.0_{\pm 18.0}$ (0%) | **79.0**$_{\pm 16.0}$ (+11%) |
| Continuity | $50.0_{\pm 3.0}$ (0%) | **81.0**$_{\pm 19.0}$ (+61%) |

Table 4: Scallop baseline with and without group-level facts. Scallop (o) uses only object-level facts, while Scallop (og) additionally receives the same group-level facts used by GRM. Adding group facts improves Scallop on four of five Gestalt principles, indicating that group-level abstraction is beneficial beyond GRM's specific rule learner.

| **Model** | **Proximity** | **Similarity** | **Closure** | **Symmetry** | **Continuity** |
|---|---|---|---|---|---|
| Scallop (o) | $0.52 \pm 0.13$ | $0.51 \pm 0.11$ | $0.70 \pm 0.17$ | $0.48 \pm 0.17$ | $0.42 \pm 0.15$ |
| Scallop (og) | $0.56 \pm 0.12$ | $0.62 \pm 0.16$ | $0.68 \pm 0.16$ | $0.54 \pm 0.15$ | $0.58 \pm 0.15$ |
| GRM | $0.71 \pm 0.17$ | $0.72 \pm 0.21$ | $0.78 \pm 0.16$ | $0.64 \pm 0.17$ | $0.78 \pm 0.17$ |

**Computational Efficiency.** GRM is substantially more efficient than GPT-5 and remains competitive with the highly stable InternVL3-78B, solving most tasks in under 10 seconds, while GPT-5 often exceeds 100 seconds (Fig. 14). A detailed analysis of how perceptual grouping quality affects induction time is provided in App. I.

**Robustness to 3D Geometry and Occlusion.** To evaluate whether GRM's grouping mechanism generalizes beyond planar 2D layouts, we construct a CLEVR-style 3D extension of ELVIS tasks by converting 10 representative tasks (two per Gestalt principle across five principles) into synthetic scenes rendered with perspective projection and object-level occlusion. The original group-centric reasoning structure is preserved, while depth compression and partial visibility introduce realistic geometric challenges absent from the 2D setting. GRM is applied without architectural modification in this setting, allowing us to isolate the robustness of the learned grouping operator under 3D transformations. Results are reported in Tab. 5. App. K provides the experiment details and the image examples.

### 4.4 REAL-IMAGE (COCO 2017) EXPERIMENT

Unlike the ELVIS 2D scenes, real-world images are 2D projections of a 3D environment. In natural images, objects at different depths may appear spatially close in the image plane, making 2D bounding boxes insufficient for reliable

Table 5: **GRM is CLEVR.** Performance of GRM on 10 CLEVR-style tasks converted from ELVIS. Each Gestalt principle contains two tasks (five principles in total).

| | Proximity | Similarity | Closure | Symmetry | Continuity |
|---|---|---|---|---|---|
| GRM-ACC | 0.98 | 0.98 | 1.00 | 0.70 | 1.00 |
| GRM-F1 | 0.98 | 0.98 | 1.00 | 0.67 | 1.00 |

grouping (See App. J). To recover missing spatial cues, we incorporate monocular depth as complementary positional evidence. and use COCO 2017 images. Since COCO lacks depth annotations, we estimate monocular depth maps using Depth Anything v3 [Yang et al., 2024]. Although imperfect, the estimated depth provides coarse relative spatial ordering that improves grouping beyond purely 2D geometry.

Due to the absence of objective group-level ground truth and the inherent ambiguity of real-world grouping, we implement the grouping module using explicit symbolic rules rather than a learned neural model. Candidate groups are constructed using Gestalt-inspired cues, including spatial proximity (depth-aware), category similarity, size similarity, and depth similarity. Importantly, GRM does not assume a fixed grouping implementation; grouping serves as an intermediate abstraction module that can be realized either neurally or symbolically, depending on the availability of supervision. The COCO case study, therefore, illustrates an alternative instantiation of this component (See app. J for more discussion). As shown in Fig. 5, different first-order rules select different plausible groups from the same image, reflecting context-dependent perceptual organization. This experiment demonstrates that GRM supports rule-conditioned perceptual abstraction in realistic inputs.

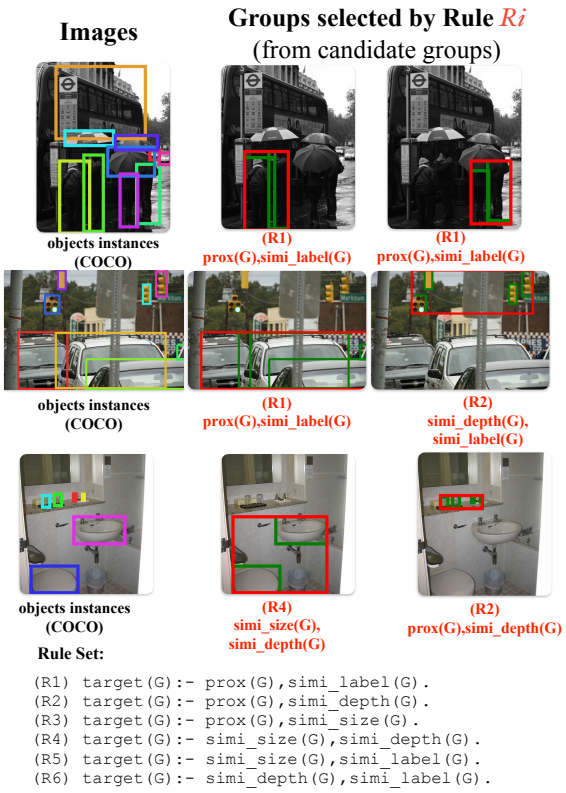



Figure 5: **Rule-based symbolic grouping in real-world images.** Object instances from COCO annotations are used to construct candidate groups based on perceptual cues (proximity, category, depth, and size). Symbolic reasoning selects a target group according to a first-order rule $R_i$. Different rules may yield alternative but plausible perceptual organizations. Predicate definitions are provided in App. E.

## 5 CONCLUSION

We introduced GRM, a neuro-symbolic framework that integrates perceptual grouping with symbolic rule induction to solve complex visual reasoning tasks grounded in Gestalt principles. The formalism of rule induction supports logically grounded and interpretable reasoning, while the scalable grouping mechanism maintains a compact symbolic representation even in complex scenes. GRM achieves competitive or superior performance compared with state-of-the-art neural, multimodal, and neuro-symbolic baselines on several Gestalt reasoning principles, while producing explicit group-based rules. These results highlight the promise of structured neuro-symbolic architectures and position GRM as a step toward cognitive systems with more robust, interpretable, and human-like structured perception.

As future work, several directions are worth further investigation. First, grouping remains the main bottleneck, as current object representations do not sufficiently encode higher-order spatial configurations such as partial occlusion, contour geometry, and orientation. Incorporating richer neural-geometric representations, such as contour embeddings or geometry-aware feature encoders, may improve the reliability of group-level abstraction. Second, real-world grouping requires information beyond 2D image coordinates; extending GRM with depth-aware or 3D object-centric representations would better align the model with human perceptual grouping in natural scenes. Finally, extending Gestalt-based neuro-symbolic reasoning to streaming data such as videos remains an important direction for dynamic structured perception, where perceptual groups may persist, split, merge, and move over time.

## ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC-3057. The TU Eindhoven authors received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute.

## References


Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling "Visual" from "Reasoning". In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

Giacomo Camposampiero, Loïc Houmard, Benjamin Estermann, Joël Mathys, and Roger Wattenhofer. Abstract visual reasoning enabled by language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Forty-second International Conference on Machine Learning (ICML)*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Willis D. Ellis. *A Source Book of Gestalt Psychology*. Routledge, 1999.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research (JAIR)*, 2018.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

Haim Gaifman. On local and non-local properties. In J. Stern, editor, *Proceedings of the Herbrand Symposium*, volume 107 of *Studies in Logic and the Foundations of Mathematics*, pages 105–135. Elsevier, 1982.

Shihui Han, Yulong Ding, and Yan Song. Neural mechanisms of perceptual grouping in humans as revealed by high density event related potentials. *Neuroscience letters*, 319(1):29–32, 2002.

Joy Hsu, Jiayuan Mao, Joshua B. Tenenbaum, and Jiajun Wu. What's left? concept grounding with logic-enhanced foundation models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Tianyu Hua and Maithilee Kunda. 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)*, 2020.

Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, and Xujie Si. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25134–25145. Curran Associates, Inc., 2021.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 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)*, 2017.

Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C. Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 4(3), 2021.

Kurt Koffka. *Principles of Gestalt Psychology*. Harcourt, Brace & World, 1935.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2034–2044, October 2025.

András Lörincz, Áron Fóthi, Bryar O. Rahman, and Viktor Varga. Deep gestalt reasoning model: Interpreting electrophysiological signals related to cognition. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshop)*, 2017.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Mikolaj Malkinski and Jacek Mandziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91, 2023.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in deepproblog. *Artif. Intell.*, 298, 2021.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence (AIJ)*, 328, 2024.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. Learning reasoning strategies in end-to-end differentiable proving. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Heimo Müller and Andreas Holzinger. Kandinsky patterns. *Artificial Intelligence (AIJ)*, 2021.

OpenAI. Introducing GPT-5. `https://openai.com/index/introducing-gpt-5/`, August 2025. Accessed: 2025-08-29.

Stephen E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA, 1999.

John C Raven and John Hugh Court. *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press, Oxford, 1998.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Jingyuan Sha, Hikaru Shindo, Kristian Kersting, and Devendra Singh Dhami. Neuro-symbolic predicate invention: Learning relational concepts from visual scenes. *Neurosymbolic Artificial Intelligence*, 2024.

Jingyuan Sha, Hikaru Shindo, Kristian Kersting, and Devendra Singh Dhami. Gestalt vision: A dataset for evaluating gestalt principles in visual perception. In *Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning (NeSy)*, 2025.

Hikaru Shindo, Viktor Pfanschilling, Devendra Singh Dhami, and Kristian Kersting. $\alpha$ilp: thinking visual scenes as differentiable logic programs. *Machine Learning (MLJ)*, 2023.

Hikaru Shindo, Manuel Brack, Gopika Sudhakaran, Devendra Singh Dhami, Patrick Schramowski, and Kristian Kersting. Deisam: Segment anything with deictic prompting. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024a.

Hikaru Shindo, Viktor Pfanschilling, Devendra Singh Dhami, and Kristian Kersting. Learning differentiable logic programs for abstract visual reasoning. *Machine Learning (MLJ)*, 2024b.

Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Gopika Sudhakaran, Hikaru Shindo, Patrick Schramwoski, Simone Schaub-Meyer, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *Proceedings of the 20th IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kristinn R Thórisson. Simulated perceptual grouping: An application to human-computer interaction. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 876–881. Routledge, 2019.

Max Wertheimer. Laws of organization in perceptual forms. In Willis D. Ellis, editor, *A Source Book of Gestalt Psychology*, pages 71–88. Routledge & Kegan Paul, 1938.

Antonia Wüst, Tim Tobiasch, Lukas Helff, Inga Ibs, Wolfgang Stammer, Devendra S Dhami, Constantin A Rothkopf, and Kristian Kersting. Bongard in wonderland: Visual puzzles that still make ai go mad? *International Conference on Machine Learning (ICML)*, 2025.

Yaoda Xu and Marvin M Chun. Visual grouping in human parietal cortex. *Proceedings of the national academy of sciences*, 104(47):18766–18771, 2007.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. CounterCurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In *Findings of the Association for Computational Linguistics (ACL)*, 2024a.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Yuyan Zhang, Derya Soydaner, Fatemeh Behrad, Lisa Koßmann, and Johan Wagemans. Investigating the gestalt principle of closure in deep convolutional neural networks. In *32nd European Symposium on Artificial Neural Networks (ESANN)*, 2024b.

Matthieu Zimmer, Xuening Feng, Claire Glanois, Zhaohui Jiang, Jianyi Zhang, Paul Weng, Dong Li, Jianye Hao, and Wulong Liu. Differentiable logic machines. *Transactions on Machine Learning Research (TMLR)*, 2023.

# Birds of a Feather Reason Together: Gestalt Grouping Meets Neuro-Symbolic Inference (Supplementary Material)

**Jingyuan Sha**[1] **Hikaru Shindo**[1] **Kristian Kersting**[1,2,3] **Devendra Singh Dhami**[4]

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

## A OBJECT DETECTOR DETAILS

The object model is a two-layer MLP: it flattens input patches, maps them to 128 hidden units with ReLU, then outputs class scores. The input patches are extracted by identifying the contours of regions in the image.

## B GROUP DETECTOR ARCHITECTURE

### B.1 SUPERVISED AFFINITY-BASED GROUPING

The grouping model is trained in a fully supervised manner using synthetically generated data derived from ELVIS patterns, which provide ground-truth group assignments under specific Gestalt principles. The objective is to learn a pairwise affinity function that predicts whether two objects belong to the same perceptual group.

Each object $o_i$ is first transformed into a fixed-length embedding $\mathbf{o}_i = f(o_i)$ using a shared encoder $f$. The encoder consists of: (i) a *point encoder*, implemented as a two-layer MLP with ReLU activations that maps each input point to a hidden representation; and (ii) a *patch encoder*, also a two-layer MLP with ReLU, which aggregates encoded points into a single object embedding.

To model group membership, the network evaluates pairwise affinities while conditioning on the full scene context. For each Gestalt principle $p$, the model estimates:

$$s_p(o_i, o_j, I) = \sigma\big(h_p(\mathbf{o}_i, \mathbf{o}_j, \mathbf{o}^*_{ij})\big)\,,$$

where $h_p$ is a principle-specific MLP, $\sigma$ denotes the sigmoid activation, and $\mathbf{o}^*_{ij}$ represents the global context embedding. The context vector is computed as the mean embedding of all other objects in scene $I$:

$$\mathbf{o}^*_{ij} = \frac{1}{|\mathcal{O}| - 2} \sum_{k \neq i,j} \mathbf{o}_k.$$

This mean pooling operation is permutation-invariant and ensures that the context representation remains stable across scenes with varying object counts [Gaifman, 1982]. It provides a simple scene-level summary that informs whether a candidate pair is typical or exceptional relative to its surroundings (see Fig. 6).

The resulting affinity score $s_p$ denotes the estimated probability that two objects belong to the same group under principle $p$. Groups are subsequently extracted by thresholding the affinity matrix and computing connected components. Group-level symbolic attributes (e.g., color diversity, shape uniformity) are then aggregated to produce structured representations for downstream reasoning.

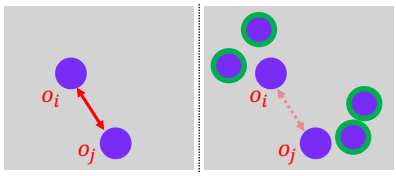



Figure 6: **Structured perception by GRMs.** Without contextual objects (left), group identification becomes ambiguous. GRMs leverage global context (right) by aggregating embeddings from all objects (green contour) into $\mathbf{o}^*_{ij}$.

Table 6: **Comparison of MLP and Transformer-based groupers.** All models are trained on the ELVIS grouping task; The evaluation metric is F1-score. GPT-Zero-Shot denotes a large VLM (GPT-5) evaluated in a zero-shot setting.

| Metric | MLP+Context | Transformer+Context | GPT-Zero-Shot |
|---|---|---|---|
| Proximity | 0.98 | 0.99 | 0.82 |
| Similarity | 0.99 | 0.99 | 0.51 |
| Closure | 0.74 | 0.67 | 0.72 |
| Symmetry | 0.66 | 0.76 | 0.27 |
| Continuity | 0.99 | 0.97 | 0.82 |
| Time/Task | 1.94s | 3.58s | 57s |
| Params | 0.5M | 3.2M | 635,000M |

## B.2 TRANSFORMER BASED GROUPING

A natural alternative to the mean-pooled MLP grouper is to use a Transformer encoder over the object set, which can in principle model variable-length contextual interactions without relying on explicit pooling.

To ensure a fair comparison, we align the interface of the Transformer with the MLP-based design. Specifically, given a candidate pair of objects $(o_i, o_j)$ and the remaining objects in the scene as context, the model predicts whether the pair belongs to the same group under a given Gestalt principle.

To validate our design choice, we perform a controlled comparison between MLP- and Transformer-based groupers on the ELVIS grouping task.

### B.2.1 Comparison of MLP and Transformer-based groupers

All learned groupers are trained to solve the same binary group-detection task: given a candidate pair of objects, predict whether they belong to the same group under a specified Gestalt principle. We consider two learned variants and one large VLM baseline: **MLP+Context**: An MLP head $h_p$ that takes $(\mathbf{o}_i, \mathbf{o}_j, \mathbf{o}^*_{ij})$, where $\mathbf{o}^*_{ij}$ is the mean-pooled embedding of all other objects in the scene. **Transformer+Context**: A lightweight Transformer encoder that takes as input the embeddings of the candidate pair together with all other object embeddings. The output tokens corresponding to the candidate pair are pooled and fed to a classifier. **GPT-Zero-Shot**: A large VLM (GPT-5) prompted in a zero-shot setting to directly predict group membership for each candidate pair, without task-specific training on ELVIS. All learned models are trained using identical data splits and loss functions as the main MLP grouper. We evaluate performance on five Gestalt principles. Runtime is measured as the average wall-clock time per ELVIS task on a single NVIDIA A100-SXM4-80GB GPU (GPT-5 latency measured separately via API).

**Results and Analysis.** Table 6 reports per-principle F1 scores, runtime, and parameter counts. The MLP+Context model achieves performance comparable to Transformer+Context while remaining substantially more efficient in both parameter count and runtime. In particular, both learned models perform strongly on principles primarily grounded in low-order symbolic cues—proximity, similarity, and continuity—achieving F1 scores above 0.97. This suggests that for principles dominated by simple relational features (e.g., Euclidean distance, attribute similarity), expressive self-attention mechanisms are not strictly necessary.

In contrast, closure and symmetry remain challenging for both models. For **closure**, certain task instances contain perceptual cues that are not explicitly represented in the symbolic object features. For example, three "pac-man"-like circular segments may form an implied triangle. Although each object is labeled as a circle in the symbolic representation, critical contour-level information—such as oriented gaps that induce illusory contours—is not encoded in features like color, shape category, or position. Consequently, neither the MLP nor the Transformer can reliably recover the higher-order geometric configuration required for closure-based grouping. For **symmetry**, the difficulty arises from the variability of the symmetry axis. Tasks include horizontal, vertical, two diagonal axes, and radial symmetry. The models must implicitly infer the symmetry axis from object positions alone. Since grouping is learned as a statistical mapping rather than implemented via explicit geometric reasoning, correctly pairing symmetric objects under varying global axes becomes non-trivial. This limitation reflects the absence of an explicit symmetry-detection mechanism or structured geometric prior.

**Implications.** These results suggest that the primary bottleneck is not the mean-pooling operation per se, but rather the expressiveness and structure of the object representation. For principles governed by simple positional or attribute relations (e.g., proximity, continuity), the current symbolic features are sufficient. However, principles such as closure require contour- or configuration-level encodings, while symmetry may demand explicit geometric reasoning or axis inference mechanisms.

Therefore, simply increasing architectural capacity (e.g., replacing MLP with Transformer) does not fundamentally resolve grouping challenges. A more promising direction is to design structured object representations or incorporate geometry-aware modules, potentially inspired by recent set-based or equivariant architectures, rather than relying solely on generic attention mechanisms.

### B.3 SYMBOLIC RULE-BASED GROUPING

Symbolic rule-based grouping is used when explicit group annotations are unavailable. In this setting, the grouping operator does not require learning; instead, groups are constructed deterministically from relational predicates defined over detected objects. Each predicate evaluates whether a pair of objects satisfies a specific perceptual relation, thereby inducing candidate groups without parameter optimization.

In real-world datasets, state-of-the-art object detectors provide object categories and spatial localization (e.g., bounding boxes or segmentation masks). In addition, monocular depth estimation methods such as Depth Anything v3 Yang et al. [2024] enable approximate recovery of relative 3D structure. These signals allow us to instantiate grouping predicates based on semantic similarity, spatial proximity, geometric consistency, and depth consistency. The grouping mechanisms are defined as follows.

**Proximity groups** are obtained by constructing a spatial graph in which two objects are connected if the Euclidean distance between their bounding-box centers (normalized by image size) is below a fixed threshold (default $0.1$). Groups are defined as connected components extracted via depth-first search. Connectivity is transitive: if $o_i$ is connected to $o_j$, and $o_j$ to $o_k$, then all three belong to the same group.

**Depth groups** are formed by partitioning objects according to depth quantization. For each image, depth values are divided into five adaptive bins (very_near, near, mid, far, very_far) determined by the 20th, 40th, 60th, and 80th percentiles of the depth distribution. Objects assigned to the same depth bin constitute a group.

**Category groups** are constructed by partitioning objects according to their semantic class labels. All objects sharing the same category are grouped together. This strategy encodes a weak prior that semantic similarity may correlate with perceptual grouping.

**Size-similarity groups** are defined through geometric consistency. Two objects are connected if both their bounding-box areas and aspect ratios are similar within a tolerance of 25%. Groups are then formed as connected components via queue-based exploration, yielding transitive geometric clusters.

All grouping strategies produce non-hierarchical groups containing at least two members. Different strategies may yield overlapping group assignments for the same objects. These candidate groups serve as inputs to first-order symbolic rules of the form

$$\texttt{target}(G) : -p_1(G), \ldots, p_k(G),$$

where $p_\ell(G)$ denotes group-level constraints (e.g., proximity, depth consistency, semantic similarity). The final selected groups are therefore fully determined by the relational structure of the scene and the specified rule, without any parameter learning.

## C DETAILS ABOUT THE CLAUSE SCORING

During beam search, each candidate clause $r$ is evaluated by a clause scorer that estimates how well $r$ explains the task's training images while avoiding spurious matches on negatives. The score depends on the type of head attached to $r$ (`image_target`, `group_target`, or `group_universal`) and is normalized to lie in $[0, 1]$.

Intuitively, the scorer rewards coverage of positive images (or groups) and penalizes violations on negative images, thereby biasing the search towards clauses that capture stable Gestalt regularities rather than accidental coincidences.

Formally, we define:

$$\texttt{score}(r) = \begin{cases} \dfrac{n_+(r)}{N_+} \cdot \Big(1 - \dfrac{n_-(r)}{N_-}\Big), & \texttt{image\_target}, \\ \dfrac{n_+^{\exists}(r)}{N_+} \cdot \Big(1 - \dfrac{n_-^{\exists}(r)}{N_-}\Big), & \texttt{group\_target}, \\ \dfrac{1}{N_+} \sum_{i=1}^{N_+} \min\Big(\dfrac{m_i}{M_i}, 1\Big) \cdot \Big(1 - \dfrac{n_-^{\forall}(r)}{N_-}\Big), & \texttt{group\_universal}. \end{cases} \tag{1}$$

## D HIGH-CONFIDENCE PRIORITIZATION STRATEGY.

Given a set of learned rules $\mathcal{R}_{\text{final}}$, each rule $r_j$ is associated with a confidence $\alpha_j \in [0, 1]$ and produces a soft match score $s_j \in [0, 1]$ on the test fact base. The final prediction score $\hat{y}_{\text{test}}$ is computed as

$$\hat{y}_{\text{test}} = \begin{cases} \frac{1}{|\mathcal{H}|} \sum_{r_j \in \mathcal{H}} s_j, & \text{if } \mathcal{H} \neq \varnothing, \\ \frac{\sum_j \alpha_j^2 \cdot s_j}{\sum_j \alpha_j^2 + \epsilon}, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{H} = \{r_j \mid \alpha_j \geq \tau\}$ is the set of rules firing with confidence above threshold $\tau$, and $\epsilon$ is a small constant for numerical stability. A higher $\tau$ enforces stricter rule selection. It can cover more positive and fewer negative cases. which provides high precision and interpretability. For example, the threshold $\tau = 0.99$ is used to retain only those rules whose confidence exceeds $99\%$, meaning the rule is highly consistent with the training examples. In the experiments, we choose $\tau = 0.99$, which keeps only the rules that reliably distinguish positive from negative examples.

If no rule fires at all, a fixed fallback prior (e.g., $0.1$) is returned. This strategy prioritizes high-confidence rules when available, while providing a smooth weighted aggregation otherwise.

## E PREDICATES DEFINITION (COCO DATASET)

Table 7: List of predicate functions used in the COCO image group detection.

| Predicate | Type | Description |
|---|---|---|
| prox | Group | Objects in group $G$ are spatially proximal. Computed using pairwise Euclidean distance between bounding box centers, normalized by image size, with a learned or fixed threshold. |
| simi_size | Group | Objects in group $G$ have similar physical size. Measured via bounding box area similarity (ratio or relative difference below threshold). |
| simi_depth | Group | Objects in group $G$ lie at similar depth. Estimated using depth maps (e.g., monocular depth prediction or RGB-D data) and enforcing low variance of depth values within $G$. |
| simi_label | Group | Objects in group $G$ share the same semantic category label (e.g., COCO class ID). |

# F SHARED BACKGROUND KNOWLEDGE FOR MODELS

For GRM, reasoning relies on a set of pre-defined predicate functions that serve as background knowledge (Tab. 8). These predicates are not learned by the model but specified in advance by the human, covering basic object- and group-level properties (e.g., shape, color, size, membership). The advantage of this design is flexibility: different tasks can be supported by simply providing different predicate sets, while the model pipeline itself remains unchanged. In this way, GRM decouples symbolic knowledge specification from the reasoning procedure, enabling interpretable and task-adaptive rule induction without modifying the architecture.

For fairness, all corresponding predicate definitions are also provided to LLM baselines in the form of natural language prompts, so that both GRM and LLMs operate with the same symbolic information. The background knowledge prompt is given as follows:

> You are given images containing multiple objects and groups. Each object and group has attributes: shape, color, size, position, and group membership. Logical patterns in the image may involve single relations (e.g., all objects have the same color) or combinations of multiple relations (e.g., objects with the same shape are grouped together and mirrored along the x-axis). You can reason about: Individual attributes: shape, color, size, position; Group properties: number of members, grouping principle; Relations: same/different shape, color, size; mirrored positions; unique/diverse attributes within groups. Analyze the image by identifying both simple and complex combinations of these relations.

Table 8: List of predicate functions used in the model.

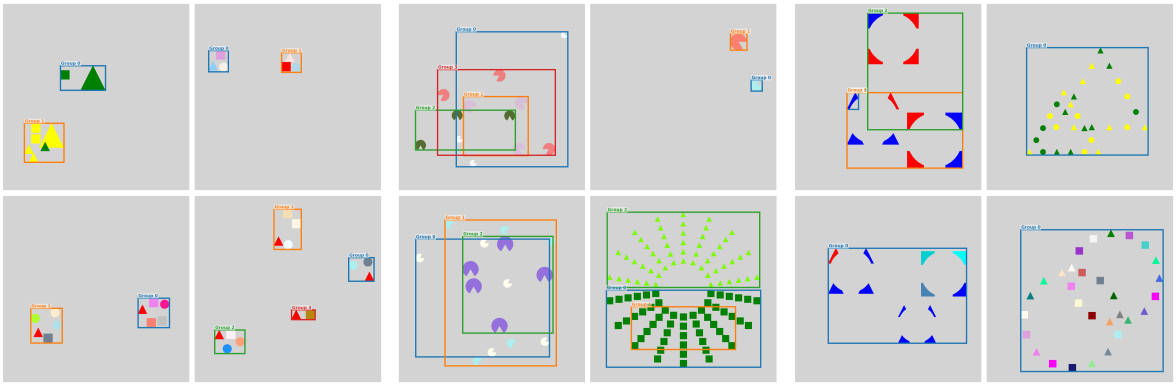
| **Predicate** | **Type** | **Description** |
|---|---|---|
| has_shape | Object | Returns shape index for each object |
| has_color | Object | Returns color index for each object |
| x | Object | Returns x position for each object |
| y | Object | Returns y position for each object |
| w | Object | Returns width for each object |
| h | Object | Returns height for each object |
| in_group | Object | Returns group membership matrix |
| not_has_shape_rectangle | Object | True if object is not a rectangle |
| not_has_shape_circle | Object | True if object is not a circle |
| not_has_shape_triangle | Object | True if object is not a triangle |
| same_shape | Object | Pairwise: True if objects have same shape |
| same_color | Object | Pairwise: True if objects have same color |
| same_size | Object | Pairwise: True if objects have same size |
| mirror_x | Object | Pairwise: True if objects are mirrored along x-axis |
| same_y | Object | Pairwise: True if objects share y-coordinate |
| group_size | Group | Returns number of members in each group |
| principle | Group | Returns grouping principle index |
| no_member_rectangle | Group | True if group has no rectangle members |
| no_member_circle | Group | True if group has no circle members |
| no_member_triangle | Group | True if group has no triangle members |
| diverse_shapes | Group | True if group contains at least two shapes |
| unique_shapes | Group | True if group contains only one shape |
| diverse_colors | Group | True if group contains at least two colors |
| unique_colors | Group | True if group contains only one color |
| diverse_sizes | Group | True if group contains at least two sizes |
| unique_sizes | Group | True if group contains only one size |
| same_group_counts | Group | True if all groups have same member count |
| group_count | Group | Returns number of groups |

## G SYMBOLIC FACT EXTRACTION PERFORMANCE

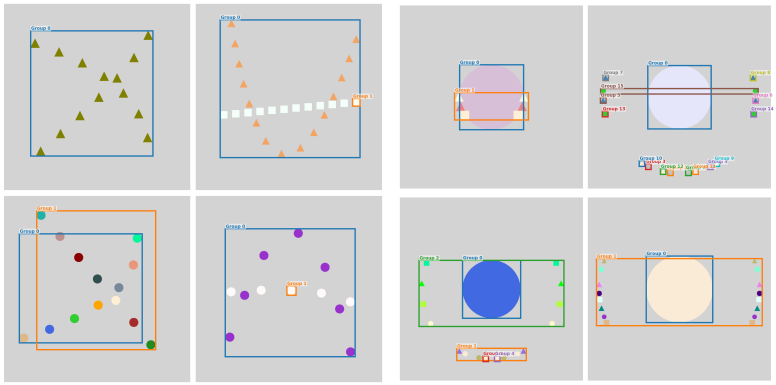
Tab. 9 reports the detailed accuracy of symbolic fact extraction across 500 ELVIS test tasks. While object-level properties such as *shape* (87%), *color* (90%), *size* (99%), and *position* (95%) are recovered with high reliability, group-level properties are more challenging. *Group label* achieves 71% accuracy, *object count* 92%, *group number* 76%, and *per-group count* only 44%. These results highlight a gap between reliable low-level perception and more complex relational grouping, motivating further improvements in symbolic abstraction. Fig. 7 shows the examples of grouping results over different gestalt principles.

Table 9: Mean accuracy (%) and standard deviation of symbolic fact extraction across 500 ELVIS test tasks.

| **Fact Type** | Shape | Color | Size | Position | Group Label | Obj. Count | Group # | Per-Group Count |
|---|---|---|---|---|---|---|---|---|
| **Accuracy** | $0.87 \pm 0.03$ | $0.90 \pm 0.06$ | $0.99 \pm 0.01$ | $0.95 \pm 0.05$ | $0.71 \pm 0.21$ | $0.92 \pm 0.08$ | $0.76 \pm 0.24$ | $0.44 \pm 0.39$ |



(a) Proximity (b) Similarity (c) Closure



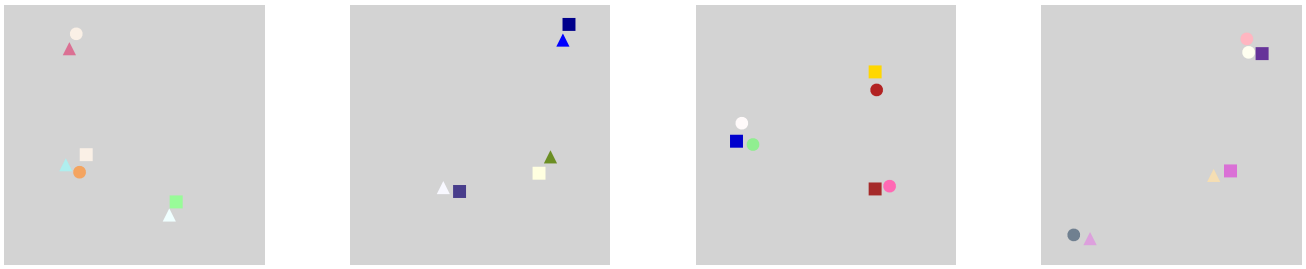
(d) Continuity (e) Symmetry

Figure 7: **Qualitative Results of Group Detection**. Visualization of predicted group structures for five Gestalt principles on randomly selected tasks from the ELVIS benchmark. Each image shows the original scene with predicted group bounding boxes overlaid, demonstrating how GRM organizes visual objects according to different perceptual grouping principles.

# H TASK EXAMPLES AND THE MODEL ANSWERS

## H.1 EXAMPLE TASK 1

This is a task called *Triangle in Groups* following *proximity* principle. The ground-truth rule is that each proximity group contains at least one triangle. Fig. 8 presents the positive and negative examples.



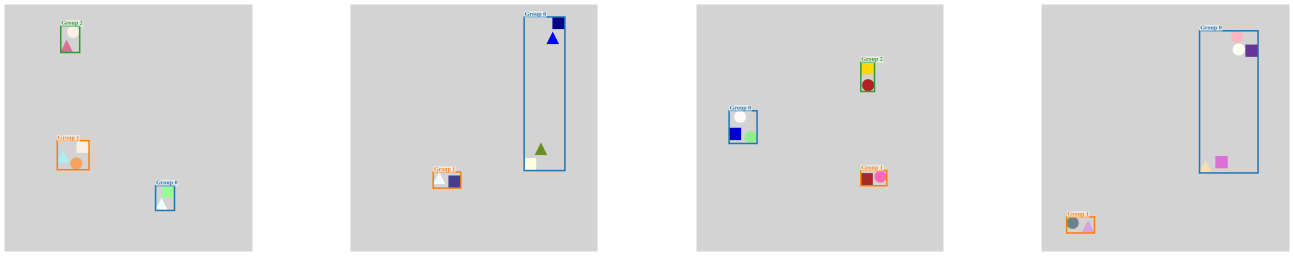
(a) Positive (b) Positive (c) Negative (d) Negative

Figure 8: Triangle in Groups: Each proximity group has at least one triangle.

**GRM.** The induced rule from GRM is shown in Listing 2. The target rule is successfully identified by GRM, but its confidence is only 0.667. This relatively low value reflects imperfect object or group detection in the images (See Fig. 9 (d)), which reduces the rule's overall confidence score.

Listing 2: Rules induced by GRM on Example Task 1

```
% Group-level rules
group_target(G,X) :- has_shape(O,0), in_group(O,G).
   [confidence=0.667, scope=universal]
```



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

Figure 9: GRM grouping results of task 1 examples

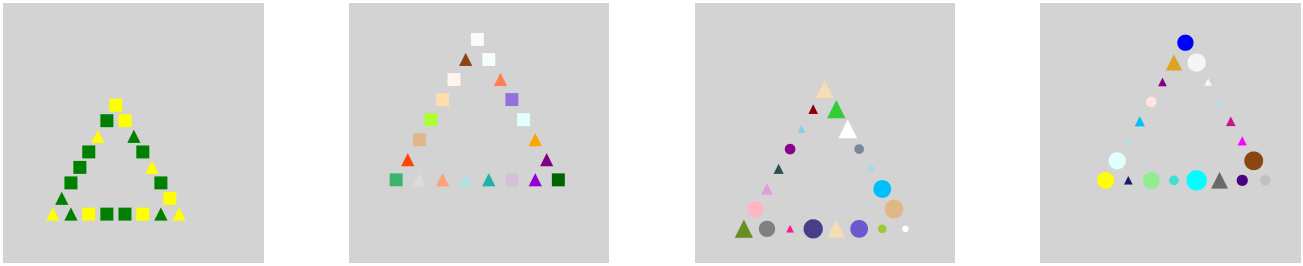
**GPT-5** The induced logic rules by GPT-5 are shown in Listing 3. The first rule correctly identifies that each proximity group must contain a triangle, but it incorrectly constrains the group size to exactly two, whereas the ground truth allows any size of two or more. The second rule does not match the task semantics and is therefore incorrect.

Listing 3: Rules induced by GPT-5 on Example Task 1 (reformated by authors)

```
Grouping by proximity.
1. Every proximity group must be a pair of exactly two shapes:
one triangle and one non-triangle (circle or square).
2. No proximity group may contain two non-triangles or
have more/less than two members.
```

### H.2 EXAMPLE TASK 2

This is a task called *Shape of Shape* following *closure* principle. The ground-truth rule is that objects form the shape of a triangle; all the objects have the same width and height. Fig. 10 presents the positive and negative examples.



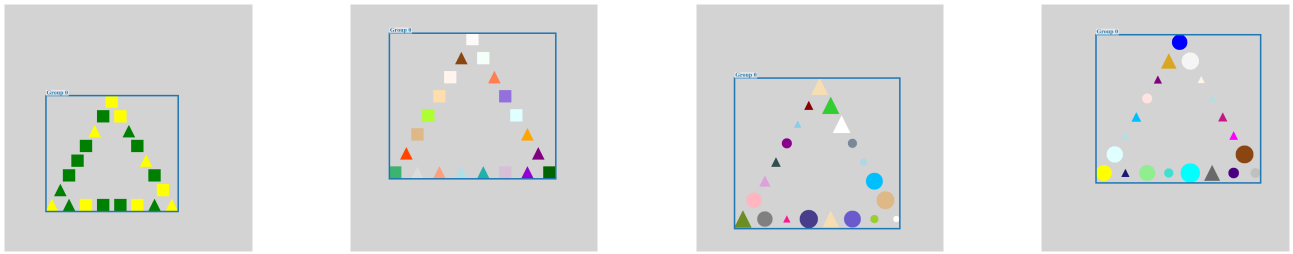
(a) Positive (b) Positive (c) Negative (d) Negative

Figure 10: Big Triangle: all the objects in the image have same size.

**GRM.** The induced rule from GRM is shown in Listing 4. The target rule is successfully identified by GRM. The groups detected by the GRM are shown in Fig. 11.

Listing 4: Rules induced by GRM on Example Task 2

```
% Image-level rules
image_target(X) :- unique_sizes(I).
   [confidence=1.000, scope=image]
group_target(G,X) :- unique_sizes(G).
   [confidence=1.000, scope=image]
% Existential rule
group_target(G,X) :- unique_sizes(G).
   [confidence=1.000, scope=existential]
```



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

Figure 11: GRM grouping results of task 2 examples

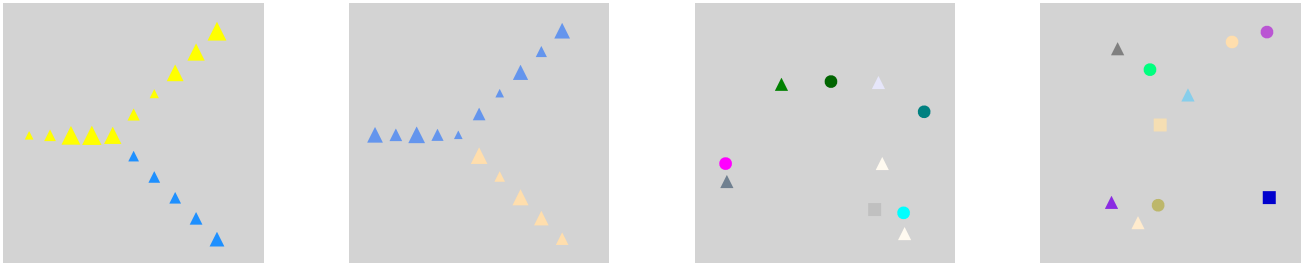
**GPT-5.** The induced rules by GPT-5 are shown in Listing 5. Rules 1–4 are logically correct but not target rules, as they hold for both positive and negative cases. Rules 5 and 6 are incorrect and do not reflect the task semantics.

Listing 5: Rules induced by GPT-5 on Example Task 2 (reformated by authors)

```
1. Objects are positioned so that their arrangement forms an implied
closed triangular contour (three sides meeting at three vertices);
2. Small gaps are allowed, but each side is perceptually continuous;
3. Most elements lie on the triangle perimeter;
4. The interior is largely empty.
5. Colors, shapes, sizes, and groupings of the objects are irrelevant.
6. Images that do not produce a clear, closed triangular outline
(scattered points, missing sides, or filled interiors) are negative.",
```

### H.3 EXAMPLE TASK 3

This is a task called *One Splits Two* following *continuity* principle. The ground-truth rule is that objects with diverse sizes are arranged along a continuous path that later splits into two directions. Fig. 12 presents the positive and negative examples.



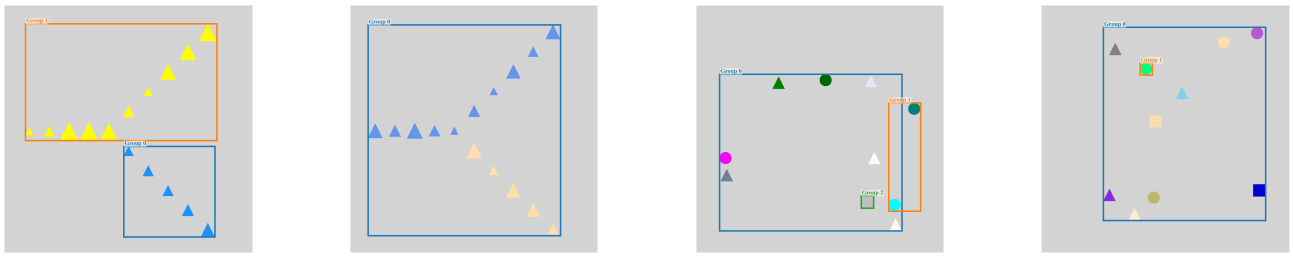
(a) Positive (b) Positive (c) Negative (d) Negative

Figure 12: Examples of task *One Splits Two*

**GRM.** The induced rules from GRM are shown in Listing 6. The first rule focuses on the $x$-positions of two objects within a group; although unexpected, it still fits both training and test sets, illustrating that high accuracy does not always imply correctness for the expected reasons. The advantage of GRM is that such rules are interpretable and auditable: unlike black-box models, unsafe or spurious rules can be manually removed or their confidence reduced by adding targeted training examples. The second rule successfully matches the target rule, and the corresponding grouping results are illustrated in Fig. 13.

Listing 6: Rules induced by GRM on Example Task 3

```
% Image-level rules
image_target(X) :- mirror_x(O1,O2), same_color(O1,O2).
   [confidence=1.000, scope=image]
% Group-level rules
group_target(G,X) :- diverse_sizes(G).
   [confidence=1.000, scope=universal]
```



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

Figure 13: GRM grouping results of task 3 examples

**GPT-5.** The induced rules by GPT-5 are shown in Listing 7. Rules 1–4 are logically valid, with Rule 2 matching the target, but the coverage is limited: precision reaches 1.0 while recall remains at 0.3, showing that even when the correct rule is discovered GPT-5 does not guarantee high recall.

Listing 7: Rules induced by GPT-5 on Example Task 3 (reformated by authors)

```
1. The scene contains exactly two groups.
2. Each group is homogeneous:
all members share the same shape and the same color.
3. Members of a group are arranged along a single smooth, continuous path
(straight or gently curved), showing clear positional continuity.
4. Along each path the sizes vary gradually in one direction
(monotonic size progression).
```

# I TASK SOLVING TIME ANALYSIS

Fig. 14 reports the average time required by GRM to induce target rules across the five Gestalt principles, with comparisons to GPT-5 and InternVL3-78B.

InternVL3-78B exhibits the most stable efficiency, requiring about 15s across all principles. GPT-5 is slower and more variable, averaging around 100s per task. On *symmetry*, GPT-5 exceeds 150s, consistent with its weaker accuracy: when uncertain, the model takes longer before committing to a prediction.

GRM is generally efficient, with rule induction on *proximity*, *closure*, and *continuity* completed in under 10s. On *symmetry*, GRM requires about 35s. The longer time is due to extended search: when no high-confidence rules are quickly available, the search process continues until a satisfactory candidate is identified *or* the maximum extended step is exceed.

The main outlier is *similarity*, where GRM averages nearly 90s. Unlike other principles, similarity depends almost entirely on color and size cues, with minimal reliance on positional features. This reduces grouping accuracy and often yields many spurious groups. A larger number of candidate groups substantially enlarges the space of possible group-level rules, thereby increasing induction time.

In summary, most GRM tasks can be solved within seconds to one minute, but tasks with weaker grouping cues or more ambiguous structures can extend to several minutes. These results highlight the impact of grouping quality on symbolic reasoning efficiency, and suggest that designing more robust symbolic features and search strategies is a promising direction for improving scalability.

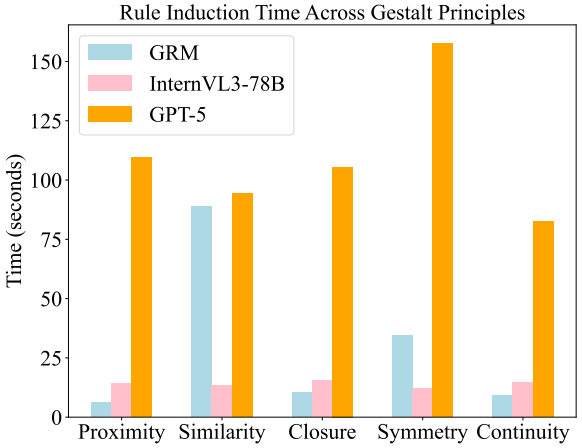



Figure 14: **Task Solving Time Comparison.** The time is measured from the start of image input to the completion of rule induction.

# J DEPTH AS A MISSING CUE FOR REAL-WORLD GROUPING

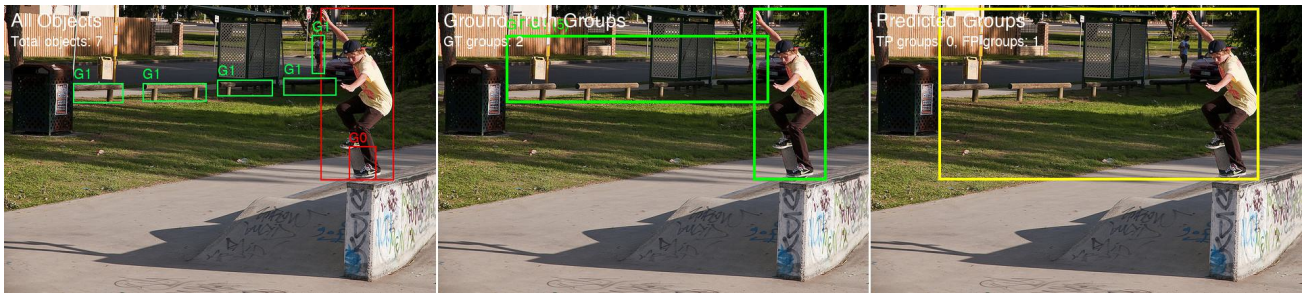



Figure 15: From left to right: labeled objects, labeled groups, predicted groups by a small MLP

In this section, we investigate the necessity of depth information for perceptual grouping in real-world images. We randomly select 100 images from the COCO dataset and keep the perception component fixed, operating directly on ground-truth bounding boxes to isolate the grouping stage.

For grouping, we employ a small *MLP-based grouper*, consistent with the GRM grouping module. The model takes as input the center coordinates, width, and height of two bounding boxes, together with a simple contextual encoding of neighboring objects, and predicts whether the pair belongs to the same group under the proximity principle.

**Qualitative Observation.** A characteristic failure mode emerges: the proximity grouper frequently collapses all objects into a single group. Fig. 15 illustrates a representative example. A person riding a skateboard appears in the foreground, while another person stands far in the background but shares a similar horizontal alignment in the 2D image plane. Four benches are positioned mid-left and also lie in the background. According to ground-truth grouping, the four benches and the distant pedestrian form one group (shared background plane), while the foreground skateboarder and skateboard form another. However, the MLP predicts that all objects belong to a single group. This failure occurs because grouping decisions rely solely on 2D geometric cues (bounding box centers and sizes), which cannot distinguish foreground–background separation.

**Limitations of 2D Image Coordinates for Capturing 3D Grouping Structure** Human Gestalt grouping in natural scenes is grounded in 3D structure: objects share supporting surfaces, occupy similar depths, or form physical assemblies (e.g., a person and their skateboard). In contrast, COCO provides only 2D bounding boxes. When annotators define groups, they implicitly rely on 3D scene understanding and semantics, whereas the grouping model receives only 2D geometric features (position, size, aspect ratio).

This creates an inherent information mismatch between the definition of ground-truth groups and the model's available inputs. In Fig. 15, the benches and distant pedestrian form a coherent 3D configuration at the far end of the scene. Yet under 2D projection, perspective effects compress depth, making foreground and background objects appear spatially close. As shown in Fig. 16, a top-view representation clearly reveals two proximity-based groups separated along the depth axis, which makes the information entirely absent from the image plane and its annotations.

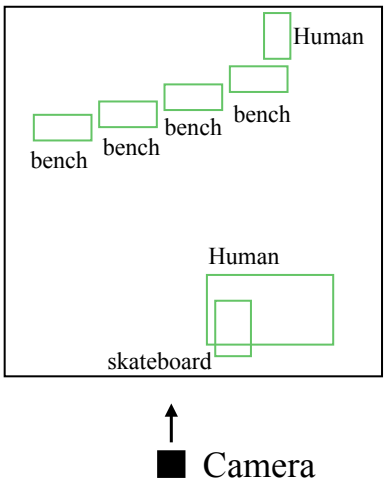



Figure 16: Top view of the Skateboard example scene in Fig. 15

Consequently, grouping performance on COCO-style images cannot be interpreted as a clean evaluation of Gestalt principles. Many apparent errors stem from missing depth cues rather than deficiencies in the grouping mechanism itself. For this reason, we use COCO only as a qualitative case study, and base our quantitative evaluation on synthetic stimuli where geometry, grouping, and supervision are fully aligned.

Extending GRM to 3D object-centric representations (e.g., via estimated depth or reconstructed scene structure) is a natural next step, and would allow real-image grouping to be evaluated under conditions closer to human perceptual grounding.

## K CLEVR-STYLE 3D EXTENSION OF ELVIS TASKS

To evaluate robustness under 3D geometry and occlusion, we convert 10 ELVIS tasks (two per Gestalt principle) into CLEVR-style synthetic scenes rendered with perspective projection and object-level occlusion. The original group-centric reasoning structure is preserved, while depth compression, viewpoint variation, and partial visibility introduce geometric challenges absent in the planar 2D setting. Representative examples are shown in Figure 17.

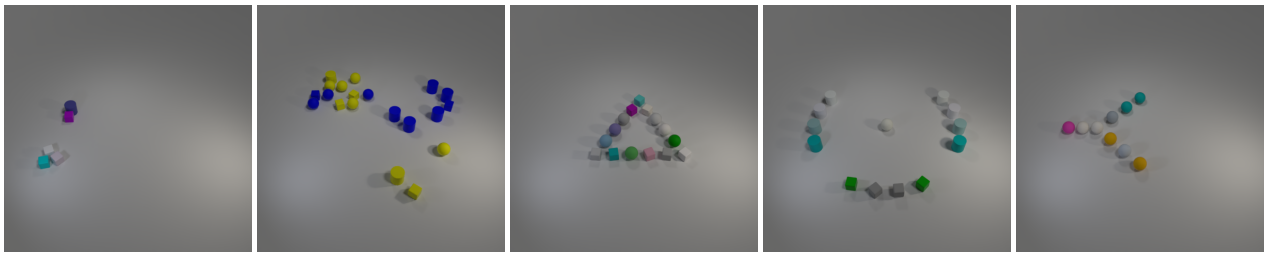


Figure 17: Representative CLEVR-style 3D scenes for proximity, similarity, closure, symmetry, and continuity. Scenes are rendered with perspective projection and include object-level occlusion, introducing depth compression and viewpoint-dependent distortion compared to the original 2D ELVIS layouts.

For each principle, two tasks are constructed in the CLEVR style. The experimental protocol mirrors the ELVIS setting: each task provides train/test splits with binary classification labels and ground-truth annotations for object attributes and group membership (10 examples per split).

The pipeline follows two stages. First, a neural group scorer is trained using ground-truth group annotations to learn pairwise object affinities under the corresponding Gestalt principle, with balanced positive and negative training pairs when available. Second, for each task, ground-truth object detections are provided to GRM; the learned group scorer predicts perceptual groupings, after which weighted logical rules are induced over object-level predicates (e.g., shape, color, position, size) and group-level predicates (e.g., group size, principle identity, diversity measures such as `diverse_shapes`, and `same_group_counts`). All the required predicates are listed in Tab. 8.

## L FIRST-ORDER LOGIC DETAILS

The symbolic reasoning component of our framework is formally grounded in First-Order Logic (FOL). In FOL, a *Language* $\mathcal{L}$ is a tuple $(\mathcal{P}, \mathcal{F}, \mathcal{C}, \mathcal{V})$, where $\mathcal{P}$ is a set of predicates, $\mathcal{F}$ a set of function symbols (functors), $\mathcal{C}$ a set of constants, and $\mathcal{V}$ a set of variables. A *term* is defined recursively: a variable is a term, a constant is a term, and if $f$ is an $n$-ary function symbol and $t_1, \ldots, t_n$ are terms, then $f(t_1, \ldots, t_n)$ is a term. An *atom* is a formula $p(t_1, \ldots, t_n)$, where $p$ is a predicate symbol (`closeby`) and $t_1, \ldots, t_n$ are terms. A *ground atom* or simply a *fact* is an atom with no variables (`closeby(obj1, obj2)`). A *literal* is an atom ($A$) or its negation ($\neg A$). A *clause* is a finite disjunction ($\vee$) of literals. A *ground clause* is a clause with no variables. A *definite clause* is a clause with exactly one positive literal. If $A, B_1, \ldots, B_n$ are atoms, then $A \vee \neg B_1 \vee \cdots \vee \neg B_n$ is a definite clause. We write definite clauses in the form of $A \leftarrow B_1, \ldots, B_n$, where the comma represents conjunction ($\wedge$), making the expression equivalent to $A \vee \neg(B_1 \wedge \ldots \wedge B_n)$. $A$ is the rule *head*, and the set $\{B_1, \ldots, B_n\}$ is its *body*. To distinguish their roles, we refer to definite clauses manipulated during intermediate learning steps simply as *clauses*, reserving the term *rules* for the final explanatory solutions.