6. SUPPLEMENTARY MATERIALS FOR "NSI-PEP: BENCHMARKING MULTIMODAL AI ON NUMBER–SHAPE INTEGRATION IN K–12"

6.1. Evaluation Details

In this study, all questions were derived from original textbook problems and subsequently adapted to suit the requirements of multimodal model inputs. Each "image + corresponding text question" pair was stored in a unified JSON format, containing the image name, a unique question_id, and the English version of the question. For example:


```
{"image": "1.png", "question_id": 0, "text": "What are the lengths of the blue bars in the
    three images?"}
```


After completing the aforementioned preparations, we designed the model inputs and processed them through a standardized pipeline. This pipeline consisted of three main steps. First, input setup: all images and JSON files were stored in a single folder and collectively served as inputs for the models. Second, model inference: under identical hardware and inference settings, the unified inputs were sequentially fed to the 13 multimodal models, and outputs were collected, including fields such as question_id and answer_text. Third, result stability check: to reduce variability in model outputs, each question was inferred twice. When both outputs were consistent, that answer was selected as the final output; otherwise, the answer whose semantics were closest to the reference answer was retained.

The evaluation procedure combined automated and manual steps to ensure reliability and reproducibility. Model outputs were first automatically scored using Doubao AI according to a pre-designed rubric aligned with reference answers. Subsequently, manual verification was performed to resolve ambiguities and confirm scoring accuracy. For reference, the full prompt used during the evaluation process is provided here:

"The answers generated by open-source multimodal large models are based on the number-shape integration dataset. Some responses are open-ended and may not exactly match the reference answers, but convey equivalent meanings. To evaluate accuracy, the grading should be based on semantic equivalence rather than string matching, with three levels: completely incorrect (0), partially correct (1), and completely correct (2)."

6.2. Knowledge Points in Each Knowledge Domain

Table 3: **Knowledge Points of Mathematics Domains**

Knowledge Domain	**Knowledge Module**	**Specific Knowledge Points**
Measurement Domain	Basic Concepts of Measurement	- Recognition and use of common measuring tools (e.g., ruler, balance, measuring cup). - Distinction between standard and non-standard units. - Understanding of measurement errors and precision (e.g., reading aligned to the specified unit).
	Length Measurement	- Standard usage of rulers and correct reading methods. - Cognition of length units (millimeter, centimeter, decimeter, meter, kilometer). - Estimation of object length (e.g., pencil, pin, needle). - Special cases of length measurement (e.g., total length of connected iron rings).
	Weight Measurement	- Recognition of weighing instruments (e.g., platform scale, spring scale). - Cognition of weight units (gram, kilogram, ton). - Estimation of object weight (e.g., cucumber, oil barrel, animal). - Weight accumulation and limit judgment (e.g., bridge load, vehicle load).
	Area and Volume Measurement	- Cognition of area units (cm^2, m^2, hectare, km^2). - Area estimation in practical contexts (e.g., extrapolating population capacity from 1 m^2 to larger areas). - Cognition of volume units (cm^3, dm^3, m^3). - Calculation of areas of regular figures (e.g., deriving the area of a square from its side length).
	Time Measurement	- Cognition of time units (second, minute, hour). - Estimation of travel time across different transportation modes (e.g., walking, cycling, driving). - Simple reasoning on the relationship between time and speed.
	Unit Conversion and Application	- Mastery of unit conversion relations (e.g., 1 ton = 1000 kg, 1 hectare = 10,000 m^2). - Rationality judgment of unit usage (e.g., 13 m vs. 13 cm for a flagpole). - Mixed-unit calculations (e.g., converting ring width in mm to total length in cm). - Matching units to practical contexts (e.g., step width 30 cm vs. 30 dm).
	Integration of Measurement and Graphics	- Graph-assisted reading of scales and measurements. - Application of graphical aids in area estimation (e.g., using grids for square meter–hectare conversion). - Interpretation of graphs indicating measurement constraints (e.g., bridge load limit).
	Application Scenarios of Measurement	- Object weighing in real-life contexts (e.g., determining whether an object exceeds the maximum scale capacity). - Transport and load-limit estimation (e.g., vehicle or coal transport limits). - Special measurement problems requiring combined reasoning (e.g., calculating the total length of an iron ring).
Fraction Operations Domain	Basic Concepts of Fractions	- Cognition of unit "1" and the relationship between whole and parts. - Meaning of fractions. - Distinction between proper and improper fractions. - Fundamental property of fractions (the value remains unchanged when numerator and denominator are multiplied/divided by the same nonzero number). - Relationship between fractions and division.
	Core Fraction Operations	- Fraction addition and subtraction (common denominator and different denominator cases). - Fraction multiplication (fraction × integer, fraction × fraction). - Fraction division (fraction ÷ integer, fraction ÷ fraction). - Mixed fraction operations with order of precedence (including brackets).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	Knowledge Module	Specific Knowledge Points
		- Fraction–decimal conversion: transforming fractions into decimals and vice versa, and comparing their magnitudes.
	Fractions in Graphical Representations	- Representing fractions through shaded parts of figures. - Identifying fractional proportions in geometric figures (e.g., tangram, shaded areas). - Representing fractions on the number line. - Exploring fractional relationships in figure transformations (e.g., scaling up or down).
	Application Scenarios of Fractions	- Engineering problems (e.g., work/rate problems such as flood discharge or road construction). - Distribution problems (e.g., dividing candies, water, or other resources). - Ratio and proportion problems (e.g., fractional area of farmland distribution). - Periodicity and least common multiple problems (e.g., bus departure intervals, running encounters).
	Number Theory Connections	- Greatest common divisor (GCD) and its applications (e.g., cutting rods, dividing rectangles into equal squares). - Least common multiple (LCM) applications (e.g., bus departure synchronization). - Judgment of common factors and multiples. - Pattern exploration in fractions (e.g., invariant properties in fraction sums, filling fraction sequences).
Probability Domain	Fundamental Probability Concepts	- Cognition of certain and random events. - Definition of likelihood and the degree of possibility. - Probability range and interpretation ($0 \leqslant P \leqslant 1$).
	Probability Judgment and Comparison	- Identifying events that are certain to occur. - Identifying events that are impossible to occur. - Comparing and ordering likelihoods across events (e.g., color sectors on a turntable, ball-drawing scenarios).
	Core Probability Calculations	- Classical probability: probability derived from equally likely outcomes (e.g., probability of rolling “2” on a 12-faced die). - Geometric probability: probability based on area/length ratios (e.g., probability of pointer landing on a colored region of a spinner). - Compound events: probability of multi-step or combined events (e.g., drawing two cards to form a complete image).
	Graphical and Tabular Representations of Probability	- Spinner design and probability analysis (coloring sectors to control likelihoods). - Tree diagram representation of multi-path processes (e.g., ants moving along branches). - Tabular listing method for enumerating all outcomes (e.g., rock–paper–scissors cases).
	Application Scenarios of Probability	- Ball/chess piece drawing problems (probability of selecting a red, blue, or green ball). - Card-related problems (e.g., probability of drawing a face card from 13 spades). - Image reconstruction problems (e.g., probability of reassembling a complete picture from fragments).
Time Domain	Basic Time Concepts	- Recognition of standard time units (hours, minutes, seconds). - Understanding the movement rules of hour and minute hands. - Cognition of 12-hour and 24-hour time systems.
	Clock Reading and Calculations	- Accurate clock reading from analog or digital clocks. - Duration calculation between two time points (e.g., movie screening length). - Relationship between rotations of hour and minute hands (e.g., hour hand 6 circles $\rightarrow$ minute hand 72 circles).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	**Knowledge Module**	**Specific Knowledge Points**
	Time Application Scenarios	- Behavior–time rationality (matching activities with appropriate time). - Stopwatch applications (e.g., judging athletes' rankings by recorded time). - Future time deduction (e.g., determining the time after 5 or 10 minutes).
	Time Rules and Matching	- Completing missing values based on observed temporal patterns. - Matching time-related information with contextual pictures or events. - Unit conversion between seconds, minutes, and hours.
Expansion of Number Concepts Domain	Algebra Foundation & Representation	- Using letters to represent numbers (e.g., rectangle perimeter 2(a+b), total fish = bags × a). - Writing and evaluating algebraic expressions (e.g., pipe volume π a(R^2r^2), surface area of open box). - Operations with polynomials and rational expressions; conditions for rational expressions to be valid (e.g., denominator $\neq$ 0, fraction equals zero). - Representing arithmetic laws symbolically (e.g., commutative law a+b=b+a).
	Equations & Inequalities	- Establishing and solving linear equations in one variable (e.g., calculating number of soccer ball panels). - Solving systems of linear equations in two variables (e.g., crop planting allocation). - Applying quadratic equations to solve geometric problems (e.g., constructing open box from square cut, area of rectangle). - Representing inequality solutions on number lines.
	Functional Relationships	- Direct proportion (e.g., fuel consumption vs. distance, t=s/v). - Inverse proportion (e.g., funnel area S=3/d). - Linear functions (e.g., ball velocity v=2t). - Quadratic functions (e.g., parabolic bridge, maximizing rectangular garden area).
	Geometry–Algebra Integration	- Algebraic expressions of geometric measures (e.g., annulus area π (R^2r^2), trapezoid passage width). - Application of Pythagorean theorem (e.g., isosceles right triangle side AB=2a). - Geometric scaling and proportion (e.g., enlarging by 4:1, golden ratio division). - Coordinates and number line problems (e.g., rolling circle displacement, relative positions of points).
	Ratios & Percentages	- Ratio-based distribution (e.g., concrete mix 2:3:5, solution dilution 1:4). - Percentage calculations (e.g., increase rate of afforestation, shaded area percentage). - Scale conversions (e.g., ladybug diagram 5mm $\rightarrow$ 5cm = 1:10, room plan 6m $\rightarrow$ 4cm = 1:150). - Judging proportional relationships (e.g., height vs. shadow length, pressure vs. contact area).
	Pattern Exploration & Applications	- Discovering geometric patterns (e.g., arranging n squares requires 3n+1 sticks, counting black dots in the nth figure). - Sequences and summation (e.g., calculating number of logs, triangular numbers). - Optimization problems (e.g., maximizing area of rectangle CDEF, minimizing square EFGH). - Balance and equilibrium (e.g., lever balance equations, weight distribution).
Mathematical Ideas and Methods Domain	Counting & Combinatorial Thinking	- Matching and combination problems (e.g., pairing clothes and pants, selecting fruits and vegetables). - Permutation problems (e.g., arranging characters, distributing books). - Addition and multiplication principles for counting (e.g., coloring problems, path counting).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	Knowledge Module	Specific Knowledge Points
	Optimization & Overall Planning	- Time planning and scheduling (e.g., arranging tea-making steps, minimizing cooking time). - Resource allocation (e.g., distributing passengers in boats, purchasing basketballs and volleyballs). - Scheme optimization strategies (e.g., minimizing moves to transfer beads, defective item identification).
	Logical Reasoning Methods	- Sudoku-type reasoning (e.g., filling numbers 1–9 without repetition). - Grid logic reasoning (e.g., ensuring uniqueness of numbers 1–4 in each row/column). - Pattern deduction (e.g., extending Pascal's triangle, algebraic expansion such as $(a+b)^2$).
	Geometric Counting & Patterns	- Counting geometric figures (e.g., number of rectangles, squares, triangles in a grid). - Tree planting problems in circular and linear arrangements (e.g., planting around a pond, along a road). - Geometric pattern exploration (e.g., number of people seated when tables are joined, distribution of colors in subdivided shapes).
	Pigeonhole Principle & Set Applications	- Basic pigeonhole principle (e.g., pencils into holders, books into drawers). - Worst-case principle (e.g., ensuring identical chopsticks selection, dart score estimation). - Set and overlap problems (e.g., filling Venn diagrams with animals, counting fruit types collected over two days).
Statistics Domain	Statistical Chart Cognition & Interpretation	- Line charts (e.g., temperature variation, growth of high-speed rail mileage). - Bar charts (e.g., vehicle numbers, breakfast preferences). - Sector charts (e.g., air composition, expenditure proportions). - Statistical tables (e.g., part length, seedling height, annual employee profit).
	Data Feature Analysis	- Central tendency (e.g., mean, median). - Dispersion degree (e.g., variance of shooting scores, defective product counts). - Data distribution (e.g., speed distribution, cucumber root counts). - Proportion analysis (e.g., proportion of qualified cities, nitrogen ratio).
	Sampling & Sample Estimation	- Random sampling (e.g., product quality checks, agricultural sampling). - Population estimation by sample (e.g., estimating average part length or tree circumference). - Sample size design (e.g., tasting by 36 people, measuring 10 wheat seedlings). - Sampling representativeness (e.g., using intersection traffic speeds to represent a road segment).
	Statistical Calculation Application	- Mean calculation (e.g., average temperature, average speed, average age). - Median determination (e.g., median in food evaluation, median part length). - Variance calculation (e.g., variance of defective machines, variance of shooting scores). - Total estimation (e.g., nitrogen volume, number of qualified cities, food expenditure).
	Data Trend & Comparison	- Trend judgment (e.g., GDP growth, fuel vehicle sales trend). - Increase/decrease range (e.g., annual mileage increments, yearly sales growth). - Inter-group comparison (e.g., comparing seedling height or defective products between two groups). - Gap analysis (e.g., urban–rural population gap, difference between fixed and mobile phone users).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	Knowledge Module	Specific Knowledge Points
Measurement and Calculation of Graphics Domain	Core Measurement of Plane Figures	- Basic figure area calculation (e.g., parallelogram, trapezoid, square). - Composite figure area calculation (e.g., shapes with cutouts or combined figures, trapezoid lawns with a pool). - Perimeter calculation (e.g., combined squares, trapezoid fence length). - Irregular figure estimation (e.g., grid method for leaf area, approximate colored area).
	Key Measurement of Solid Figures	- Volume and surface area of solids (e.g., combined cubes, semi-cylinders with cubes). - Solid unfolding and folding (e.g., mapping cuboid faces from nets, constructing geometric solids). - Counting colored faces (e.g., 2/3 faces colored on combined cubes). - Cylinder and cone measurement (e.g., semi-cylinder + cube surface area, cone slant height).
	Geometric Properties & Application	- Pythagorean theorem (e.g., original rod length, ladder sliding distance, distance between two points). - Equal area transformations (e.g., triangles with same base and height, shaded triangles in parallelograms). - Axial symmetry (e.g., completing symmetric figures, counting symmetry axes). - Translation and rotation (e.g., moving flags, transforming figure A to B).
	Measurement Calculation Methods	- Inverse formula use (e.g., calculating height from known parallelogram area, distances from trapezoid area). - Cut-and-complement / transformation methods (e.g., computing shadow area, estimating log volumes with trapezoid formulas). - Unit conversion (e.g., m^2 ha, cm m). - Extreme value calculation (e.g., largest square cut from rectangle, largest parallelogram cut from trapezoid).
	Graphic Observation & Operation	- Three views (e.g., front/left/top view of solids, calculating number of small cubes from views). - Graphic counting (e.g., counting shapes in tangram, basic shapes in composite figures). - Pattern deduction (e.g., number of sticks for n hexagons, black dot count in nth figure). - Graphic comparison (e.g., comparing perimeters/areas, assessing size of colored regions).
Relationships of Graphics Domain	Core Relationships of Lines & Angles	- Parallel line judgment and properties (e.g., proving parallelism using corresponding angles, calculating alternate interior angles). - Perpendicular line properties (e.g., lines perpendicular to the same line are parallel, finding angles with $EO \perp AB$). - Special angle relationships (e.g., vertical angles equality, sum of adjacent angles 180°). - Three-line eight-angle recognition (e.g., listing corresponding, alternate, and interior angles, identifying $\angle AOD$).
	Key Types of Graphic Transformations	- Translation transformation (e.g., moving paths to calculate lawn area, translating quadrilaterals for constructions). - Homothety (scale) transformation (e.g., drawing a figure with similarity ratio 0.75 from center Q, judging scale factor in similar squares). - Axial symmetry transformation (e.g., completing symmetric figures, folding rectangles to form isosceles triangles). - Projection transformation (e.g., drawing orthographic projections of cylinders, connecting objects and projections).
	Graphic Similarity & Symmetry	- Similar triangles judgment and properties (e.g., $DE \parallel BC$ proving similarity, finding heights or side lengths).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	Knowledge Module	Specific Knowledge Points
		- Similar polygons judgment and properties (e.g., calculating side ratios of pentagons, determining similarity of quadrilaterals). - Congruent figure applications (e.g., proving equality in equilateral triangles, perpendicular bisector properties). - Similarity ratio calculations (e.g., scaling line segments in similar figures, comparing $\triangle$COD and $\triangle$AOB). - Axial symmetry figure judgment (e.g., determining symmetry of pentagons, rectangles, regular polygons). - Symmetry axis drawing and counting (e.g., drawing all axes of symmetry, deducing number of axes for regular n-gons). - Axial symmetry property applications (e.g., constructing triangles symmetric about a line, measuring angles in mirror images). - Symmetric figure correspondence (e.g., finding coordinates after reflection over x-axis, identifying symmetric figures).
	View Relationships of Solid Figures	- Three-view recognition and drawing (e.g., identifying shapes and counting cubes from front/left/top views, drawing three-view diagrams). - Three-view and unfolding diagram transformation (e.g., constructing nets from three-view diagrams, calculating volume and surface area). - View and object relationship (e.g., computing steel plate area from can three-view, matching three-view diagrams to real objects). - View and projection connection (e.g., drawing orthographic projections along projection lines, matching three-view diagrams to object projections).
Recognition of Graphics Domain	Core Cognition of Plane Figures	- Basic figure recognition (e.g., triangles, parallelograms, trapezoids, circles, squares, rectangles). - Special angle recognition (e.g., right angles, acute and obtuse angles, straight angles, central and inscribed angles). - Polygon features (e.g., equilateral triangle, regular hexagon, interior angle sums, symmetry of squares). - Composite plane figure cognition (e.g., overlapping circles and squares, figures including sectors).
	Key Cognition of Solid Figures	- Basic solid figure recognition (e.g., cube, cuboid, cylinder, cone, prism). - Solid figure composition (e.g., cubes composed of smaller cubes, cylinders formed by rotating rectangles). - Solid figure views (e.g., front/top/side observation, matching three-view diagrams to real objects). - Unfolding and folding (e.g., analyzing cube nets, developing cylinder side surfaces into rectangles).
	Graphic Elements & Relationships	- Line relationships (e.g., identifying and constructing parallel and perpendicular lines, distinguishing segments, rays, and lines). - Angle relationships (e.g., vertical angles equality, complementary and supplementary angles, corresponding and alternate interior angles). - Graphic position relationships (e.g., shortest distance from point to line, tangency or intersection between circle and line). - Congruence and similarity of figures (e.g., determining congruent triangles, analyzing properties of similar polygons).
	Graphic Operation & Transformation	- Graphic construction (e.g., drawing specified angles, perpendicular lines through points, completing rectangles or parallelograms). - Graphic folding and cutting (e.g., folding rectangles to form isosceles triangles, cutting squares to form regular octagons). - Graphic rotation and translation (e.g., rotating rectangles to form cylinders, translating line segments to form plane figures). - Graphic symmetry (e.g., drawing the other half of axisymmetric figures, determining number of axes of symmetry).
	Graphic Measurement & Calculation	- Plane figure measurement (e.g., areas of triangles, parallelograms, circles; perimeters of rectangles, squares, circles).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	Knowledge Module	Specific Knowledge Points
		- Solid figure measurement (e.g., volumes and surface areas of cubes, cuboids, cylinders; lateral area of cones). - Special graphic measurement (e.g., area of circular rings, sectors, perimeters of regular polygons). - Practical scene measurement (e.g., volume of oil barrels, area of chimney caps, length of pedestrian bridge ramps).
Movement of Graphics Domain	Graphic Translation	- Translation recognition (e.g., determining whether small houses or fish overlap after translation, distinguishing translation from rotation). - Translation direction and distance (e.g., number of grid steps for moving small animals, translating a boat 4 grids to the right). - Translation construction (e.g., moving figures up 3 grids and left 8 grids, assembling basic shapes using translation). - Translation properties (e.g., shape and size remain unchanged after translation, corresponding segments are parallel and equal).
	Graphic Rotation	- Rotation recognition (e.g., identifying whether a pattern is obtained by rotation, determining windmill rotation direction). - Rotation three elements (e.g., identifying center, angle, and direction of rotation, such as clockwise 90° around point O). - Rotation construction (e.g., rotating a triangle 30° clockwise around point S, rotating around point A 90°). - Rotation properties (e.g., shape and size remain unchanged, distances from points to rotation center are equal). - Rotation overlap angle (e.g., smallest angle to rotate a pattern so that it coincides with itself).
	Graphic Axial Symmetry	- Axial symmetry recognition (e.g., identifying whether a figure is axisymmetric, distinguishing symmetric from non-symmetric figures). - Axis of symmetry (e.g., drawing axes of symmetry, determining the number of axes). - Axial symmetry construction (e.g., completing a figure according to its axis, drawing the other half of an axisymmetric figure). - Axial symmetry properties (e.g., the axis perpendicularly bisects corresponding lines, corresponding segments and angles are equal).
	Graphic Central Symmetry	- Central symmetry recognition (e.g., identifying whether a figure is centrally symmetric). - Symmetric center and points (e.g., locating symmetric points of quadrilaterals, identifying the center of symmetry). - Central symmetry construction (e.g., drawing a figure symmetric about point O). - Central symmetry properties (e.g., the center bisects corresponding lines, 180° rotation maps the figure onto itself).
	Comprehensive Application of Graphic Movements	- Movement combination (e.g., combining translation and rotation to form a square, transforming triangle ABC to DEC using translation, rotation, or axial symmetry). - Paper cutting and folding (e.g., folding paper three times to cut a pattern, matching cut figure to original paper). - Practical applications (e.g., analyzing rotational correspondence of rods, using rotation to prove segment equality).
Graphics and Coordinates Domain	Coordinate Basics	- Coordinate representation of points (e.g., expressing positions of pieces or objects using ordered pairs (x,y)). - Plotting points (e.g., plotting A(4,4), B(2,2) on a coordinate plane). - Coordinate system recognition (e.g., distinguishing horizontal/vertical axes and the origin). - Reading and writing coordinates (e.g., recording positions of chess pieces, fruits, or objects as (x,y) pairs).
	Coordinates & Graphic Position	- Vertices coordinates (e.g., representing the vertices of a triangle with ordered pairs).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	Knowledge Module	Specific Knowledge Points
		- Position determination by coordinates (e.g., shading regions, marking points, and connecting them into closed figures). - Point deduction (e.g., deducing the coordinates of point A from B and C). - Association between coordinates and direction (e.g., determining relative directions between buildings). - Features of special point coordinates (e.g., deriving unknown vertex coordinates from given points).
	Coordinates & Graphic Movement	- Translation and coordinate change (e.g., computing new vertex coordinates after right 2/up 3 translation). - Coordinate verification after translation (e.g., checking if a point moves to (1,7) after translation). - Coordinate description of graphic movement (e.g., describing the translation of figure 1 to figure 2 using coordinate changes).
	Coordinates & Practical Scenarios	- Positioning by direction and distance (e.g., plotting a city B east-northeast 40° from city A, or a museum relative to a TV tower). - Routes and coordinates (e.g., computing visit paths of a character, describing typhoon trajectories). - Practical application of coordinates (e.g., calculating meeting times in races using start/return point coordinates). - Coordinate conversion in floor plans (e.g., computing positions given small square diagonals, directions, and distances).
Decimal Operations Domain	Decimal Basics	- Integer range judgment of decimals (e.g., determining between which two integers a decimal lies). - Decimal-fraction link (e.g., expressing shaded proportions as decimals, converting between fractions and decimals). - Decimal positioning on coordinates (e.g., marking decimals on a number line or Cartesian coordinate axis).
	Decimal Size & Representation	- Decimal comparison (e.g., comparing animal speeds, product prices, or shaded areas). - Graphic representation of decimals (e.g., shading areas to represent given decimals). - Decimal unit conversion (e.g., converting 0.4 km to 400 m, unifying speed units).
	Decimal Four Fundamental Operations	- Decimal division (e.g., computing unit price of 12 shuttlecocks costing 19.4 yuan, calculating per second speed for 400 m in 75 s). - Decimal addition & subtraction (e.g., combining lengths of bamboo poles, summing fruit box prices to check budget). - Mixed decimal operations (e.g., calculating exceeding parts in segmented billing scenarios).
	Practical Application of Decimals	- Segmented charging problems (e.g., calculating parking fees for 12.5 yuan, taxi fare for 6.3 km). - Shopping discount calculation (e.g., comparing total cost of 30 milk boxes in two supermarkets, computing discounted prices). - Travel & total quantity calculation (e.g., total distance fallen in 4 seconds, shortest path across multiple routes). - Container packaging problems (e.g., filling 2.5 kg sesame oil into 0.4 kg bottles, calculating number of bottles needed).
	Decimal Patterns Exploration	- Decimal arrangement & combination (e.g., forming all possible decimals with 4 cards). - Decimal operation patterns (e.g., deducing subsequent results based on patterns in earlier calculations). - Decimal sequence patterns (e.g., filling missing decimals according to visual patterns, ordering figures based on length patterns).
Rational Numbers and Related Operations Domain	Rational Number Basics	- Concept of rational numbers (e.g., standard quantity vs. deviation, such as 50 kg wheat as baseline, ball mass deviation).

Continued on Next Page

Table 3 – Continued from Previous Page

Knowledge Domain	**Knowledge Module**	**Specific Knowledge Points**
		- Rational number classification (e.g., distinguishing positive and negative numbers, exceeding standard = positive, insufficient = negative). - Standard quantity and deviation (e.g., using 50 kg or standard ball mass as reference to determine positive/negative).
	Number Axis & Representation	- Number axis drawing (e.g., drawing number axis and marking specified rational numbers). - Point-number correspondence on axis (e.g., determining the rational number represented by points A, B, C, D). - Positional features on the number axis (e.g., positive numbers on the right, negative numbers on the left; deviations' signs correspond to axis directions).
	Application of Rational Number Properties	- Absolute value application (e.g., comparing absolute values to find the smallest, judging closeness of ball mass to standard). - Rational number comparison (e.g., using number axis or absolute value to compare deviations and find closest to standard). - Positive-negative meaning application (e.g., distinguishing "overweight" vs. "insufficient" using positive/negative deviations).
	Practical Operations of Rational Numbers	- Addition & subtraction with rational numbers (e.g., calculating total mass of 10 wheat bags: sum deviations + 50×10). - Practical total calculation (e.g., combining standard quantity and positive/negative deviations to determine actual total mass). - Operation verification (e.g., verifying results via number axis or real-world interpretation).
Integer Operations Domain	Integer Basics	- Counting and representation of integers (e.g., counting objects, using sticks/blocks, distinguishing tens and ones). - Comparison of integers (e.g., comparing numbers of animals/objects, ordering numbers). - Composition of integers (e.g., decomposing numbers into tens and ones).
	Four Fundamental Operations of Integers	- Addition & subtraction of integers (e.g., calculating remaining or total quantities, such as 10 pears minus 4). - Multiplication & division (e.g., summing repeated addends, average distribution, e.g., 12 bamboo shoots divided among 4 plates). - Mixed operations (e.g., combined operations with brackets, following correct order of operations).
	Practical Applications of Integer Operations	- Shopping and consumption problems (e.g., calculating total price for 8 balls, comparing unit prices). - Distribution and loading problems (e.g., dividing fruits evenly, calculating container capacity for oil bottles). - Resource planning problems (e.g., determining number of trips, optimizing purchase schemes, e.g., buying seedlings with budget constraints).
	Integer Patterns & Optimization	- Pattern exploration (e.g., identifying periodic patterns, deducing subsequent results from arithmetic sequences). - Integer classification (e.g., distinguishing multiples of 2/3/5, prime vs. composite numbers). - Optimization & simple calculation (e.g., using arithmetic laws to simplify computation, "make tens" method, 24-point game strategies).

End of Table

6.3. Examples of Each Difficulty Level

Easy Level:

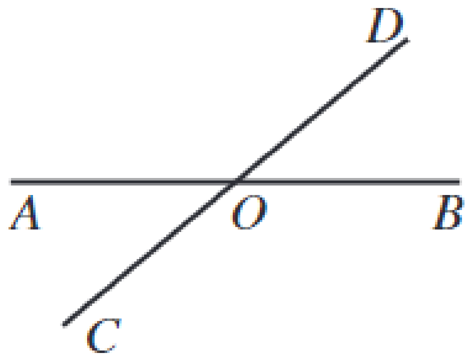

Fig. 6: **Easy Problem.**

question:If the measure of angle AOC is 50 degrees, what is the measure of angle BOD?

Medium Level:

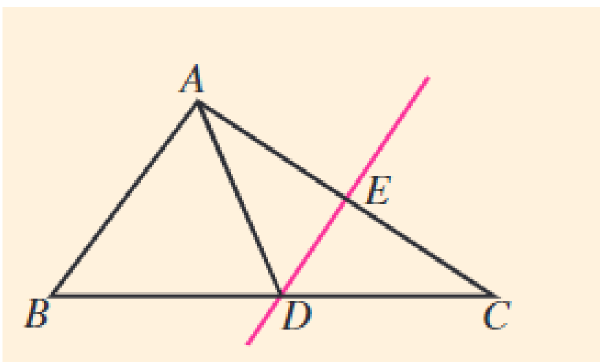

Fig. 7: **Medium Problem.**

question:In triangle ABC, DE is the perpendicular bisector of AC, AE is 3 cm, and the perimeter of triangle ABD is 13 cm. What is the perimeter of triangle ABC?

Hard Level:

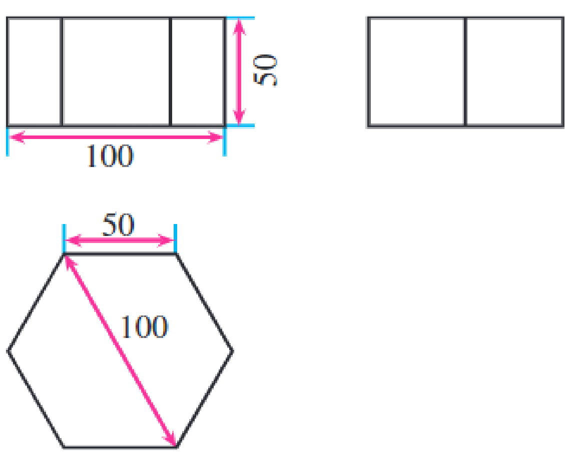

Fig. 8: **Hard Problem.**

question:Observe the picture. It shows a three-view drawing of a sealed can. According to the three-view drawing, calculate the area of the steel plate required to make each sealed can.

6.4. Distribution of Question Types for Each Knowledge Domain

The situation of the 15 knowledge areas is shown in Table 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

Table 4: **Measurement-26 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	16	61.5%	"In the figure, a cow weighs 500 kg, an oil barrel weighs 100 kg, a cement bag weighs 50 kg, and a tiger weighs 250 kg. How many of each are needed to make exactly 1 ton?"
Formula/Equation	0	0%	
Multiple Choice and Judgement	3	26.9%	"Is the height of a cat greater than or less than 1 meter?"
Open-Ended Response	7	11.5%	"In the figure, a ruler has a thickness of 1 cm. Is this reasonable? Why?"
Drawing Design	0	0%	

Table 5: **Fraction Operations-98 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	65	66.3%	"A reservoir plans to open floodgates to adjust the water level. If only Gate A is opened, the task can be completed in 8 hours; if only Gate B is opened, it can be completed in 6 hours. How many hours will it take if both gates are opened simultaneously?"
Formula/Equation	7	7.1%	"Write down the equation and calculate the result."
Multiple Choice and Judgement	5	5.1%	"Based on the figure, fill the circle with greater than, less than, or equal to."
Open-Ended Response	6	6.1%	"In the tangram, what fraction is occupied by Shape 7 and Shape 4 together, and why?"
Drawing Design	15	15.3%	"Shade three-fifths of the objects in each figure."

Table 6: **Probability-21 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	6	28.57%	"Observe the figure: An ant starts at the lower end of a branch and looks for food at the top. If the ant randomly chooses a path at each fork, what is the probability it will reach the food?"
Formula/Equation	0	0%	
Multiple Choice and Judgement	10	47.62%	"Which boxes in the picture make it impossible to draw a red ball?"
Open-Ended Response	1	4.76%	"List all possible outcomes of rock-paper-scissors in a table."
Drawing Design	4	19.05%	"Color the figure as required: the outcome must be a blue ball."

6.5. Model Information

The relevant information of the model is shown in Table 19, Fig. 9

6.6. Accuracy and Confidence Interval Calculation

We use overall accuracy and its confidence interval as the primary metrics to measure the performance of multimodal large models on K-12 number-shape integration tasks. The definitions of them are provided in Eq. (1) and (2), respectively:

$$\hat{p} = \frac{k}{N} \tag{1}$$

$$CI = \frac{\hat{p} + \frac{z^2}{2N} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{2}$$

where N is the number of questions in the test and k is the number of the correctly answered by the model. Since the accuracy commonly follows a binomial distribution $B(N, p)$, we compute its binomial confidence interval as shown in Eq. (2), where z is the 95% quantile of the standard normal distribution and $z = 1.96$. Compared with the traditional normal approximation interval, binomial confidence intervals are more accurate when the sample size is limited and the probability is low.

Table 7: **Time-19 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	15	78.9%	"When the minute hand moves from 12 to 6, how many minutes have passed? When the hour hand moves from 12 to 6, how many minutes have passed?"
Formula/Equation	0	0%	
Multiple Choice and Judgement	2	10.5%	"If the stopwatches shown belong to three athletes, which one indicates first place and which one indicates last place?"
Open-Ended Response	2	10.5%	"Are the actions of the person in the picture consistent with the time shown on the clock? Why?"
Drawing Design	0	0%	

Table 8: **Expansion of Number Concepts-144 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	65	45.1%	"One person makes 3 successful shots out of 5, and another makes 4 successful shots out of 6. What is the shooting percentage of each, and whose is higher?"
Formula/Equation	45	31.3%	"A ball starts rolling down a slope from rest, with its speed increasing by 2 each second. Find the function expression of velocity v with respect to time t."
Multiple Choice and Judgement	12	8.33%	"Which of the images is an enlargement of Image A by a ratio of 2:1?"
Open-Ended Response	14	9.7%	"Originally, there were two identical balls, and the balance was level as shown in the left picture. If one ball is removed from each side, as in the right picture, will the balance remain level? Why?"
Drawing Design	8	5.6%	"In the picture, each grid represents 1 meter. Point A is at -1, and point B is 3 meters away from point A. Mark the position of B in the picture."

Table 9: **Mathematical Ideas and Methods-42 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	34	81.0%	"A shirt and a pair of pants are matched. How many possible combinations are there in total?"
Formula/Equation	1	2.4%	"Observe the picture and directly write the expanded result of $(a+b)^2$."
Multiple Choice and Judgement	1	2.4%	"Two people are playing a card game comparing values. If Xiaohong has the set of cards shown below, does she have a chance to win?"
Open-Ended Response	5	11.9%	"In a darts competition, a player scores 41 points with 5 darts. Explain why at least one dart must score no less than 9 points."
Drawing Design	1	2.4%	"Based on the four images shown, draw the fifth image."

Table 10: **Statistics-34-Question Type Distribution(Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	26	76.5%	"To check the quality of a batch of parts, 10 were randomly selected. Their lengths were measured. Based on the data, calculate the average length of the batch."
Formula/Expression	0	0%	
Multiple Choice/Judgment	3	8.8%	"In the figure, the left column shows years, the middle column shows fuel vehicle sales, and the right column shows new energy vehicle sales. Is it correct that fuel vehicle sales increased year by year? Why?"
Open-ended Response	3	8.8%	"A food company surveyed the saltiness of a newly developed snack by inviting 36 people to taste it and give feedback: A = too salty, B = slightly salty, C = moderate, D = slightly bland, E = too bland. What is the median evaluation, and what did most people think of the taste?"
Graphing/Design	2	5.9%	"The categories on the x-axis represent breakfast types: milk, soy milk, and porridge. The y-axis represents the number of people: 6 like milk, 12 like soy milk, 24 like porridge. Draw the corresponding bar chart."

Table 11: **Measurement and Calculation of Graphics-118-Question Type Distribution(Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	95	80.5%	“A parallelogram-shaped parking space has a base of 6 meters and a height of 2.5 meters. What is its area?”
Formula/Expression	1	0.8%	“Using sticks to form hexagons: in the figure, the left column shows the number of hexagons, the middle column shows the arrangement, and the right column shows the number of sticks used. If n hexagons are formed, how many sticks are needed in total?”
Multiple Choice/Judgment	7	5.9%	“Are the shaded areas of the four figures in the picture equal? Why?”
Open-ended Response	11	9.3%	“The figure shows a rectangular wooden frame, 18 cm long and 15 cm wide. If it is stretched into a parallelogram, will its perimeter and area change? Why?”
Graphing/Design	4	3.4%	“First trace the perimeters of the three figures in yellow, then shade the areas of the three figures in green.”

Table 12: **Relationships of Graphics-149 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	48	32.2%	“To reinforce a house, a beam DE is added parallel to BC. If $\angle ABC = 31°$, what is $\angle ADE$?”
Formula/Equation	0	0%	
Multiple Choice and Judgement	41	27.5%	“Are the two angles in the figure symmetric about an axis?”
Open-Ended Response	23	15.4%	“List the corresponding angles, alternate interior angles, and same-side interior angles in the figure.”
Drawing/Design	37	24.8%	“Draw one axis of symmetry for each figure.”

Table 13: **Recognition of Graphics-441 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	205	46.5%	“A cubic fish tank has an edge length of 3 dm and no cover. What is the minimum glass area (in square decimeters) required to make it?”
Formula/Equation	5	1.1%	“The lateral surface of a cylinder is a square when unfolded. Find the ratio of the base diameter to the height of the cylinder.”
Multiple Choice and Judgement	40	9.1%	“In the figure, AB is the diameter of circle O, and AD is perpendicular to the tangent at point C. Does AC bisect $\angle DAB$?”
Open-Ended Response	81	18.1%	“Between the two types of fences shown, which is more stable, and why?”
Drawing/Design	110	25.2%	“Draw the corresponding views of the object from the top, front, and left side.”

Table 14: **Movement of Graphics-65 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	8	12.3%	“By how many degrees does the pointer in the figure rotate clockwise around point O?”
Formula/Equation	0	0%	
Multiple Choice and Judgement	14	21.5%	“Observe the figure: is the shape in the figure centrally symmetric?”
Open-Ended Response	19	29.2%	“The pattern in the right figure is formed by translating or rotating the 4 cards on the left. Describe the movement process of each card.”
Drawing/Design	24	36.9%	“Complete the axisymmetric figure in the image based on the symmetry axis.”

Table 15: **Graphics and Coordinates-26 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	7	26.9%	"In a long-distance race, the left side of the figure is the starting point and the right side is the turning point. A runner reaches 3 km from the start and returns. The leading runner runs 310 m per minute, the last runner runs 290 m per minute. After how many minutes do they meet?"
Formula/Equation	0	0%	
Multiple Choice and Judgement	2	7.7%	"Observe the figure: translating the triangle 2 units to the right and 3 units up, one vertex becomes (1,7). Is this correct?"
Open-Ended Response	10	38.5%	"How can Figure 2 be obtained from Figure 1 through translation?"
Drawing/Design	7	26.9%	"The center of the plane in the figure is City A. Draw City B located 45 km away from A in the direction 40° east of north."

Table 16: **Decimal Operations-46 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	21	45.7%	"An object falls from a height and hits the ground after 4 seconds. The distance fallen in the first second is 4.9 m, and each subsequent second it falls 9.8 m more than the previous second. What is the total height from which it fell?"
Formula/Equation	0	0%	
Multiple Choice and Judgement	8	17.4%	"The speeds of the animals in the figure are: elephant 0.4 km/min, hare 1200 m/min, horse 1170 m/min, cheetah 1.85 km/min. Which is fastest and which is slowest?"
Open-Ended Response	8	17.4%	"Compare the shaded areas of the two figures relative to the whole. Which proportion is larger and why?"
Drawing/Design	9	19.6%	"Color the areas in the figure to represent the given decimals."

Table 17: **Rational Numbers and Related Operations-6 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	0	0%	
Formula/Equation	0	0%	
Multiple Choice and Judgement	4	66.7%	"Which of the points A, B, C, D on the number line has the smallest absolute value?"
Open-Ended Response	0	0%	
Drawing/Design	2	33.3%	"Draw a number line and mark the numbers from the figure on it."

Table 18: **Integer Operations-334 Question Type Distribution (Example)**

Question Type	Quantity	Proportion	Typical Example
Numerical Calculation	140	41.9%	"11 sticks, forming a triangle with every three sticks. How many triangles can be formed?"
Formula/Equation	100	29.9%	"There are 10 pears, 4 are taken away. How many remain in the basket? Write the calculation formula."
Multiple Choice and Judgement	50	15.0%	"Which row of animals has fewer in the upper or lower row?"
Open-Ended Response	30	9.0%	"Why is the quotient of this calculation closest to the middle number?"
Drawing/Design	14	4.2%	"Color 368 squares."

Table 19: **Model Overview**

Model Name	Parameters	Model Type	Source / Official Team	Reason for Selection
LLaVA-v1.5-7B	7B	Instruction-tuned Vision-Language Model	UC Berkeley et al.	Community benchmark model, stable performance, moderate resource usage
LLaVA-v1.5-13B	13B	Instruction-tuned Vision-Language Model	UC Berkeley et al.	Large-scale version, stronger reasoning, suitable for high-resource environments
MiniGPT-4(Vicuna-7B)	7B	BLIP-2 + Vicuna (Lightweight)	HKU	Lightweight deployment, early classic model, easy comparison
MiniGPT-4(Vicuna-13B)	13B	BLIP-2 + Vicuna (Large)	HKU	Stronger reasoning capability, suitable for comparison with 7B
InstructBLIP(Vicuna-7B)	7B	Instruction-tuned BLIP-2	Salesforce	Instruction-aligned version, tailored for educational tasks
InstructBLIP(Vicuna-13B)	13B	Instruction-tuned BLIP-2	Salesforce	Stronger reasoning, suitable for scale comparison
BLIP-2	3B	Base Vision-Language Model	Salesforce	Classic benchmark, source for most improved models
MobileVLM-3B	3B	Lightweight Vision-Language Model	Meituan-AutoML	Lightweight and efficient, suitable for local/mobile deployment
MobileVLM-V2-7B	7B	Lightweight Vision-Language Model	Meituan-AutoML	V2 version, optimized performance, balance of efficiency and effectiveness
mPLUG-Owl2	7B	General Vision-Language Model	Alibaba DAMO Academy	Emphasizes complex reasoning, good open-source performance
IDEFICS-9B-Instruct	9B	Multimodal Dialogue Model	HuggingFace	Instruction fine-tuned, ecosystem-friendly, easy integration
VisualGLM-6B	6B	Chinese-optimized Vision-Language Model	Tsinghua KEG & Zhipu	Advantage in Chinese scenarios, suitable for number-shape integration tasks
MultiModal-GPT	9B	General Vision-Language Model	Open-source community	General framework, easily extendable for educational tasks

Fig. 9: **Model Theory Ranking**

6.7. Model Accuracy in Each Domain

The performance of different models in different knowledge fields is shown in Fig. 20.

Table 20: **Experimental results of different models in different knowledge domains.**

model	ME	FO	Pro	Time	ENC	MIM	Sta	MCG	RG1	RG2	MG	GC	DO	RNRO	IO
LLaVA-v1.5-13B	26.92%	10.20%	28.57%	10.53%	12.50%	16.67%	14.71%	12.71%	30.87%	18.14%	16.92%	7.69%	15.22%	0%	21.56%
MobileVLM-V2-7B	15.38%	10.20%	14.29%	10.53%	9.03%	11.90%	5.88%	7.63%	21.55%	15.42%	9.23%	0%	6.52%	0%	13.17%
LLaVA-v1.5-7B	26.90%	4.08%	14.29%	0%	4.17%	9.52%	2.94%	6.78%	14.10%	15.42%	23.08%	11.54%	10.87%	0%	12.87%
MiniGPT-4(Vicuna-13B)	7.69%	3.06%	14.29%	0%	16.67%	11.90%	2.94%	1.69%	10.07%	9.98%	15.38%	15.38%	19.57%	16.67%	14.67%
MobileVLM-3B	19.20%	4.08%	33.33%	10.53%	6.94%	14.29%	2.94%	5.08%	22.15%	14.97%	23.08%	3.85%	0%	0%	4.50%
mPLUG-Owl2	15.38%	4.08%	4.76%	10.53%	3.47%	9.52%	2.94%	6.78%	16.78%	12.70%	4.62%	15.38%	8.70%	0%	12.87%
MultiModal-GPT	7.69%	3.06%	14.29%	10.53%	4.86%	19.05%	5.88%	5.93%	16.78%	14.74%	6.15%	7.69%	6.52%	0%	8.38%
MiniGPT-4(Vicuna-7B)	7.69%	15.31%	23.81%	10.53%	7.64%	11.90%	14.71%	4.24%	14.09%	8.62%	9.23%	3.85%	6.52%	0%	10.78%
BLIP-2	11.54%	4.08%	9.52%	0%	6.25%	9.52%	0%	1.69%	11.41%	9.52%	21.54%	15.38%	10.87%	0%	9.64%
InstructBlip(Vicuna-13B)	19.23%	5.10%	4.76%	10.53%	6.25%	4.76%	0%	4.24%	15.44%	8.62%	10.77%	7.69%	6.52%	0%	8.08%
VisualGLM-6B	0%	2.04%	9.52%	10.53%	6.94%	4.76%	8.82%	4.24%	22.82%	9.75%	4.62%	0%	2.17%	0%	5.39%
InstructBlip(Vicuna-7B)	11.54%	10.20%	23.81%	10.53%	4.86%	9.52%	5.88%	5.08%	18.12%	6.58%	1.54%	3.85%	4.35%	0%	7.49%
IDEFICS-9B-Instruct	15.38%	11.22%	23.81%	10.53%	3.47%	4.76%	8.82%	3.39%	8.05%	1.82%	9.23%	3.85%	2.17%	0%	9.94%

Measurement (ME), Fraction Operations (FO), Probability (Pro), Time, Expansion of Number Concepts (ENC), Mathematical Ideas and Methods (MIM), Statistics (Sta), Measurement and Calculation of Graphics (MCG), Relationships of Graphics (RG1), Recognition of Graphics (RG2), Movement of Graphics (MG), Graphics and Coordinates (GC), Decimal Operations (DO), Rational Numbers and Related Operations (RNRO), and Integer Operations (IO).

6.8. Comparison of Excellent and Poor Models

To intuitively demonstrate the performance characteristics of multimodal large models on number–shape integration tasks in primary and secondary school mathematics, this study selects five representative cases covering the following problem types: Integer Operations, Graphics Recognition, Decimal Operations, Graphics and Coordinates, and Rational Number Operations. These categories not only correspond to the knowledge domains emphasized in subsequent analyses but also collectively reflect the differences in model capabilities across diverse number–shape integration tasks. Each case consists of the problem image, reference answer, model output, and error type annotation. For comparative purposes, the model with the highest overall accuracy (LLaVA-v1.5-13B) and the model with the lowest overall accuracy (IDEFICS-9B-Instruct) are highlighted to illustrate the performance gap.

Case 1: Integer Operations

Table 21: **Comparison of LLaVA-v1.5-13B and IDEFICS-9B-Instruct on Integer Operations (Number-Line Problem)**

Category	**Content**
Question	If 3 cubes make a cuboid, how many cuboids can be made?
Reference Answer	The number of cuboids that can be formed is $15 \div 3 = 5$.
LLaVA-v1.5-13B	If 3 cubes make a cuboid, we can make a cuboid by connecting three cubes in a row. In the image, there are 15 cubes, so we can make 5 cuboids by connecting three cubes in a row.
IDEFICS-9B-Instruct	The answer is 27.
Error Type	Disconnection between numbers and shapes

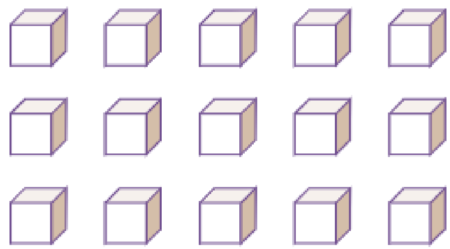

Fig. 10: **Integer Operations (Number-Line Problem).**

Error Analysis: The IDEFICS-9B-Instruct model directly provides an answer, which is incorrect, revealing its weakness in the "image-to-numerical conversion" step. The LLaVA-v1.5-13B model answers correctly, indicating that intuitive problems are still well-suited for some models.

Case 2: Recognition of Graphics

Table 22: **Comparison of LLaVA-v1.5-13B and IDEFICS-9B-Instruct on Graphics Recognition (Geometric Identification).**

Category	**Content**
Question	From which direction is the right image viewed relative to the left object?
Reference Answer	Viewed from above, the object appears as two connected rectangles, matching the right image.
LLaVA-v1.5-13B	The right image is viewed from above.
IDEFICS-9B-Instruct	The right image is viewed from the left.
Error Type	Spatial Relationship Error

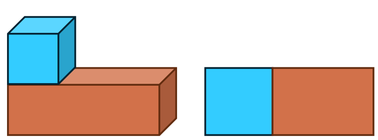

Fig. 11: **Recognition of Graphics (Geometric Identification).**

Error Analysis: In the geometric recognition (spatial perspective judgment) task, the model is required to determine the viewing direction of a three-dimensional object based on its projection. The main challenge lies in understanding spatial geometric relations. LLaVA-v1.5-13B produced the correct answer by identifying the perspective as "top view," whereas IDEFICS-9B-Instruct misclassified it as "left view," reflecting a typical error in spatial relationship perception.

Case 3: **Decimal Operations**

Table 23: **Comparison of LLaVA-v1.5-13B and IDEFICS-9B-Instruct on Decimal Operations**

Category	Content
Question	The left box of apples costs 58.2 yuan, the middle box of pears costs 39.6 yuan, and the right box of bananas costs 42.8 yuan. Can you buy the four boxes of fruit shown in the picture with 200 yuan?
Reference Answer	Yes, you can buy the four boxes of fruit shown in the picture with 200 yuan.
LLaVA-v1.5-13B	Yes, you can buy the four boxes of fruit shown in the picture with 200 yuan. The total cost of the four boxes is 58.2 + 39.6 + 42.8 + 42.8 = 193.4 yuan. Since the total cost is less than 200 yuan, you can afford to buy all four boxes.
IDEFICS-9B-Instruct	No, you cannot buy the four boxes of fruit shown in the picture with 200 yuan.
Error Type	Numerical Calculation Error

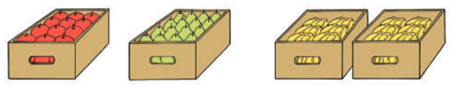

Fig. 12: **Decimal Operations(Graphics data processing).**

Error Analysis: The IDEFICS-9B-Instruct model lacks the reasoning and calculation process, directly giving a "cannot" answer, exposing deficiencies in its logic and computation; the chain from visual-to-numerical transformation to logical calculation is incomplete.

Case 4 : **Graphics and Coordinates**

Table 24: **Comparison of LLaVA-v1.5-13B and IDEFICS-9B-Instruct on Graphics and Coordinates**

Category	Content
Question	Observe the image. If the triangle is translated 2 units to the right and 3 units upward, will one point coordinate change to (1, 7)?
Reference Answer	Yes
LLaVA-v1.5-13B	Yes, if the triangle is translated 2 units to the right and 3 units upward, one point's coordinate will change to (1, 7).
IDEFICS-9B-Instruct	Yes, the point's coordinate will change to (1, 7) if the triangle is translated 2 units to the right and 3 units upward.
Error Type	No Error

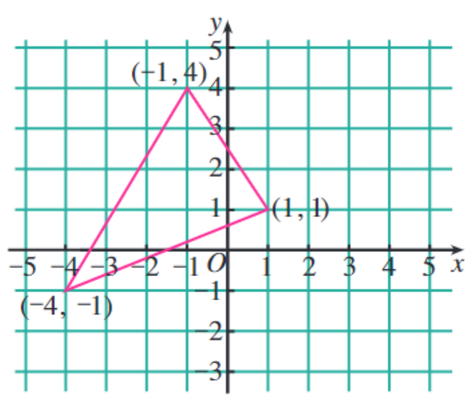

Fig. 13: **Graphics and Coordinates (Abstract Transformation).**

Error Analysis: Both LLaVA-v1.5-13B and IDEFICS-9B-Instruct provide similar analyses for the same problem, reflecting that their geometry-to-algebra transformation ability is relatively good.

Case 5 : **Rational numbers and related operations**

Table 25: **Comparison of LLaVA-v1.5-13B and IDEFICS-9B-Instruct on Rational Number Operations**

Category	Content
Question	The weights of 10 bags of wheat are recorded in the picture on the right, unit: kilograms. What is the total weight of the 10 bags? If each bag is supposed to weigh 50 kilograms, is the total weight of the 10 bags heavier or lighter than the standard?
Reference Answer	The total weight of the 10 bags of wheat is 502.5 kg, which is 2.5 kg heavier than the standard total of 10 bags.
LLaVA-v1.5-13B	The total weight of the 10 bags of wheat is 500 kilograms. If each bag is supposed to weigh 50 kilograms, the total weight of the 10 bags should be 500 kilograms. In this case, the total weight of the 10 bags is heavier than the standard.
IDEFICS-9B-Instruct	The total weight of the 10 bags is 500 kilograms. If each bag is supposed to weigh 50 kilograms, the total weight of the 10 bags is heavier than the standard.
Error Type	Ignored Key Constraint

Fig. 14: **Rational numbers and related operations (abstract logic).**

Error Analysis: Almost all models fail on rational number problems. Common errors include not using the corresponding quantities in the image (i.e., ignoring key constraints) and not fully understanding the problem, highlighting serious limitations of current multimodal models in multi-object quantity recognition and complex number-shape integration tasks.

6.9. Error Classification of All Models

The error conditions of all models are shown in Fig. 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27.

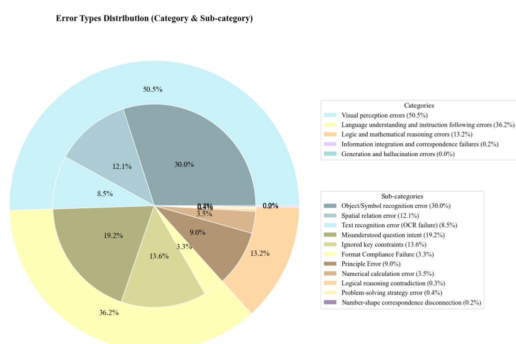

Fig. 15: **llava-13b**

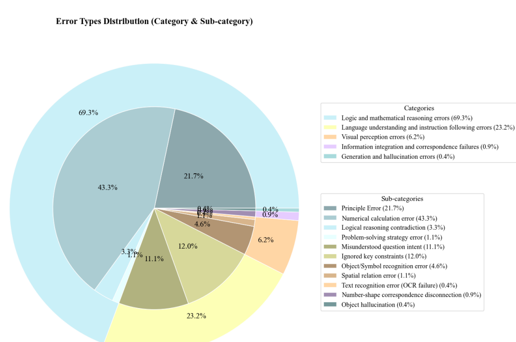

Fig. 16: **llava-7b**

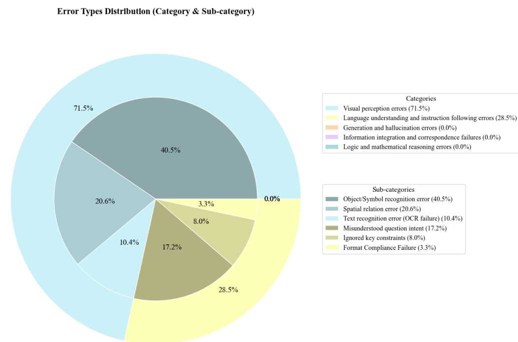

Fig. 17: **minigpt4-13b**

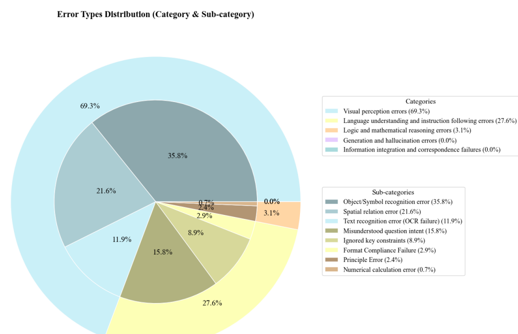

Fig. 18: **minigpt4-7b**

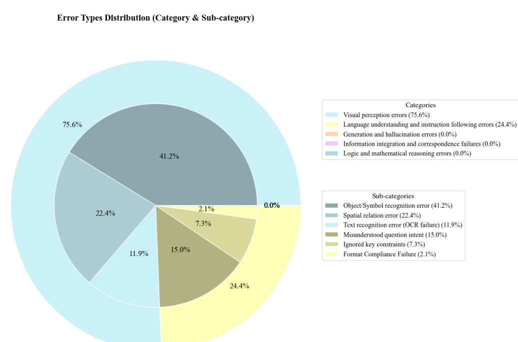

Fig. 19: **instructblip-13b**

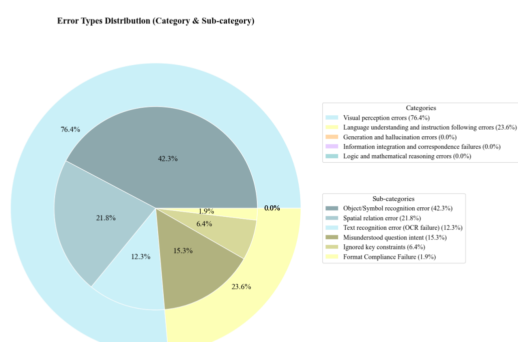

Fig. 20: **instructblip-7b**

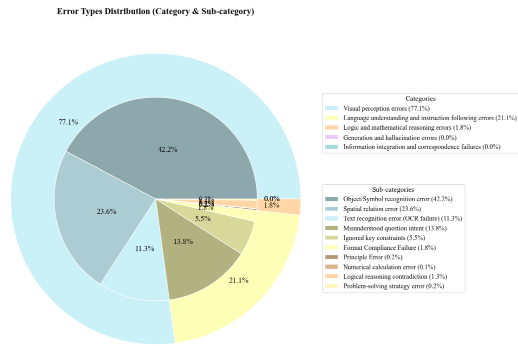

Fig. 21: bilp-2

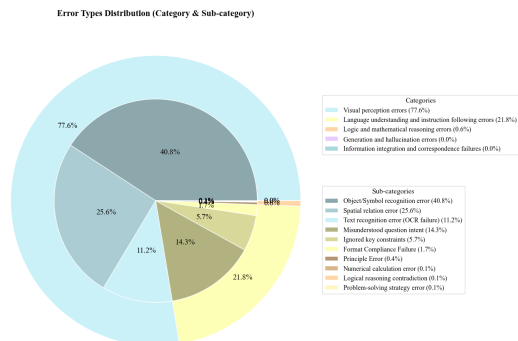

Fig. 22: mobilevlm-7b

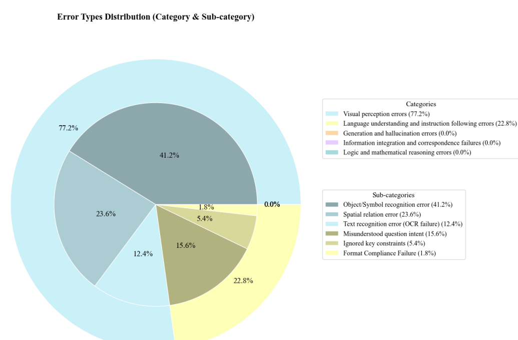

Fig. 23: mobilevlm-3b

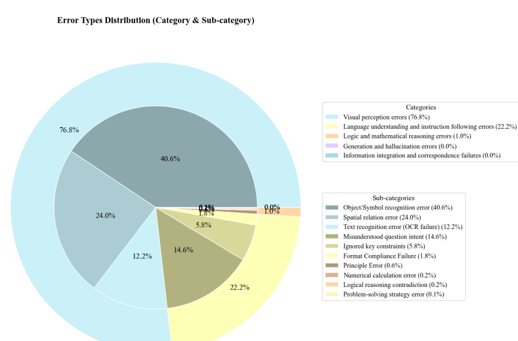

Fig. 24: mPLUG-Owl2

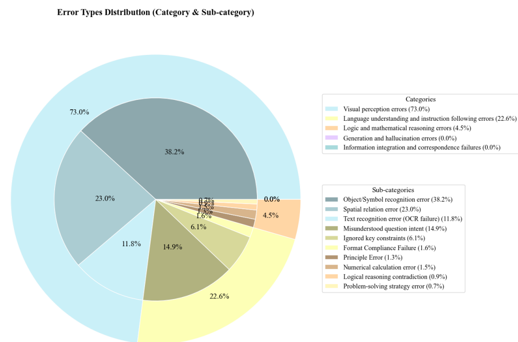

Fig. 25: **IDEFICS-9B-Instruct**

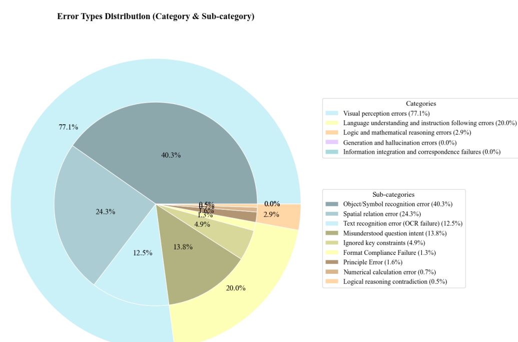

Fig. 26: **MultiModal-GPT**

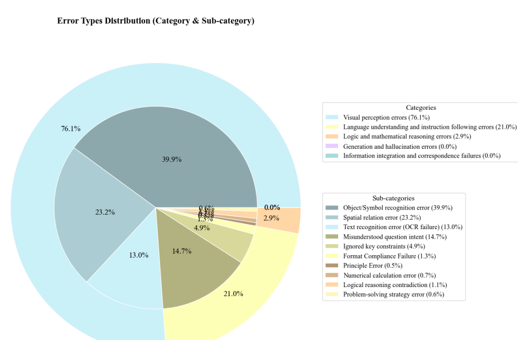

Fig. 27: **VisualGLM-6B**

6.10. Error Case of All Type

Visual Perception Errors. The first source of errors arises from difficulties in recognizing and interpreting fundamental visual elements. These errors include misidentification of objects or symbols (e.g., bars, points, lines, or numeric labels in charts), misjudgment of spatial relations such as relative positions or alignments, and failures in text recognition (OCR), where embedded labels, numbers, or scale values are incorrectly read.The following are examples:

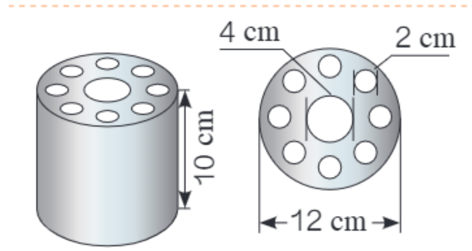

Fig. 28: **misidentification of objects or symbols**

question:A cylindrical metal part has 9 cylindrical holes. What is the approximate volume of metal used for this part in cubic decimeters?

model answer:To determine the approximate volume of metal used for this part in cubic decimeters, we need to know the dimensions of the cylinder and the size of the holes. Unfortunately, the image does not provide enough information to calculate the volume. However, we can use the measurements provided to determine the volume if we have the dimensions of the cylinder and the size of the holes.

reference:The metal usage of this part is approximately 0.8478 cubic decimeters.

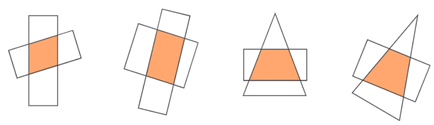

Fig. 29: **misjudgment of spatial relations**

question:What is the shape of the shaded area in each of the four figures in the image?

model answer:The shape of the shaded area in each of the four figures in the image is a triangle.

reference:A parallelogram has two pairs of parallel sides, meeting the definition of a parallelogram. A rectangle has four right angles and equal and parallel opposite sides, making it a rectangle. A trapezoid has only one pair of parallel opposite sides, making it a trapezoid. A triangle is formed by three line segments connected end to end.

Language and Instruction Understanding Errors. A second category concerns the model's ability to correctly parse and interpret textual input. Such errors manifest when the model misinterprets the intent of the question, disregards critical constraints (for instance, "list all combinations" or "explain reasoning"), or produces outputs that deviate from the required format, such as omitting units, lists, or structured responses.The following are examples:

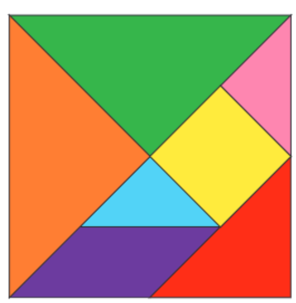

Fig. 30: **misinterprets the intent of the question**

question:What shapes are included in a tangram set, and how many of each shape are there?

model answer:A tangram set typically includes five different shapes: a square, a triangle, a parallelogram, a pentagon, and a hexagon. Each shape is made up of four smaller squares, which are called tiles. In the image, there are four different colored tiles, which are part of the tangram set. The tiles are arranged in a way that forms a colorful pattern, showcasing the versatility and creativity of the tangram set.

reference:The tangram consists of seven pieces, including 5 triangles (including 2 large triangles, 1 medium triangle and 2 small triangles), 1 square and 1 parallelogram.

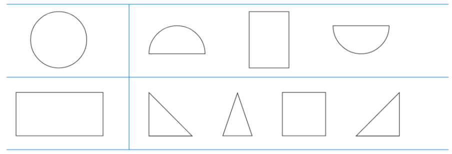

Fig. 31: **disregards critical constraints**

question:In each row, which shapes on the right can be combined to form the shape on the left?

model answer:In each row, the shapes on the right can be combined to form the shape on the left by following the pattern of the shapes. For example, in the first row, the triangle on the right can be combined with the square on the right to form the triangle on the left. In the second row, the square on the right can be combined with the triangle on the right to form the square on the left. In the third row, the triangle on the right can be combined with the square on the right to form the triangle on the left. In the fourth row, the square on the right can be combined with the triangle on

reference: First Row Analysis: The left side is a circle. The first shape on the right is a semicircle, and the third shape is also a semicircle. These two semicircles can be joined to form a circle, so the first and third shapes on the right can be used to form the left side.

Second Row Analysis: The left side is a rectangle. The first shape on the right is a right triangle, the third shape is a square, and the fourth shape is a right triangle. By joining the first and fourth right triangles to opposite sides of the third square (with the right-angled sides coinciding with the square sides), we can form a rectangle. Therefore, the first, third, and fourth shapes on the right can be used to form the left side.

Fig. 32: **disregards critical constraints**

question:How many apples are there in total in the image? Write a multiplication expression.

model answer:There are a total of 9 apples in the image.

reference:There are 9 apples in the picture, and the multiplication formula is 3×3 = 9.

Logical and Mathematical Reasoning Errors. Even when information is correctly extracted, models frequently fail during the reasoning stage. Errors of this type include applying incorrect formulas, theorems, or methods; committing numerical miscalculations within otherwise correct procedures; producing logically inconsistent reasoning; or adopting inefficient or inappropriate problem-solving strategies.The following are examples:

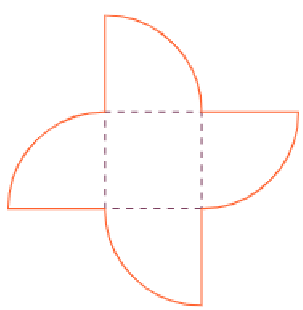

Fig. 33: **applying incorrect formulas, theorems, or methods**

question:A shape has a square with side length 1 cm in the center and four 90-degree sectors around it. What is the total area of the entire shape?

model answer:The total area of the entire shape is 16 square centimeters. This can be calculated by multiplying the side length of the square (1 cm) by the number of sectors (4) and then adding the area of the square to the sum.

reference:The area of the entire figure is 4.14 square centimeters.

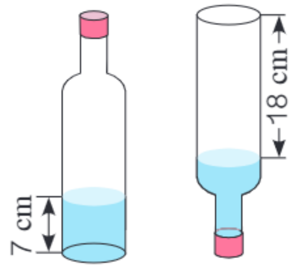

Fig. 34: **applying incorrect formulas, theorems, or methods**

question:A bottle with an internal bottom diameter of 8 cm contains water to a height of 7 cm. After sealing the cap and turning the bottle upside down, the air-filled part becomes a cylinder 18 cm tall. What is the total volume of the bottle?

model answer:To find the total volume of the bottle, we need to consider the volume of the water inside the bottle and the volume of the air inside the bottle. The volume of the water inside the bottle is 7 cm x 8 cm = 56 cm̆00b3. The volume of the air inside the bottle is 18 cm x 7 cm = 126 cm̆00b3. The total volume of the bottle is the sum of these two volumes: 56 cm̆00b3 + 126 cm̆00b3 = 182 cm̆00b3.

reference:The volume of this bottle is 1256 cubic centimeters

Information Integration and Correspondence Failures. Another major challenge lies in linking visual inputs with abstract mathematical concepts. Failures include the inability to map visual features (such as length or area) to corresponding numerical values, difficulty in synthesizing visual and textual information to construct a solution, or neglecting crucial contextual cues such as titles, legends, or axis labels. The following are examples:

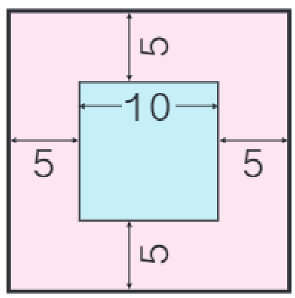

Fig. 35: **difficulty in synthesizing visual and textual information to construct a solution**

question:The side length of the small square in the image is 10. What is the perimeter of the large square?

model answer:The perimeter of the large square is 50, as it is a 5x5 square.

reference:The perimeter of the large square is 80.

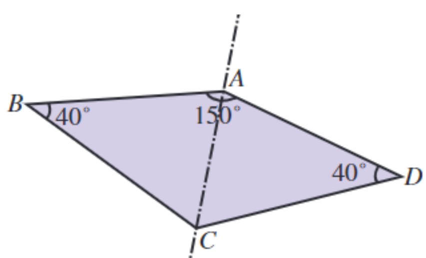

Fig. 36: **the inability to map visual features (such as length or area) to corresponding numerical values**

question:The image shows a symmetrical quadrilateral ABCD. Based on the image, what is the measure of angle C?

model answer:Unfortunately, I cannot provide an accurate answer to your question without more information or a clearer image. The image description does not provide enough details to determine the measure of angle C.

reference:Angle C is 130 degrees.

Generation and Hallucination Errors. Finally, some errors are attributable to the generation process itself, where the model introduces information not grounded in the input. This includes fabricating numbers or measurements (numerical hallucination), describing objects or relations absent from the figure (object hallucination), or introducing incorrect domain knowledge and commonsense facts (factual hallucination).The following are examples:

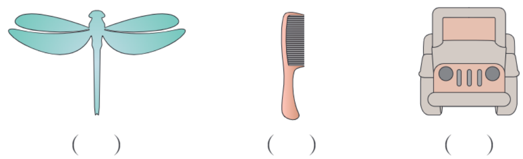

Fig. 37: **describing objects or relations absent from the figure (object hallucination)**

question:Which are axisymmetric shapes? Indicate in the brackets.

model answer:[Dragonfly, Car, Jeep].

reference:A dragonfly shape: When folded along the vertical line, the two sides of the straight line completely overlap, making it an axially symmetrical shape. A comb shape: No matter which line it is folded along, the two sides of the straight line never completely overlap, making it not an axially symmetrical shape. A car shape: When folded along the vertical line, the two sides of the straight line completely overlap, making it an axially symmetrical shape.

Fig. 38: **describing objects or relations absent from the figure (object hallucination)**

question:Choose the correct object from the brackets based on the pattern in the image and fill in the blank.

model answer:The correct object from the brackets is a ̈doughnutäs it is the only one that has a hole in the middle.

reference:The third object in the brackets.