Appendix

A Question-Type Breakdown

We disaggregate diagnostic gaps across visual content types and cognitive levels, averaging across all open-source models.

Visual content type. Molecular structure questions exhibit the largest Integration Gap (IG= +9.4) and Perception Fidelity gap ($O_H - O_M = 8.2$), confirming bond-line structures as the hardest visual content for current VLMs to both perceive and integrate. Reaction mechanism diagrams follow closely (IG= +8.7), as models struggle to track multi-step transformations across arrow-connected intermediates. Among Biology content types, phylogenetic trees show the largest IG (+7.1), likely because tree topology requires relational reasoning over branching structures. Anatomical illustrations show the smallest IG (+4.2), suggesting that spatially labeled diagrams are the most integration-friendly format. Cell and organelle diagrams fall in between (IG= +5.8), and ecological diagrams show moderate gaps (IG= +5.3).

Cognitive level. Recall-level questions show the smallest IG (+5.1), consistent with simple extraction tasks being easier to perform end-to-end. Application-level (IG= +7.6) and Analysis-level (IG= +9.8) questions show progressively larger gaps, indicating that integration failure worsens as the reasoning chain lengthens and visual features are progressively diluted by language-model priors.

B Control Experiment Details

Diagnostic Logic. We define four possible outcomes for the Residual Integration Gap (R-IG = $\text{Acc}(O_M) - \text{Acc}(\text{CoT})$): (1) R-IG$\approx$ 0: CoT closes the gap, bottleneck is reasoning depth; (2) 0 < R-IG < IG: both reasoning depth and integration contribute; (3) R-IG$\approx$IG: bottleneck is integration, not reasoning depth; (4) 2P-Img$\approx O_M \gg$V+T: advantage is structural (task decomposition), not modality-specific.

Native thinking mode analysis. For the Qwen2.5-VL family with native extended thinking, Think mode adds +1.4 pp (72B: 65.9% $\rightarrow$ 67.3%) and +1.8 pp (7B: 51.3% $\rightarrow$ 53.1%) beyond CoT. However, R-IG against Think remains positive: Qwen2.5-VL-72B at +2.1 pp, 7B at +4.1 pp. Maximal reasoning affordances within V+T cannot substitute for the modality-switching that the Model Oracle performs.

Closed-source circularity concern. Closed-source models may perform implicit chain-of-thought or internal verbalization, making their small IG potentially circular. The CoT and Think controls partially address this: open-source models *with* CoT still show substantial R-IG while closed-source do not, confirming the behavioral distinction even after controlling for reasoning depth. We interpret the comparison as behavioral rather than a definitive architectural claim—closed-source models may have bridged the gap through better projection layers, implicit multi-pass reasoning, or training-time exposure to verbalize-then-reason data.

C Prompt Templates

All prompts are deterministic (temperature = 0) with identical system instructions within each mode across all models. No model-specific prompt engineering was applied.

C.1 Mode 1: Vision+Text (V+T)

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

Your task is to answer the following question
based on the provided image and question text.

Rules:
- If the question is multiple choice, respond
  with ONLY the option letter (e.g., "A" or "B").
  Do not include any explanation.
- If the question requires a numerical answer,
  respond with ONLY the number. Include units
  only if explicitly asked.
- If the question requires a short text answer,
  respond in at most one sentence.
- Do not hedge, qualify, or say "I think".
  Commit to a single answer.

[User]
<image>
{image_data}
</image>

Question: {question_text}
Options (if applicable): {options}

Your answer:
```

C.2 Mode 2: Text-Only (T)

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

Your task is to answer the following question
based ONLY on the question text provided.
No image is available.

Rules:
- If the question is multiple choice, respond
  with ONLY the option letter (e.g., "A" or "B").
  Do not include any explanation.
- If the question requires a numerical answer,
  respond with ONLY the number. Include units
  only if explicitly asked.
- If the question requires a short text answer,
  respond in at most one sentence.
```

```
- Do not hedge, qualify, or say "I think".
  Commit to a single answer.
- If the question references a diagram, figure,
  or image that you cannot see, use your best
  scientific judgment based on the textual
  information available.

[User]
Question: {question_text}
Options (if applicable): {options}

Your answer:
```

C.3 Mode 3: Vision-Only (V)

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

You will receive a single image that contains
BOTH a scientific diagram AND the question text
rendered within the image. No separate text
input is provided.

Your task:
1. Read the question text embedded in the image.
2. Examine the scientific content in the image.
3. Answer the question.

Rules:
- If the question is multiple choice, respond
  with ONLY the option letter (e.g., "A" or "B").
  Do not include any explanation.
- If the question requires a numerical answer,
  respond with ONLY the number.
- Do not reproduce or restate the question.
  Only provide the answer.

[User]
<image>
{composite_image_with_embedded_question}
</image>

Your answer:
```

C.4 Mode 4: Human Oracle (O_H)

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

Your task is to answer the following question.
You are provided with:
(a) The original scientific image, AND
(b) A detailed expert annotation that describes
    all visual content in the image, including
    labeled structures, numerical values, spatial
    relationships, colors, and symbolic notation.

Use BOTH the image and the annotation to answer
the question. The annotation resolves any
perceptual ambiguity in the image.

Rules:
- If the question is multiple choice, respond
  with ONLY the option letter (e.g., "A" or "B").
  Do not include any explanation.
- If the question requires a numerical answer,
  respond with ONLY the number.
- Do not hedge, qualify, or say "I think".
  Commit to a single answer.

[User]
<image>
{annotated_image}
</image>

Expert annotation:
{human_annotation_text}

Question: {question_text}
Options (if applicable): {options}

Your answer:
```

C.5 Mode 5: Model Oracle (O_M) — Pass 1 (Perception)

```
[System]
You are a scientific image analyst with
expertise in Chemistry and Biology diagrams.

Your task is to produce a DETAILED, STRUCTURED
DESCRIPTION of all visual content in the
provided image. A student will later use your
description (without seeing the image) to
answer a science question.

CRITICAL RULES:
- Do NOT answer the question. Only describe
  what you observe in the image.
- Do NOT speculate about what the answer might
  be or provide any reasoning toward an answer.
- Be exhaustive: the student will have NO access
  to the image and must rely entirely on your
  description.

Describe the following in order:

1. OVERALL LAYOUT
   - What type of scientific diagram is this?
   - How many distinct visual components are
     present? How are they arranged spatially?

2. STRUCTURES AND SHAPES
```

```
   - For Chemistry: identify all atoms, bonds,
     ring systems, functional groups, stereo-
     chemistry indicators, and charge symbols.
   - For Biology: identify all organelles, tissue
     types, organs, organisms, or structural
     components visible.

3. TEXT AND LABELS
   - Transcribe ALL text visible in the image.
   - Note the position of each label relative to
     the structure it annotates.

4. ARROWS, LINES, AND FLOW
   - Describe all arrows with their direction,
     start point, and end point.

5. COLORS AND VISUAL ENCODING
   - Note any color coding, shading, hatching,
     or highlighting.

6. NUMERICAL AND QUANTITATIVE DATA
   - Transcribe all numerical values, units,
     measurements, angles, or coordinates.

7. SPATIAL RELATIONSHIPS
   - Describe relative positions between key
     components.

[User]
<image>
{image_data}
</image>

A student needs to answer the following question
about this image (but you must NOT answer it
yourself --- only describe what you see):

"{question_text}"

Provide your structured description now:
```

C.6 Mode 5: Model Oracle (O_M) — Pass 2 (Reasoning)

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

Your task is to answer the following question
using ONLY the image description provided below.
You do NOT have access to the original image.

The description was written by a scientific
image analyst who examined the original image.
Treat the description as your sole source of
visual information.

Rules:
- If the question is multiple choice, respond
  with ONLY the option letter (e.g., "A" or "B").
  Do not include any explanation.
- If the question requires a numerical answer,
  respond with ONLY the number.
- Base your answer strictly on the information
  in the description. If the description does
  not contain sufficient information to answer
  confidently, select the most likely answer
  given the available information.
- Do not hedge, qualify, or say "I think".
  Commit to a single answer.

[User]
An image was described by a scientific analyst
as follows:

--- BEGIN DESCRIPTION ---
{model_generated_description_from_pass1}
--- END DESCRIPTION ---

Question: {question_text}
Options (if applicable): {options}

Your answer:
```

C.7 Control: V+T with Chain-of-Thought (V+T-CoT)

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

Your task is to answer the following question
based on the provided image and question text.

Think step by step before answering:
1. First, carefully examine the image and
   identify all relevant visual information.
2. Then, reason through the problem using
   the visual information and your scientific
   knowledge.
3. Finally, provide your answer.

Rules:
- Show your reasoning step by step.
- After your reasoning, write "FINAL ANSWER:"
  followed by ONLY the option letter (e.g., "A")
  or the numerical answer.
- Do not hedge or say "I think". Commit to a
  single answer.

[User]
<image>
{image_data}
</image>
```

```
Question: {question_text}
Options (if applicable): {options}

Think step by step, then provide your final
answer:
```

C.8 Control: Two-Pass with Image (2P-Img) — Pass 2

Pass 1 is identical to the Model Oracle Pass 1 (§C.5). Pass 2 differs by including the original image alongside the description:

```
[System]
You are an expert science examiner evaluating
a student's knowledge of Chemistry and Biology.

Your task is to answer the following question.
You are provided with:
(a) The original scientific image, AND
(b) A detailed description of the image written
    by a scientific analyst.

Use BOTH the image and the description to answer
the question.

Rules:
- If the question is multiple choice, respond
  with ONLY the option letter (e.g., "A" or "B").
  Do not include any explanation.
- If the question requires a numerical answer,
  respond with ONLY the number.
- Do not hedge, qualify, or say "I think".
  Commit to a single answer.

[User]
<image>
{image_data}
</image>

An image analyst described this image as follows:

--- BEGIN DESCRIPTION ---
{model_generated_description_from_pass1}
--- END DESCRIPTION ---

Question: {question_text}
Options (if applicable): {options}

Your answer:
```

C.9 Answer Extraction and Normalization

For all modes:

(1) **Option letter extraction:** For multiple-choice questions, extract the first occurrence of a single capital letter (A–E) via `\b([A-E])\b`. No match ⇒ invalid.
(2) **Numerical normalization:** Strip units, whitespace, commas; convert fractions and scientific notation to decimal; round to four significant figures.
(3) **Invalid response handling:** Refusals, contradictory answers, or no extractable answer scored as incorrect.

For V+T-CoT, extract from text following "FINAL ANSWER:"; if absent, fall back to the standard regex pipeline.

D Dataset Samples: Five Modes Applied to One Question

To concretely illustrate how DISSECT transforms a single question into five diagnostic inputs, we present one Biology and one Chemistry example, each shown across all five evaluation modes. This demonstrates how the same underlying question isolates different failure dimensions depending on the input construction.

D.1 Biology Example: Human Female Reproductive System and Ovulation

Original question: *Study the diagram given below.* The image shows a labeled diagram of the human female reproductive system (with structures A, B, and C marked) alongside an ovarian follicle development cycle.
Sub-questions: (a) What is the hormone responsible for ovulation? (b) What happens to part B if fertilization does not occur? (c) Describe the role of the corpus luteum.

D.1.1 Mode 1: Vision+Text (V+T). **Input:** Diagram only (Figure 1) + question text as separate channels.

> Study the diagram given below.
> (a) What is the hormone responsible for ovulation?
> (b) What happens to part B if fertilization does not occur?
> (c) Describe the role of the corpus luteum.

What this mode tests. Standard multimodal baseline. The model must visually identify the labeled structures (A, B, C) in the reproductive system diagram, interpret the ovarian follicle development cycle showing the progression from primordial follicle through ovulation to corpus luteum, and apply reproductive biology knowledge to answer all three sub-questions. The diagram is essential for part (b), since the model must identify that "part B" refers to the uterus. All other modes are compared against this accuracy.

D.1.2 Mode 2: Text-Only (T). **Input:** Question text only. No image.

> *(Same question text as Mode 1 above.)*

What this mode tests. Without the diagram, the model cannot determine what structures A, B, and C refer to. Sub-question (a) ("hormone responsible for ovulation") is answerable from parametric knowledge alone—the answer is luteinizing hormone (LH)—exposing low visual dependency. However, sub-question (b) is *unanswerable*: "What happens to part B?" is meaningless without the image identifying B as the uterus. Sub-question (c) is also answerable from textbook knowledge. This question thus has *mixed* visual dependency across its sub-parts: (a) low, (b) high, (c) low. The LPG captures this at the question level, but per-sub-question analysis reveals finer-grained patterns.

D.1.3 Mode 3: Vision-Only (V). **Input:** Single image (Figure 2) only. No separate text channel.

Table 1: Biology sample: input construction across all five modes.

Mode	Name	Image Input	Text Input
1	V+T	Diagram only (Fig. 1)	Question text
2	T	None	Question text only
3	V	Full image with questions (Fig. 2)	None
4	O_H	Diagram only (Fig. 1)	Human annotation + question text
5	O_M	Diagram only (Fig. 1)	Pass 2: model description + question

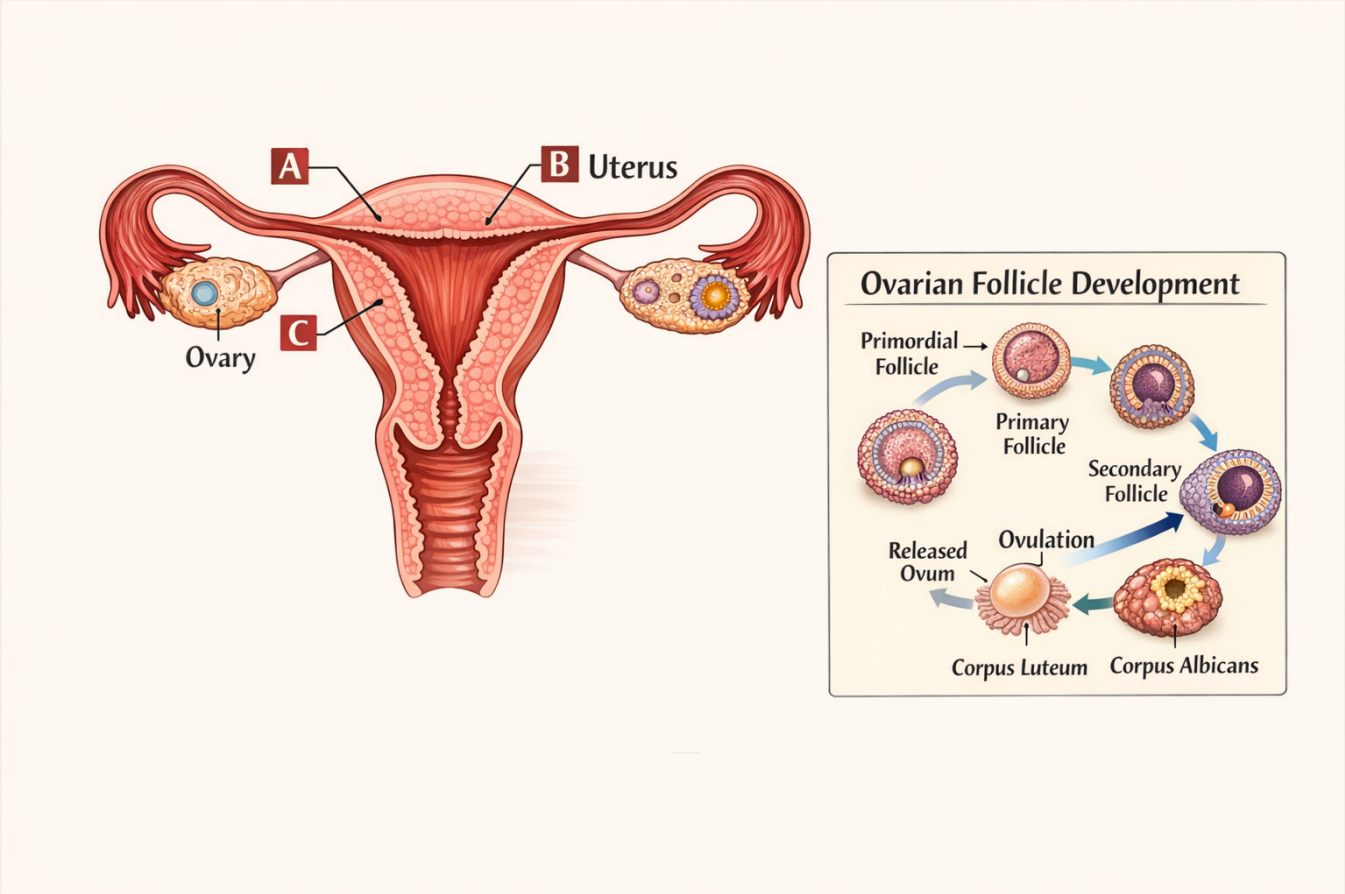

Figure 1: Biology sample, Modes 1/4/5 — Diagram of the human female reproductive system with labeled structures (A = Fallopian tube / Oviduct, B = Uterus, C = Ovary) and the ovarian follicle development cycle (primordial follicle → primary → secondary → ovulation → corpus luteum → corpus albicans). Question text is provided as a *separate* text channel.

What this mode tests. The model must: (1) OCR all three sub-questions from the image, including the italicized formatting and label references ("part B"); (2) distinguish the rendered question text at the bottom from the diagram's own labels ("Ovary," "Uterus," "Primordial Follicle," "Corpus Luteum," etc.); and (3) connect the label "B" in the question to its position in the anatomical diagram. Biological diagrams are particularly challenging for Mode 3 because they contain dense label text that overlaps spatially with rendered question text. The model must parse which text is a diagram annotation and which is the question to be answered.

D.1.4 Mode 4: Human Oracle (O_H). **Input:** Diagram only (Figure 1) + human annotation text + question text, all as separate channels.

Human annotation text provided:

> *Structure identification:* The diagram shows a frontal view of the human female reproductive system. Structure A = Fallopian tube (oviduct), connecting ovary to uterus. Structure B = Uterus (womb), a pear-shaped muscular organ with a thick endometrial lining. Structure C = Ovary, the primary female reproductive organ, shown with a small follicle on its surface.

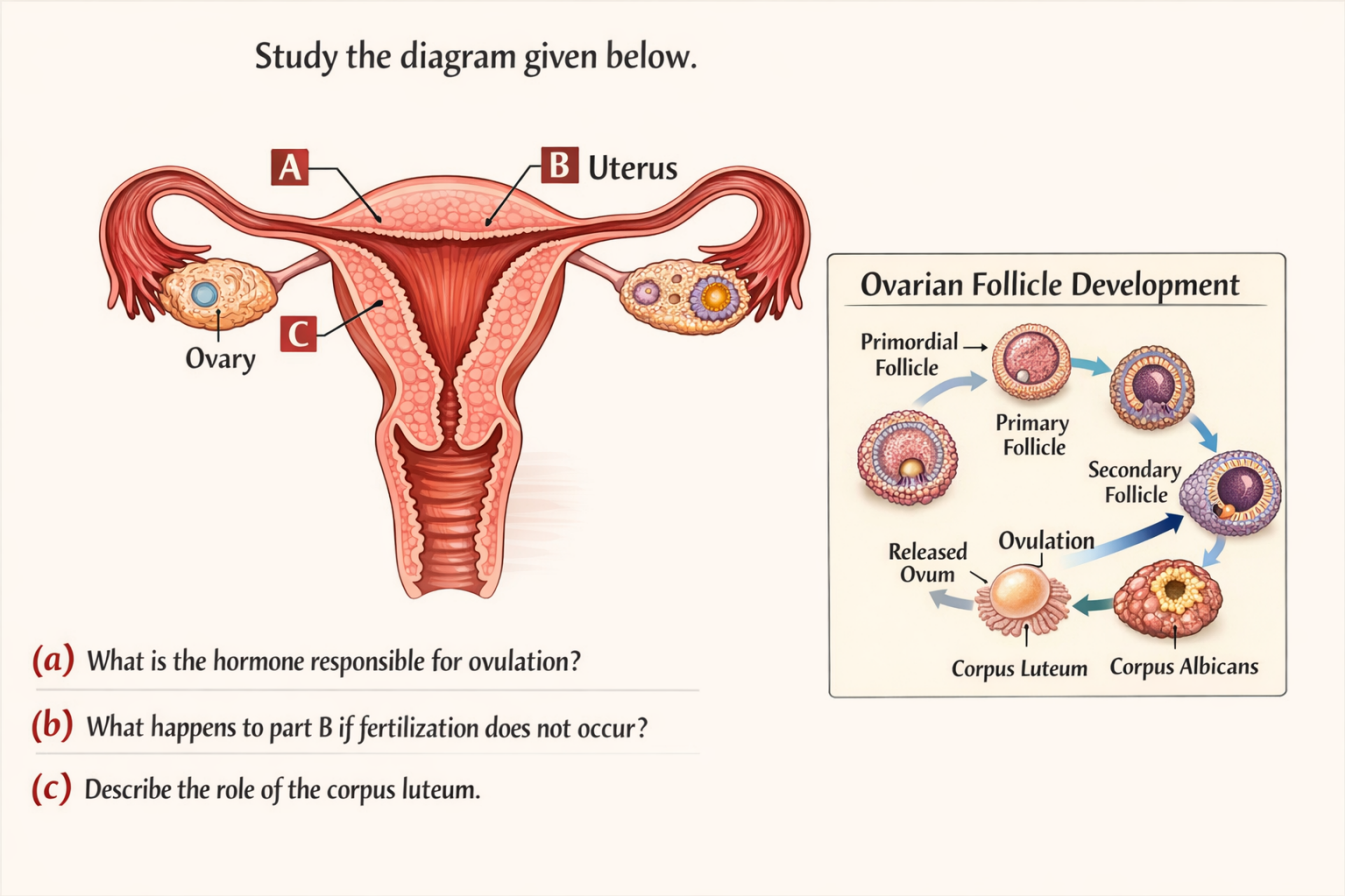

Figure 2: Biology sample, Mode 3 (V) — The complete image including the diagram *and* the three sub-questions (a), (b), (c) rendered within the image. This single image is the *only* input; no separate text channel is provided. The model must OCR the questions from the image while simultaneously interpreting the anatomical diagram and follicle development cycle.

> *Ovarian follicle development cycle (right panel):* Circular diagram showing: Primordial follicle → Primary follicle → Secondary follicle → Ovulation (release of ovum) → Corpus luteum (yellow body, secretes progesterone) → Corpus albicans (degenerated corpus luteum). Blue arrows indicate the progression sequence clockwise. The "Released Ovum" is shown departing from the secondary follicle stage.
>
> *Spatial relationships:* Ovary (C) is lateral, connected to uterus (B) via Fallopian tube (A). The released ovum travels from ovary through the Fallopian tube to the uterus. The follicle development cycle occurs *within* the ovary.
>
> *Relevant physiological context:* LH surge triggers ovulation. Corpus luteum secretes progesterone to maintain uterine lining. If no fertilization, corpus luteum degenerates into corpus albicans, progesterone drops, and the endometrial lining of the uterus sheds (menstruation).

What this mode tests. The annotations resolve all perceptual ambiguity: the model no longer needs to visually identify what A, B, and C are, or interpret the follicle development cycle diagram. In particular, the critical mapping B = Uterus is stated explicitly, making sub-question (b) answerable. The only remaining task is biological reasoning—linking the absence of fertilization to corpus luteum degeneration, progesterone withdrawal, and endometrial shedding. If the model fails here, the bottleneck is reasoning or knowledge. If it succeeds here but fails under V+T, the bottleneck is definitively perceptual (it could not identify the labeled structures from the diagram).

D.1.5 Mode 5: Model Oracle (O_M). **Input:** Two-pass procedure using the same VLM.
Pass 1: Diagram only (Figure 1) + question text.
Pass 2: Model's self-generated description + question text (no image).

Pass 1 (Perception) — Expected output:

> *The image contains two components. On the left is a frontal anatomical diagram of the human female reproductive system. Three structures are labeled with red boxes: A points to a curved tubular structure connecting the ovary to the uterus (Fallopian tube/oviduct), B points to the large central pear-shaped organ (uterus), and C points to the smaller oval organ on the left side (ovary), which has a small circular*

structure on its surface. The label "Ovary" appears near C, and "Uterus" appears near B.
On the right is a circular diagram titled "Ovarian Follicle Development." It shows a clockwise progression: Primordial Follicle → Primary Follicle → Secondary Follicle → Ovulation (with a "Released Ovum" departing) → Corpus Luteum → Corpus Albicans, which cycles back to the primordial stage. Blue arrows indicate the direction of progression.

Pass 2 (Reasoning): Receives *only* the above description and the original question text (no image).

What this mode tests. The model itself extracts visual information in Pass 1. A critical test is whether the model correctly maps the labels: does it identify B as the uterus? If Pass 1 describes "B points to the large central organ" but fails to name it, Pass 2 cannot answer sub-question (b). Comparing results: O_M > V+T means the model perceives the labels correctly but struggles with joint reasoning in a single pass; $O_M < O_H$ means the model's self-generated description is less complete than human annotations—e.g., it might describe the follicle cycle without noting that the corpus luteum secretes progesterone, or it might miss the corpus albicans stage entirely; $O_M \approx O_H$ means the model's perceptual extraction for this anatomical diagram matches human quality. A common Pass 1 failure for this type of labeled biological diagram is correctly transcribing the visible labels ("Ovary," "Uterus") but failing to associate them with the letter labels (A, B, C) in the red boxes.

D.2 Chemistry Example: Electrolysis of $SnSO_4$ (Chem Q30)

Original question: *If 0.50 L of a 0.60 M $SnSO_4$ solution is electrolyzed for 30.0 min using a current of 4.60 A with inert electrodes, what is the final concentration of Sn^{2+} remaining in the solution?* [at. wt. of Sn = 119]

(1) 0.342 M
(2) 0.544 M
(3) 0.389 M
(4) 0.514 M

Correct answer: (d) 0.514 M

D.2.1 Mode 1: Vision+Text (V+T). **Input:** Image (Figure 3) + question text as separate channels.

> If 0.50 L of a 0.60 M $SnSO_4$ solution is electrolyzed for a period of 30.0 min using a current of 4.60 A. If inert electrodes are used, what is the final concentration of Sn^{2+} remaining in the solution? [at. wt. of Sn = 119]
> (1) 0.342 M
> (2) 0.544 M
> (3) 0.389 M
> (4) 0.514 M

What this mode tests. The model must read the cell diagram to identify electrode types (inert), the cathode reaction ($Sn^{2+} + 2e^- \rightarrow$ Sn), ion migration direction, and all quantitative parameters, then apply Faraday's law of electrolysis. The diagram provides visual confirmation of the setup that complements the numerical data in the question text.

D.2.2 Mode 2: Text-Only (T). **Input:** Question text only. No image.

(Same question text and options as Mode 1 above.)

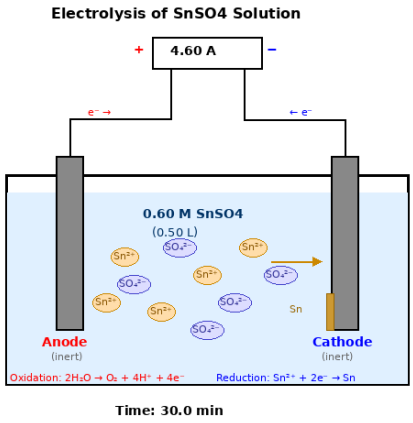

Figure 3: Chem Q30 — Electrolysis cell diagram showing 0.60 M $SnSO_4$ solution (0.50 L), inert anode and cathode, Sn^{2+} and SO_4^{2-} ions, electrode reactions, Sn deposit on cathode, and 4.60 A current source. This image is used in Modes 1 (V+T), 4 (O_H), and 5 (O_M Pass 1).

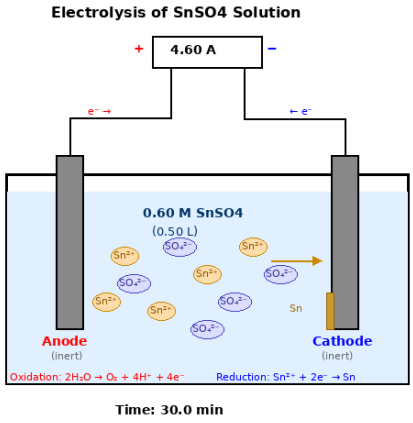

Figure 4: Chem Q30 — Mode 3 (V) composite image: the electrolysis cell diagram with the complete question text and all four options rendered below. This single image is the *only* input; no separate text channel is provided.

What this mode tests. This question is *largely answerable* from text alone: all numerical parameters (molarity, volume, current, time, atomic weight) are in the question text. The calculation follows

Table 2: Chem Q30 input construction across all five modes.

Mode	Name	Image Input	Text Input
1	V+T	Cell diagram (Fig. 3)	Question text
2	T	None	Question text only
3	V	Composite image (Fig. 4)	None
4	O_H	Cell diagram (Fig. 3)	Human annotation + question text
5	O_M	Cell diagram (Fig. 3)	Pass 2: model description + question

directly from Faraday's law:

$$Q = I \times t = 4.60 \times 1800 = 8280 \text{ C}$$

$$n_{e^-} = \frac{8280}{96500} \approx 0.0858 \text{ mol}$$

$$n_{\text{Sn}^{2+}\text{reduced}} = \frac{0.0858}{2} = 0.0429 \text{ mol}$$

$$[\text{Sn}^{2+}]_{\text{final}} = \frac{0.30 - 0.0429}{0.50} \approx 0.514 \text{ M}$$

A high Text-Only accuracy here exposes that the diagram adds minimal information beyond what the text provides—a classic case of low visual dependency. The LPG for this question is expected to be ≈ 0, flagging it as a question where V+T accuracy overestimates visual grounding.

D.2.3 Mode 3: Vision-Only (V). **Input:** Single composite image (Figure 4) only. No separate text channel.

What this mode tests. Chemistry questions pose unique OCR challenges: the model must parse subscripts ($SnSO_4$), superscripts with charges (Sn^{2+}), decimal values (0.60 M, 4.60 A), and units (mol, L, min) from rendered text. It must also distinguish the rendered question text from the diagram's own labels—both contain "Sn^{2+}", "0.60 M", and "4.60 A". Numerical OCR errors (misreading "0.60" as "0.80") propagate directly into incorrect Faraday's law calculations, making this mode especially sensitive to text extraction fidelity in quantitative chemistry.

D.2.4 Mode 4: Human Oracle (O_H). **Input:** Original image (Figure 3) + human annotation text + question text, all as separate channels.

Human annotation text provided:

> *Apparatus identification:* Electrolytic cell with two inert (non-reactive) electrodes. Power source supplies 4.60 A direct current. Left electrode = Anode (oxidation occurs). Right electrode = Cathode (reduction occurs). Solution: 0.50 L of 0.60 M $SnSO_4$.
>
> *Electrode reactions:* Cathode: $Sn^{2+} + 2e^- \rightarrow Sn(s)$. n-factor = 2 (two electrons per Sn^{2+} ion reduced). Anode: $2H_2O \rightarrow O_2 + 4H^+ + 4e^-$ (water is oxidized; O_2 gas evolved).
>
> *Quantitative data extracted from diagram:* Current = 4.60 A. Time = 30.0 min. Volume = 0.50 L. Initial concentration $[Sn^{2+}]$ = 0.60 M. Initial moles of Sn^{2+} = 0.60×0.50 = 0.30 mol.
>
> *Ion movement:* Sn^{2+} cations migrate toward cathode (arrow visible in diagram moving rightward). SO_4^{2-} anions migrate toward anode. Electrons flow from anode to cathode through external circuit.
>
> *Visual indicators:* Golden/brown deposit on cathode surface = solid Sn metal being deposited. Orange ovals = Sn^{2+} ions in solution. Purple ovals = SO_4^{2-} ions. Blue shading = aqueous $SnSO_4$ solution.

What this mode tests. The annotations explicitly provide the cathode reaction with its n-factor, pre-computed initial moles, and all numerical values extracted from the diagram. The original image is still provided for reference. The only remaining task is applying Faraday's law arithmetic. If the model fails here, the bottleneck is quantitative reasoning (stoichiometry, unit conversion), not perception. Failure would indicate a fundamental gap in electrochemistry knowledge.

D.2.5 Mode 5: Model Oracle (O_M). **Input:** Two-pass procedure using the same VLM.
Pass 1: Original image (Figure 3) + question text.
Pass 2: Model's self-generated description + question text (no image).

Pass 1 (Perception) — Expected output:

> *The image shows an electrolysis setup. A power source labeled "4.60 A" is connected to two grey electrodes immersed in a blue solution. The left electrode is labeled "Anode (inert)" with the reaction "Oxidation: $2H_2O \rightarrow O_2 + 4H^+ + 4e^-$". The right electrode is labeled "Cathode (inert)" with "Reduction: $Sn^{2+} + 2e^- \rightarrow Sn$". The solution is labeled "0.60 M $SnSO_4$ (0.50 L)". Orange ovals represent Sn^{2+} ions and purple ovals represent SO_4^{2-} ions. An arrow shows Sn^{2+} migrating toward the cathode. A golden deposit is visible on the cathode surface. Time is given as 30.0 min.*

Pass 2 (Reasoning): Receives *only* the above description and the original question text (no image).

What this mode tests. For chemistry, Pass 1 must extract precise numerical values and chemical notation from the diagram. A common failure is misreading concentrations or charges—e.g., describing "Sn^{2+}" as "Sn^+" would change the n-factor from 2 to 1, doubling the moles reduced and yielding an incorrect final concentration of 0.428 M instead of 0.514 M. The Perception Fidelity gap ($O_H - O_M$) captures exactly these numerical extraction errors, which are particularly consequential in quantitative chemistry where small perceptual mistakes cascade into large calculation errors.

D.3 Cross-Subject Comparison

Table 3 summarizes the diagnostic signals across both subjects. The cross-subject comparison reveals a key difference in how perceptual failures manifest. In Biology, perception failures tend to

Table 3: Diagnostic signals for each mode across Biology and Chemistry examples. The same five-mode framework reveals different failure patterns depending on the subject and question type.

Mode	Biology (Reproductive System)	Chemistry (Chem Q30: Electrolysis)
1 (V+T)	Requires identifying labeled structures (A, B, C) + interpreting follicle development cycle + reproductive biology reasoning	Requires reading cell setup + Faraday's law calculation
2 (T)	Mixed dependency: (a) and (c) answerable from knowledge; (b) unanswerable—"part B" is meaningless without the diagram	Largely answerable: all numerical data is in the text
3 (V)	Dense diagram labels ("Ovary," "Uterus," "Corpus Luteum") overlap with question text; model must parse which is which	Numerical OCR errors (misreading 0.60 as 0.80) cascade into incorrect calculations
4 (O_H)	Annotations map B = Uterus, resolving the critical label; remaining task is physiological reasoning about menstruation	Annotations provide n-factor and pre-computed moles; remaining task is arithmetic
5 (O_M)	Perceptual risk: correctly reading labels "Ovary" and "Uterus" but failing to associate them with letter labels A, B, C	Perceptual risk: misreading charges or concentrations; small errors cause large calculation errors

be *categorical*—misidentifying a structure leads to a qualitatively wrong answer. In Chemistry, perception failures tend to be *quantitative*—misreading a number or charge leads to a numerically wrong answer through otherwise correct reasoning. Both failure types are invisible to standard V+T evaluation but are decomposed by DISSECT's five-mode framework.

E Human Oracle Annotation Guidelines

Human annotators followed a structured protocol to construct oracle annotations:

(1) **Structural identification:** Label all discrete visual entities (molecules, organelles, apparatus components) with their standard scientific names.
(2) **Quantitative extraction:** Transcribe all numerical values, measurements, and symbolic notation visible in the image.
(3) **Spatial relationships:** Describe relative positions, connections, and directional indicators (arrows, flow lines).
(4) **Implicit properties:** Annotate properties that require visual interpretation but not domain reasoning (e.g., "lines indicate double bond," "blue shading indicates deoxygenated blood").
(5) **Question-relevance filtering:** Prioritize annotations relevant to the question, but include all identifiable visual content to avoid introducing annotator bias about the solution path.

Annotators were undergraduate and graduate students in Chemistry and Biology. Each annotation was independently verified by a second annotator, with disagreements resolved by a subject-matter expert.