OMNIREASONING: PUSHING THE LIMITS OF AUDIO-VISUAL JOINT REASONING

Anonymous authors
Paper under double-blind review

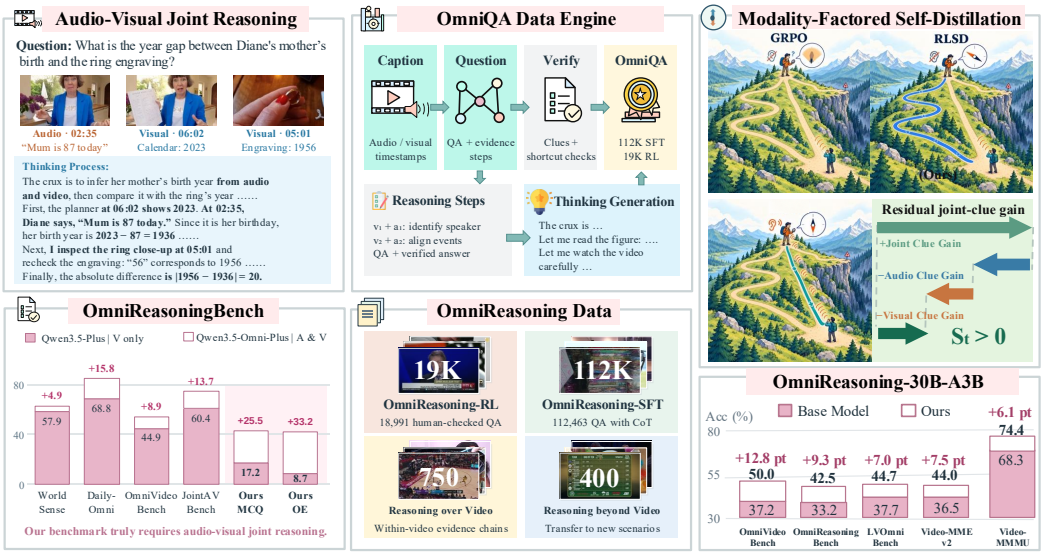


Figure 1: **OmniReasoning: a benchmark, data engine and learning method for audio-visual joint reasoning.** Unlike previous benchmarks, our benchmark OmniReasoningBench truly requires both audio and visual inputs for joint reasoning. Besides this benchmark, our data engine OmniQA additionally produces large-scale training data with evidence-grounded questions. Further, our learning method MFSD leverages the gain from cross-modality joint clues to assign credit at token level, enabling effective exploration along audio-visual joint reasoning. In comparison, previous method GRPO offers only outcome-level guidance, and RLSD does not consider cross-modality interaction. With our training data and learning method, our model achieves large improvements over base model on audio-visual, long video, and general video benchmarks.

ABSTRACT

Recent advances have enabled unified omni-modal models in understanding audio, vision, and language. However, existing benchmarks, training data, and learning methods largely treat the modalities independently, leaving the capability of audio-visual joint reasoning poorly evaluated and insufficiently elicited. We address this gap with a benchmark, data engine, and learning method. First, we introduce **OmniReasoningBench**, a benchmark where both audio and visual evidence are indispensable. It comprises 1,150 multiple-choice and open-ended questions across two tasks, *reasoning over video* and *reasoning beyond video*. Second, we develop a data engine **OmniQA**. It automatically constructs evidence-grounded QA pairs that explicitly necessitate audio-visual joint reasoning, together with time-stamped clue chains that guide the annotation of thinking process. Besides our benchmark, this engine produces training data OmniReasoning-SFT-112K and OmniReasoning-RL-19K. Finally, we propose an on-policy self-distillation method **Modality-Factored Self-Distillation (MFSD)**. It evaluates each sampled response under modality-specific clue contexts, disentangling the contributions of individual clues and their cross-modal interactions for token-level credit assignment. With our training data and learning method, our model **OmniReasoning-30B-A3B** achieves 50.0% on OmniVideoBench and 42.5% on OmniReasoning-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Bench, improving the base model Qwen3-Omni-30B-A3B-Thinking by 12.8 and 9.3 percentage points, respectively. Moreover, it delivers substantial gains on general and long-video benchmarks, including Video-MME-v2. We hope our work offers a solid step for facilitating future research in omni-modal joint reasoning.

1 INTRODUCTION

Understanding real-world videos often requires reasoning across what has been heard and what has been seen. Consider the example in Figure 1: to determine the gap between a mother's birth year and the year engraved inside a ring, a model has to combine a spoken age with an observed calendar date, infer the corresponding birth year, and then compare it with the engraving. Neither modality alone provides sufficient evidence. The answer emerges only by connecting information across modalities. We refer to this capability as *audio-visual joint reasoning*. Recent omni-modal large language models (Omni-LLMs), such as Qwen-Omni (Xu et al., 2025a;b) and Nemotron 3 Nano Omni (NVIDIA, 2026), have made substantial progress toward unified audio-vision-language understanding. However, whether these omni models can reliably perform such cross-modal reasoning remains largely unclear.

A fundamental obstacle is that existing benchmarks do not consistently make joint reasoning necessary. Several recent audio-visual benchmarks (Hong et al., 2025; Zhou et al., 2025; Li et al., 2025; Chao et al., 2026) position themselves as evaluations of omni-modal understanding and conduct modality ablation experiments. However, ablation studies do not fully validate that a benchmark actually requires information from multiple modalities to answer the questions. In our experiments, Qwen3.5-Plus (Qwen Team, 2026a) achieves 57.85%, 68.76%, 44.90%, and 60.43% accuracy on WorldSense, Daily-Omni, OmniVideoBench, and JointAVBench, respectively, when provided with *visual-only inputs* and *no audio*. These results indicate that a substantial fraction of the questions remain answerable without audio, limiting the extent to which such benchmarks can distinguish genuine audio-visual joint reasoning from strong single-modality understanding.

Motivated by these findings, we introduce **OmniReasoning**, a unified framework for evaluating and eliciting omni-modal joint reasoning through explicit modality-specific evidence. At its core is **OmniReasoningBench**, a benchmark of 1,150 multiple-choice and open-ended questions spanning two settings: *reasoning over video*, which requires connecting observations across events within a video, and *reasoning beyond video*, which applies information derived from a video to a new scenario or figure. The questions are deliberately constructed around dependencies between audio and visual observations, thereby minimizing the possibility of solving them from a single modality alone. Take Figure 1 as an example, when Qwen3.5-Plus model evaluated without audio, the accuracy drops to 17.2% and 8.7% on the multiple-choice and open-ended questions, respectively. Moreover, our ablation studies show that audio-visual joint reasoning provides gains beyond either modality alone, with improvements exceeding the gains obtained from simply combining the questions solved independently by the two modalities. Together, these results demonstrate that the benchmark is able to test the capability of integrating complementary evidence across modalities.

Making such questions at scale, however, presents a second challenge: training data has to preserve the same evidence dependencies rather than merely pair arbitrary audio, visual, and textual content. To address this, we develop a data engine **OmniQA**, which automatically constructs evidence-grounded questions that explicitly link audio and visual observations. In addition to generating question-answer pairs, OmniQA produces timestamped clues and dependency chains that connect observed evidence to intermediate inferences and the final answer (Figure 3). After verification, these reasoning chains are used to synthesize thinking processes for supervised fine-tuning, yielding *OmniReasoning-SFT-112K*. The modality-specific clue annotations are also retained as structured supervision for reinforcement learning, resulting in *OmniReasoning-RL-19K*. Thus, OmniQA provides not only the benchmark and large-scale training data, but also an explicit representation of how evidence from different modalities contributes to a reasoning process.

The same evidence structure further enables a more targeted learning objective. Existing reasoning-oriented post-training methods (Shao et al., 2024; Zheng et al., 2025; Yu et al., 2025) typically assign credit to a response according to its overall quality, without distinguishing whether a prediction is supported by audio evidence, visual evidence, or their interaction. We therefore propose **Modality-Factored Self-Distillation (MFSD)**, an on-policy self-distillation method that factors token-level

credit assignment according to modality-specific evidence. As in Figure 1, for each sampled response, the model evaluates the token likelihood under four types of clue context: no clues, audio clues, visual clues, and joint audio-visual clues. The resulting likelihood difference quantifies both the contribution of individual modalities and the non-additive interaction between them. Specifically, MFSD measures the interaction by subtracting the individual audio and visual gains from the gain obtained under joint clues, and combines this interaction with the overall support provided by the joint clues to derive fine-grained token-level supervision through the weighting mechanism of RLSD (Yang et al., 2026). This design explicitly encourages the model to generate reasoning steps that rely on complementary cross-modal evidence rather than exploiting one modality in isolation.

Combining the OmniQA training data with MFSD yields our model **OmniReasoning-30B-A3B**. It achieves 50.0% accuracy on OmniVideoBench and 42.5% on OmniReasoningBench, improving over the baseline (Qwen3-Omni-30B-A3B-Thinking) by 12.8 and 9.3 percentage points, respectively. Beyond the targeted evaluation, the model also delivers substantial improvements on general and long-video understanding benchmarks, including LVOmniBench, Video-MMMU, and Video-MME-v2. These results suggest that explicitly constructing, annotating, and optimizing for cross-modal evidence dependencies can strengthen not only audio-visual joint reasoning, but also broader video understanding. We hope that **OmniReasoning** provides a unified foundation for studying and improving joint reasoning across modalities in future omni-modal models.

Our contributions are summarized as follows:

1. **A benchmark for genuine audio-visual joint reasoning**. We introduce OmniReasoning-Bench, a benchmark where audio and visual observations are jointly necessary, enabling a more rigorous evaluation of omni-modal reasoning.
2. **An evidence-grounded data engine for joint reasoning**. We develop OmniQA, an automated data engine that constructs audio-visual QA pairs together with timestamped clues and dependency chains, providing explicit supervision for cross-modal training.
3. **A modality-aware learning method for joint reasoning**. We propose MFSD, a reinforcement learning method that highlights cross-modality interactions and offers token-level credit assignment. Combined with OmniQA data, MFSD substantially improves both audio-visual joint reasoning and general video understanding.

2 Related Work

Omni-LLMs and audio-visual joint reasoning. Recent omni-modal models support native audio-visual understanding (Google, 2026; ByteDance, 2026; Meta, 2026; Xiaomi, 2026; Xu et al., 2025a;b; Qwen Team, 2026b; Tang et al., 2025; NVIDIA, 2026). WorldSense (Hong et al., 2025), Daily-Omni (Zhou et al., 2025), OmniVideoBench (Li et al., 2025), and LVOmniBench (Tao et al., 2026) benchmark this capability. JointAVBench (Chao et al., 2026) explicitly evaluates cross-modality dependence. AV-Reasoner (Lu et al., 2025) studies clue-grounded counting and Video-MMMU (Hu et al., 2025) tests knowledge acquisition from instructional videos. In contrast, our benchmark necessitates joint reasoning over audio and visual observations.

Audio-visual instruction data. OmniVideo-100K (Cai et al., 2026) preserves cross-segment associations and audio-visual correspondence through entity-anchored scripts and clue-guided questions. OmniVideo-R1 (Chen et al., 2026) filters LLaVA-Video and Video-Vista questions for cross-modal grounding and fusion. Our OmniQA leverages explicit evidence for question construction, dependency chains for thinking generation, and separate audio/visual clues for RL scoring, while extending video-derived knowledge to new inputs through *reasoning beyond video*.

Reinforcement learning and self-distillation. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) develops group-relative outcome rewards while Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025) considers sequence-level importance weighting. Video-R1 (Feng et al., 2025) and OmniVideo-R1 (Chen et al., 2026) apply RL to video and audio-visual reasoning, and Video-KTR (Wang et al., 2026) addresses key-token attribution. On-policy self-distillation (Zhao et al., 2026; Hübotter et al., 2026) converts additional context into token-level feedback, which RLSD (Yang et al., 2026) maps to bounded, sign-preserving advantage weights. We propose MFSD which retains RLSD's weighting, outcome verifier, and policy objective, combining a centered interaction contrast from separate and joint clues with the joint-clue gain.



Figure 2: **OmniReasoningBench tasks and examples.** *Reasoning over video* connects observations across events; *reasoning beyond video* applies video-derived knowledge to a new scenario. Orange and blue mark audio and visual clues, with numbered markers linking evidence to timestamps.

3 Benchmark, Data Engine, and Learning Method

Our work, OmniReasoning, seeks to advance audio-visual joint reasoning with three components: an evaluation benchmark, a data engine, and a learning method. They are connected through explicit dependencies between audio and visual evidence. Specifically, the OmniQA data engine constructs both the questions of OmniReasoningBench and the training corpora, and its modality-specific clue annotations further provide the supervision signal for our learning method, MFSD.

3.1 OmniReasoningBench: A Benchmark for Audio-Visual Joint Reasoning

OmniReasoningBench evaluates whether models can connect complementary audio and visual evidence to answer a question. It contains 750 *reasoning over video* and 400 *reasoning beyond video* questions, each with a reference answer and an annotated evidence chain (Figure 2).

Reasoning over video. These questions ask about the content within videos. Every answer can be derived by chaining multi-hop audio and visual evidence, and no single modality suffices. A spoken reference can identify which object to inspect, while a visible action can identify the relevant utterance. Correct answers therefore require joint reasoning over audio and visual evidence. This setting covers ten task types and includes 375 multiple-choice and 375 open-ended questions.

Reasoning beyond video. Omni reasoning can go beyond the understanding of a given video and transfer to new scenarios. The questions in this setting therefore ask models to carry knowledge acquired from a video into a new situation, figure, or numerical condition. For example, a spoken explanation and a visual demonstration establish a rule that must then be applied to a new diagram. In this case, the evidence chain still exists in the video, while the reasoning extends beyond it. There are in total 400 questions spanning nine task types, with 250 multiple-choice and 150 open-ended questions. Appendix A.3 provides the input-format breakdown.

Evidence of joint reasoning. On *reasoning over video*, joint inputs outperform a single-modality oracle that counts a question as solved whenever either the audio-only or the visual-only run answers it correctly. The gains range from 9.9 to 26.7 percentage points across three models and both question formats (Appendix A.4).

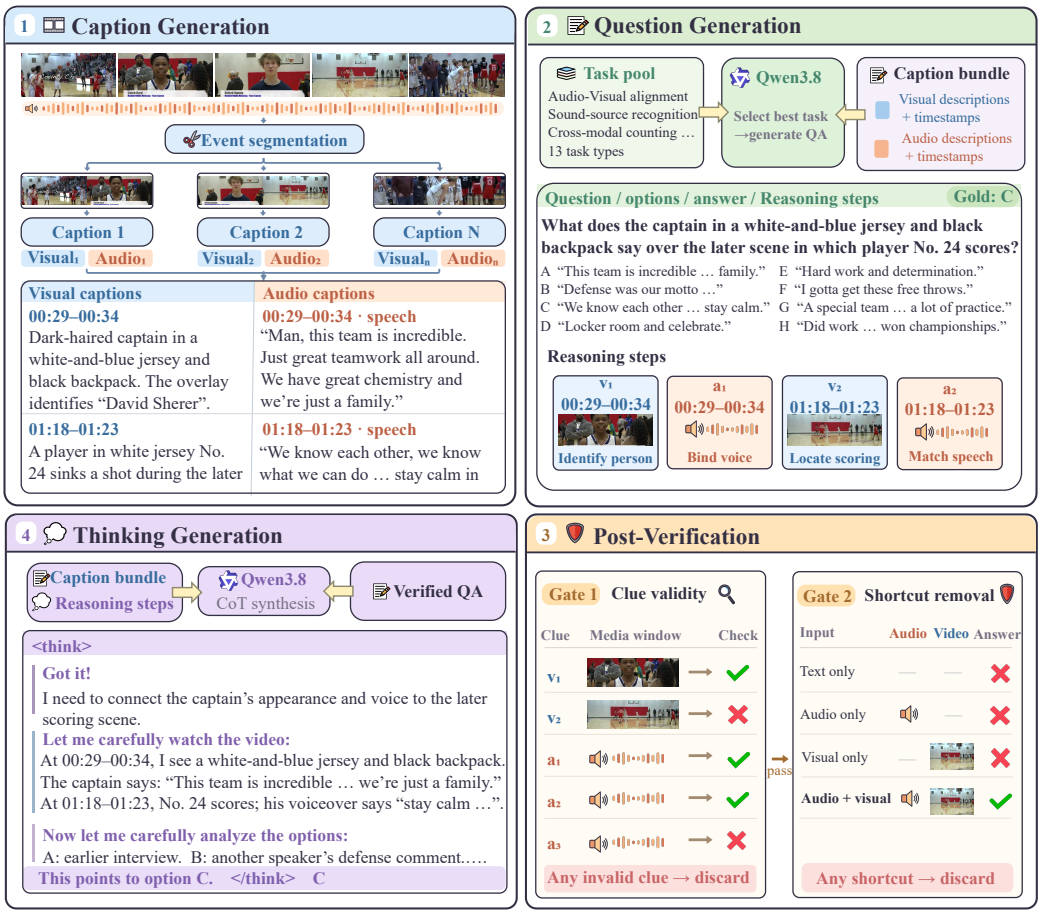


Figure 3: **OmniQA data engine.** Gemini-3.1-Pro annotates timestamped audio-visual descriptions. Qwen3.8 generates QA pairs and reasoning steps, followed by clue validation and shortcut screening. Timestamped captions, verified QA pairs, and evidence chains then guide thinking generation.

3.2 OmniQA: A Data Engine for Audio-Visual Joint Reasoning

To preserve these evidence dependencies in training data, **OmniQA** constructs QA pairs together with timestamped audio and visual clues. The verified evidence chains guide thinking generation for SFT and provide modality-specific context for MFSD during RL (Figure 3).

Constructing questions from evidence. OmniQA segments each video into events and generates separate audio and visual descriptions with timestamps. Conditioned on these descriptions and a task specification, the question generator produces a question, a reference answer, and a dependency chain that links observations to intermediate inferences and the final answer. For multiple-choice questions, it also generates candidate options.

Verifying questions and clues. Structural checks validate the dependency graph, caption references, and timestamps. A caption-conditioned solver checks the reference answer, and a media verifier checks each clue against its supporting audio or visual clip. To screen for modality shortcuts, we evaluate each question under question-only, audio-only, visual-only, and joint audio-visual conditions. We retain a candidate only when its clues are valid, the joint condition yields the correct answer, and none of the restricted conditions does.

Generating thinking processes. The thinking generator receives timestamped audio and visual captions, the verified QA pair, the evidence chain, and any additional figure, but not the source video. It expands the chain into observations, intermediate inferences, and a final answer, comparing multiple-choice options and showing numerical calculations (Appendix F).

Training datasets. The released datasets comprise **OmniReasoning-SFT-112K**, with 112,463 samples containing synthesized thinking processes, and **OmniReasoning-RL-19K**, with 18,991 human-

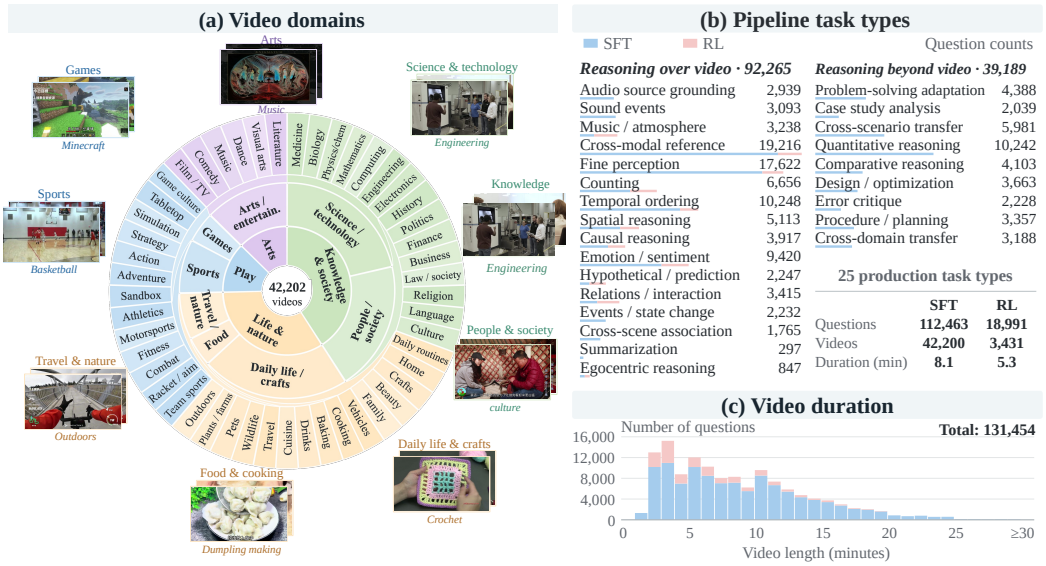


Figure 4: **OmniReasoning released training-data distributions.** The SFT and RL corpora span eight content domains, 25 production task types, and varied video durations.

validated questions and evidence annotations. Both cover *reasoning over video* and *reasoning beyond video* (Figure 4). During the SFT stage, the model learns from the synthesized responses. During the RL stage, the retained audio and visual clues, r_A and r_V, instead provide privileged context for evaluating the model's sampled responses.

3.3 Modality-Factored Self-Distillation: A Learning Method

The modality-specific clues created by OmniQA provide not only data annotations but also a learning signal for reinforcement learning. Our method stems from RLSD (Yang et al., 2026), a widely adopted RL approach that re-scores a sampled response under privileged clue contexts and converts the resulting likelihood gains into token-level advantage weights. However, a likelihood gain under joint audio-visual clues does not by itself indicate joint reasoning. The gain may come from one modality alone, so a response that exploits a single modality can receive the same guidance as one that genuinely integrates both. **Therefore, we propose MFSD, which extends RLSD with cross-modality interaction guidance.** Besides the joint-clue support used by RLSD, MFSD disentangles the likelihood gain that emerges only when audio and visual clues are combined, and uses the interaction signal for fine-grained token-level credit assignment (Figure 5).

MFSD. Given a question with its original audio-visual input, we first sample a group of responses without privileged clues, as in standard group-based reinforcement learning. The same actor then scores each sampled response under four clue contexts: no clues, audio clues, visual clues, and joint audio-visual clues. The likelihood changes across these contexts yield two complementary evidence signals: the overall support provided by the joint clues, and the cross-modality interaction that neither modality explains individually. Both signals are detached and converted into bounded token-level weights that modulate the outcome advantage.

Scoring the same response under different clue contexts. Let x contain the question and its original audio-visual input. We sample G responses $y^{(i)}$ from $\pi_{\text{old}}(\cdot \mid x)$ without privileged clues. Outcome rewards define $A_i = (R_i - \mu_R)/(\sigma_R + \varepsilon)$, where μ_R and σ_R are the group reward mean and standard deviation. Given OmniQA's audio and visual text clues r_A and r_V, define $r_\varnothing = \varnothing$ and $r_{AV} = r_A \oplus r_V$. The actor scores each sampled token as

$$\ell_{i,t}^{M} = \log \pi_\theta(y_t^{(i)} \mid x, r_M, y_{<t}^{(i)}), \qquad M \in \{\varnothing, A, V, AV\}. \tag{1}$$

The original media, sampled prefix, and actor weights remain fixed across views. The detached gains are $\Delta_{i,t}^{M} = \text{sg}(\ell_{i,t}^{M} - \ell_{i,t}^{\varnothing})$ for $M \in \{A, V, AV\}$, where sg denotes stop-gradient.

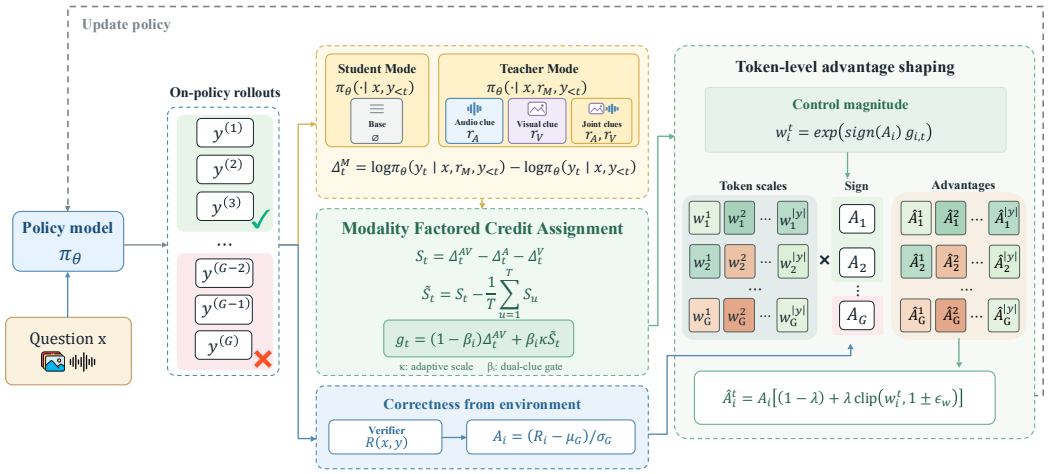


Figure 5: **Modality-Factored Self-Distillation.** The actor scores the same response under four clue contexts. Joint-clue support and non-additive audio-visual interaction determine bounded token-level advantage weights.

Separating joint support from interaction. MFSD subtracts the individual clue gains from the joint gain:

$$S_{i,t} = \Delta_{i,t}^{AV} - \Delta_{i,t}^{A} - \Delta_{i,t}^{V} = \mathrm{sg}\big(\ell_{i,t}^{AV} - \ell_{i,t}^{A} - \ell_{i,t}^{V} + \ell_{i,t}^{\varnothing}\big)\,. \tag{2}$$

Intuitively, $S_{i,t}$ is positive only when the two clue sets together raise a token's likelihood by more than the sum of their individual effects. This is a non-additivity comparison in log-likelihood: a positive $S_{i,t}$ does not by itself imply that the joint clues raise the token's likelihood over the no-clue context, since the joint gain and the interaction can differ in sign. It is zero whenever the joint gain is fully explained by a single modality. For example, audio clues alone may account for the joint gain while visual clues add nothing. Under the joint-consistency assumption in Appendix B, this contrast is a token-wise increment in conditional pointwise mutual information (PMI). Suppressing the rollout index,

$$S_t = \mathcal{I}_t - \mathcal{I}_{t-1}, \qquad \mathcal{I}_t = \log \frac{P(r_A, r_V \mid x, y_{\le t})}{P(r_A \mid x, y_{\le t})P(r_V \mid x, y_{\le t})}. \tag{3}$$

where $y_{\le 0} = \varnothing$. This interpretation concerns the model's response to clue conditioning; it does not certify reasoning-step correctness.

Combining the two evidence signals. We center interaction scores within each response using the valid-token mask $m_{i,t}$ and $T_i = \sum_t m_{i,t} > 0$:

$$\bar{S}_i = \frac{1}{T_i}\sum_t m_{i,t} S_{i,t}, \qquad \widetilde{S}_{i,t} = m_{i,t}(S_{i,t} - \bar{S}_i). \tag{4}$$

The combined token score is

$$g_{i,t} = (1-\beta_i)\Delta_{i,t}^{AV} + \beta_i \kappa \widetilde{S}_{i,t}, \qquad \beta_i = \beta\, \mathbf{1}[r_A \neq \varnothing \wedge r_V \neq \varnothing]. \tag{5}$$

The two signals play complementary roles: joint support retains useful evidence even when it comes from a single modality, while centered interaction emphasizes tokens with above-average cross-modal support. Centering removes the response-level mean of the interaction score, so the interaction term redistributes credit within a response rather than rescaling the response as a whole (Appendix B). The coefficient β balances these signals, and κ calibrates their scales (Appendix A).

Assigning token-level policy credit. Following RLSD, we convert $g_{i,t}$ into bounded advantage weights:

$$\begin{aligned} w_{i,t} &= \exp(\mathrm{sign}(A_i) g_{i,t})\,, \\ \widehat{A}_{i,t} &= A_i\,[(1-\lambda) + \lambda\, \mathrm{clip}(w_{i,t}, 1-\epsilon_w, 1+\epsilon_w)]\,. \end{aligned} \tag{6}$$

Higher $g_{i,t}$ strengthens positive advantages and reduces negative penalties without reversing their signs. The clipped policy objective uses $\widehat{A}_{i,t}$ (Appendix A.1). Detached scores and weights prevent gradients through clue conditioning; rollouts exclude privileged clues. The three clue-conditioned

Table 1: **OmniReasoningBench accuracy (%).** MCQ and OE denote multiple-choice and open-ended questions. Our model is initialized by Qwen3-Omni-30B-A3B-Thinking.

		Reasoning over video		*Reasoning beyond video*		
Model	**Modality**	**MCQ**	**OE**	**MCQ**	**OE**	**Overall**
Proprietary models						
Gemini-3.7-Flash	Omni	57.6	47.7	72.0	58.0	57.6
Gemini-3.8-Flash	Omni	55.2	49.6	69.6	60.0	57.1
Gemini-3.5-Flash	Omni	53.3	46.9	66.4	51.3	53.8
Gemini-3.1-Pro	Omni	51.7	46.9	64.0	49.3	52.5
Doubao-2.0-Lite	Omni	54.7	47.2	57.2	40.7	51.0
Qwen3.5-Omni-Plus	Omni	42.7	41.9	60.4	40.7	46.0
Muse-Spark-1.2	Omni	38.1	36.0	60.0	58.0	44.8
Muse-Spark-1.1	Omni	36.4	35.9	56.0	46.0	41.8
Open-source models						
MiMo-V2.5	Omni	42.7	40.6	54.0	44.0	44.6
Kimi-K3	Visual	21.4	11.7	50.8	35.3	26.5
Qwen3.8-Max	Visual	19.7	10.5	46.8	34.7	24.6
Qwen3.5-Plus	Visual	17.2	8.7	46.0	30.0	22.5
Nemotron-3-Nano-Omni	Omni	25.6	22.9	36.0	13.3	25.4
Qwen3-Omni-30B-A3B-Thinking	Omni	31.9	25.9	46.4	32.0	33.2
Ours						
OmniReasoning-30B-A3B	Omni	48.0	30.1	58.8	32.7	42.5

scoring views require no extra rollouts or separate teacher parameters. Relative to RLSD, the modification is that the token score $g_{i,t}$ now carries cross-modality interaction guidance. Tokens whose likelihood increases only when audio and visual clues are combined receive larger positive advantages and smaller penalties, whereas tokens supported by a single modality alone are guided mainly by joint-clue support.

4 Experiments

Our experiments answer three questions: (1) Can OmniQA training data improve audio-visual joint reasoning? (2) Can the benefits transfer to general video understanding? (3) Does the cross-modality interaction guidance of MFSD improve over existing reinforcement learning methods?

Setup. We initialize our model OmniReasoning-30B-A3B from Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b) and perform SFT on all 112,463 examples. GRPO, RLSD, and MFSD start from the same SFT checkpoint with the same RL budget: 150 steps, 8 responses per question, and a global batch of 256 generated responses per update. MFSD uses $\lambda = \beta = 0.5$ after warm-up, and the audio and visual encoders remain frozen throughout. At evaluation time, reference evidence chains are withheld, and open-ended questions require producing the answer directly rather than selecting among options. We compare against recent proprietary omni-modal models (Google, 2026; ByteDance, 2026; Qwen Team, 2026b; Meta, 2026) and open models (Xiaomi, 2026; Xu et al., 2025b; Kimi Team, 2026; Qwen Team, 2026c;a; NVIDIA, 2026), which together span omni-modal and vision-only inputs (Table 1). Appendix A provides training and evaluation details.

4.1 Audio-Visual Joint Reasoning and Generalization

Results on OmniReasoningBench. On OmniReasoningBench (Table 1), our model reaches 42.5% overall, 9.3 points above its base model. The gains are consistent across both settings and both question formats: 16.1 and 4.2 points on the multiple-choice and open-ended questions of *reasoning over video*, and 12.4 and 0.7 points on those of *reasoning beyond video*. Nevertheless, a clear gap to the strongest proprietary model (57.6% for Gemini-3.7-Flash) remains, indicating that audio-visual joint reasoning is far from solved.

Generalization to public benchmarks. The improvement is not limited to OmniReasoningBench. Our model reaches 50.0% on OmniVideoBench, 12.8 points above the base model, with gains spanning every audio-type and duration group (Table 4, Appendix A.5). It also improves by 6.1–7.5

Table 2: **Accuracy on public video benchmarks (%)** (Li et al., 2025; Tao et al., 2026; Hu et al., 2025; Fu et al., 2026). Our model is initialized by Qwen3-Omni-30B-A3B-Thinking.

Model	OmniVideoBench	LVOmniBench	Video-MMMU	Video-MME-v2
Proprietary models				
Gemini-3.5-Flash	66.5	66.9	86.1	67.6
Gemini-3.1-Pro	60.7	60.9	84.3	53.4
Qwen3.5-Omni-Plus	53.8	53.2	84.6	47.9
Muse-Spark-1.1	55.8	52.3	85.2	59.5
MiMo-V2.5	51.6	52.8	79.1	51.8
Open-source models				
Nemotron-3-Nano-Omni	42.1	32.1	73.1	38.4
Qwen2.5-Omni-7B	29.3	32.0	53.9	25.3
Qwen3-Omni-30B-A3B-Thinking	37.2	37.7	68.3	36.5
Ours				
OmniReasoning-30B-A3B	50.0	44.7	74.4	44.0

Table 3: **Comparison of RL methods (accuracy, %).** GRPO, RLSD, and MFSD start from the same SFT checkpoint and use the same RL training budget.

Method	OmniReasoning Bench	OmniVideo Bench	LVOmniBench	Video-MME-v2	Video-MMMU
Qwen3-Omni-30B-A3B-Thinking	33.2	37.2	37.7	36.5	68.3
+ SFT	36.1	47.0	40.2	40.3	73.7
+ GRPO	38.3	44.8	39.8	39.1	72.9
+ RLSD	39.2	45.4	40.1	39.8	72.0
+ MFSD	42.5	50.0	44.7	44.0	74.4

points on the other three general and long-video benchmarks (Table 2). Training on evidence-grounded joint-reasoning data thus strengthens broader video understanding rather than overfitting.

4.2 Contribution of MFSD

Finally, we validate whether the cross-modality interaction guidance in MFSD contributes beyond existing RL methods, comparing GRPO, RLSD, and MFSD under an identical budget from the same SFT checkpoint (Table 3). On OmniReasoningBench, GRPO and RLSD outperform SFT by 2.2 and 3.1 points, respectively, whereas MFSD improves by 6.4 points. On existing benchmarks, GRPO and RLSD even fall below SFT, while MFSD consistently outperforms SFT by 0.7–4.5 points. These results indicate that outcome rewards and coarse privileged clues alone fit the targeted pattern but hurt general capability. By factoring credit by modality-specific evidence and explicitly rewarding cross-modality interaction, MFSD improves both joint reasoning and general video understanding.

5 Conclusion

We present **OmniReasoning**, a unified framework for evaluating and improving audio-visual joint reasoning in omni-modal large language models. We first introduce **OmniReasoningBench**, which constructs questions that require complementary evidence from both audio and vision, providing a more stringent evaluation of genuine cross-modal reasoning. We then develop the **OmniQA** data engine to automatically generate evidence-grounded reasoning data with timestamped modality-specific clues and dependency chains, yielding **OmniReasoning-SFT-112K** and **OmniReasoning-RL-19K**. Finally, we propose **Modality-Factored Self-Distillation (MFSD)**, which exploits these structured clues to provide modality-aware, token-level credit assignment for reinforcement learning. Training with OmniQA and MFSD produces our model **OmniReasoning-30B-A3B**, substantially improving audio-visual joint reasoning on both existing and our proposed benchmarks, while also benefiting general and long-video understanding. These results highlight the importance of explicitly modeling cross-modal evidence dependencies in both evaluation and learning. We hope OmniReasoning can serve as a foundation for future omni-modal research toward joint reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets. Specifically, generative AI assisted in constructing the OmniQA dataset. We did not use generative AI tools for the other tasks with required disclosure, including developing the core scientific contributions, deriving mathematical proof, designing the research methodology or experiments, conducting experimental analysis, selecting or verifying references, or making scientific claims; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to aid or polish writing, including language editing, readability improvements, and suggestions of alternative wording. We have reviewed all AI-assisted work. All AI-assisted text was carefully edited and approved by the authors; synthetic data were checked using the validation procedures described in the paper; and all technical content, claims, citations, and reported results were manually verified. We take responsibility for the final content of this work, including text, claims, code, data, annotations, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete training details, including data generation pipeline, model configurations, optimization objectives, hyperparameters, and implementation settings. These details, together with the evaluation protocols, are documented in the main text and appendices to enable reproduction of our results.

REFERENCES

ByteDance. Seed2.0, February 2026. URL https://github.com/ByteDance-Seed/Seed2.0.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. OmniVideo-100K: A Dataset for Audio-Visual Reasoning through Structured Scripts and Evidence Chains. *arXiv preprint arXiv:2606.14702*, 2026. URL https://arxiv.org/abs/2606.14702.

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. JointAVBench: A Benchmark for Joint Audio-Visual Reasoning Evaluation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.12772.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, et al. OmniVideo-R1: Reinforcing Audio-visual Reasoning with Query Intention and Modality Attention. *arXiv preprint arXiv:2602.05847*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, et al. Video-R1: Reinforcing Video Reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, et al. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding. *arXiv preprint arXiv:2604.05015*, 2026. URL https://arxiv.org/abs/2604.05015.

Google. Gemini Models, 2026. URL https://deepmind.google/models/gemini/.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. *arXiv preprint arXiv:2501.13826*, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, et al. Reinforcement Learning via Self-Distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Kimi Team. Kimi K3: Open Frontier Intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, et al. OmniVideoBench: Towards Audio-Visual Understanding Evaluation for Omni MLLMs. *arXiv preprint arXiv:2510.10689*, 2025.

Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. AV-Reasoner: Improving and Benchmarking Clue-Grounded Audio-Visual Counting for MLLMs. *arXiv preprint arXiv:2506.05328*, 2025.

Meta. Introducing Muse Spark, April 2026. URL https://about.fb.com/news/2026/04/introducing-muse-spark-meta-superintelligence-labs/.

NVIDIA. Nemotron 3 Nano Omni: Efficient and Open Multimodal Intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026b.

Qwen Team. Qwen3.8-Max: A New Bar for Coding and Cowork, August 2026c. URL https://qwen.ai/blog?id=qwen3.8.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, et al. LVOmniBench: Pioneering Long Audio-Video Understanding Evaluation for Omnimodal LLMs. *arXiv preprint arXiv:2603.19217*, 2026.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-KTR: Reinforcing Video Reasoning via Key Token Attribution. *arXiv preprint arXiv:2601.19686*, 2026.

Xiaomi. Introducing MiMo-V2.5, April 2026. URL https://mimo.xiaomi.com/mimo-v2-5/.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, et al. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025b.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-Distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, et al. Group Sequence Policy Optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A Implementation and Evaluation Details

Supervised Fine-Tuning. SFT uses the full 112,463 SFT training corpus, with learning rate 7×10^{-6}, microbatch size 2, and global batch size 128. The configuration samples video at 2 fps with at most 256 frames and 200,704 pixels per frame. Visual and audio encoders are frozen.

RL configuration. The GRPO, RLSD and MFSD runs are trained for 150 RL steps using 64 GPUs, a global batch size of 256, 8 rollout trajectories per question, and learning rate 10^{-6}. The global batch size counts generated responses per optimizer update. The binary verifier extracts the selected option and compares it to the reference letter. The KL coefficient is 10^{-3}. MFSD uses text clue factorization, actor weights for every scoring view, sequence-wise centering, $\epsilon_w = 0.2$, and EMA coefficient 0.9 for κ with maximum 10. The coefficients λ and β warm up to 0.5 over 10 and 20 steps. No additional global advantage whitening is enabled. Constant-reward groups receive zero reward advantage.

Inference and scoring. Evaluation on OmniReasoningBench, OmniVideoBench, LVOmniBench, Video-MME-v2, and Video-MMMU uses 2 fps video sampling with at most 512 frames and no audio truncation.

A.1 Policy objective

We substitute $\widehat{A}$ into the clipped policy objective, with $\rho_{i,t} = \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)}) / \pi_{\text{old}}(y_t^{(i)} \mid x, y_{<t}^{(i)})$:

$$\mathcal{L}_{\text{MFSD}} = -\mathbb{E}\Big[\Big\langle \min\Big(\rho_{i,t}\widehat{A}_{i,t}, \text{clip}(\rho_{i,t}, 1-\epsilon_p, 1+\epsilon_p)\widehat{A}_{i,t}\Big)\Big\rangle_m\Big] + \eta\mathcal{L}_{\text{KL}}. \tag{7}$$

Here $\langle\cdot\rangle_m$ denotes the masked response-token reduction and $\mathcal{L}_{\text{KL}}$ anchors the actor to the SFT reference policy. Teacher scores and advantage weights are detached. MFSD adds three scoring passes, reuses media features, and requires no extra rollouts or teacher parameters.

A.2 MFSD implementation and component analysis

Adaptive scale calibration. The joint-clue gain Δ^{AV} and centered interaction $\widetilde{S}$ can have different magnitudes. MFSD therefore calibrates their scales before mixing them in Equation 5. At update b, let $\widehat{\sigma}_b$ denote the standard deviation over valid response tokens in the batch; padding and non-response positions are excluded. We compute

$$\begin{aligned} q_b &= \min\left(\frac{\widehat{\sigma}_b(\Delta^{AV})}{\widehat{\sigma}_b(\widetilde{S}) + 10^{-6}},\, 10\right), \\ \kappa_b &= 0.9\,\kappa_{b-1} + 0.1\,q_b. \end{aligned} \tag{8}$$

This is the explicit update corresponding to $\text{EMA}_{0.9}$ of the capped standard-deviation ratio. The 10^{-6} term stabilizes the denominator, the cap limits the target scale when interaction variance is small, and the EMA smooths batch-to-batch fluctuations. Calibration approximately aligns the two signal scales; clipping and smoothing need not yield exact variance matching. The same κ_b is shared across response tokens in that update.

Centering and clue availability. Centering gives $\sum_t m_{i,t}\widetilde{S}_{i,t} = 0$, so the interaction term redistributes credit within a response without changing the sum of its token scores. It also removes response-wide additive offsets in S (Appendix B). The joint-clue term retains useful evidence even when it comes from one modality: if $\Delta^{AV}_{i,t} = \Delta^{A}_{i,t} > 0$ and $\Delta^{V}_{i,t} = 0$, then $S_{i,t} = 0$ despite positive clue support. When either clue set is absent, the gate sets $\beta_i = 0$ and uses joint-clue-only weighting.

Mixing and bounded weighting. The coefficient β balances joint support and calibrated interaction, whereas λ controls how strongly the resulting token weights modify the outcome advantage. They warm up to 0.5 over 20 and 10 steps, respectively. Setting $\beta = 0$ removes the interaction term; setting $\lambda = 0$ recovers the unweighted outcome advantage. These are algebraic limits of the weighting rule.

Table 4: **OmniVideoBench accuracy (%) by audio type and video duration** (Li et al., 2025).

Model	Audio Type			Video Duration (min)				Avg.
	Music	Sound	Speech	(0,1]	(1,5]	(5,10]	(10,30]	
Proprietary models								
Gemini-3.5-Flash	56.9	69.3	70.1	69.7	68.0	63.9	64.8	66.5
Gemini-3.1-Pro	51.1	65.8	61.9	67.3	64.2	59.5	53.0	60.7
Gemini-3-Pro	56.2	54.1	55.7	61.0	56.4	52.9	52.5	55.5
Qwen3.5-Omni-Plus	42.3	64.9	53.3	55.8	55.2	50.7	53.4	53.8
MiMo-V2.5	46.2	47.6	53.0	56.6	57.4	47.6	44.3	51.6
Gemini-2.0-Flash	29.7	40.3	43.2	49.4	43.2	41.1	34.9	41.5
Open-source models								
Nemotron-3-Nano-Omni	33.0	42.9	43.0	48.8	41.7	41.5	38.9	42.1
Qwen2.5-Omni-7B	23.1	25.3	30.7	41.6	27.4	25.3	26.7	29.3
Qwen3-Omni-30B-A3B-Thinking	26.4	37.2	38.5	46.8	35.6	35.5	35.2	37.2
Ours								
OmniReasoning-30B-A3B	39.6	46.3	52.0	52.4	49.9	49.8	48.9	50.0

Scoring and gradient separation. All four views use the same actor weights, original media, sampled response, and prefixes; only the privileged text clues vary. The three clue-conditioned views reuse the rollout and media features, with no additional response generation or separate teacher parameters. Evidence scores and advantage weights are detached, so their clue-conditioned derivatives do not enter the policy gradient.

A.3 Benchmark evaluation protocols

OmniReasoningBench contains 750 *reasoning over video* and 400 *reasoning beyond video* questions. *Reasoning over video* has 375 multiple-choice and 375 open-numeric questions. *Reasoning beyond video* has 250 multiple-choice and 150 open-numeric questions, with 288 video-plus-figure and 112 video-plus-text inputs.

At evaluation time, models receive the source video, question, and any additional figure; reference evidence chains are withheld. For *reasoning beyond video* questions, the additional figure or table is part of the question input: modality ablations hold it fixed and vary only the source video's audio and visual streams, so that ablation scores isolate the contribution of the video's own modalities rather than mixing in the removal of the new figure.

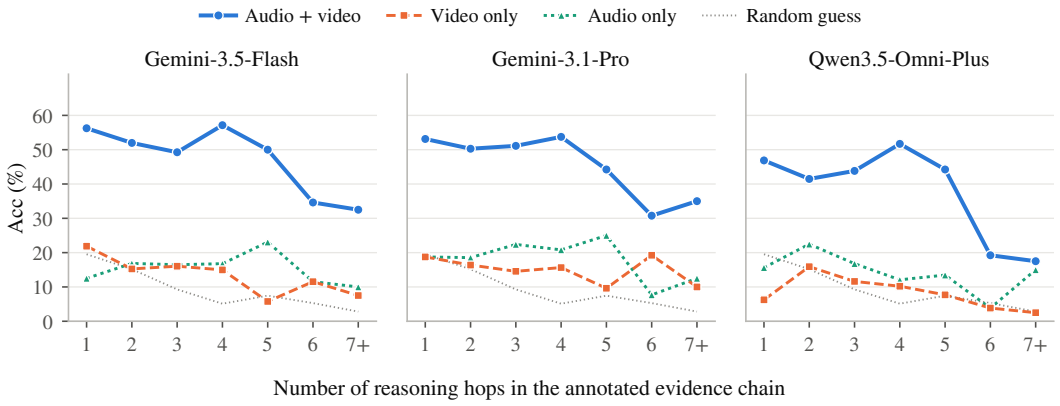


Figure 6: **Modality ablations on *reasoning over video*.** Audio-visual joint inputs outperform either modality alone for all three advanced Omni-LLMs in every evidence-chain-length group, supporting the importance of audio-visual evidence integration.

A.4 MODALITY ABLATIONS

To isolate modality effects, we evaluate each model with joint, visual-only, and audio-only inputs (Figure 6). Across the three evaluated Omni-LLMs, accuracy on *reasoning over video* falls from 42.3–50.1% with joint inputs to 10.9–14.9% with video alone and 16.3–20.4% with audio alone, supporting the importance of joint evidence.

We further compare Qwen3.5-Plus with visual inputs alone against Qwen3.5-Omni-Plus with joint audio-visual inputs, on four prior audio-visual benchmarks and on our *reasoning over video* questions (Table 5). Although these benchmarks study cross-modal understanding and report modality ablations, joint inputs add only 4.9–15.8 points over the visual-only model, and 44.9–68.8% of their questions remain answerable without audio; on our multiple-choice and open-ended questions, the visual-only model drops to 17.2% and 8.7%, while joint inputs recover 42.7% and 41.9%.

Table 5: **Visual-only versus joint-input accuracy across audio-visual benchmarks.** Qwen3.5-Plus receives visual inputs alone, while Qwen3.5-Omni-Plus receives joint audio-visual inputs.

Benchmark	**Qwen3.5-Plus**	**Qwen3.5-Omni-Plus**
WorldSense	57.9	62.8
Daily-Omni	68.8	84.6
OmniVideoBench	44.9	53.8
JointAVBench	60.4	74.1
OmniReasoningBench, *reasoning over video* (MCQ)	17.2	42.7
OmniReasoningBench, *reasoning over video* (OE)	8.7	41.9

To test whether the gains of joint inputs merely combine questions that either modality can answer alone, we perform a paired analysis on the same *reasoning over video* questions (Table 6). For each model, we compare joint audio-visual accuracy with a single-modality oracle that credits a question when either the audio-only or the visual-only run answers it correctly. Joint inputs exceed this oracle by 9.9–26.7 percentage points across the three Omni-LLMs and both question types. For Qwen3.5-Omni-Plus, 212 questions are answered correctly only with joint inputs, while 77 questions are answered by the single-modality oracle but missed with joint inputs. These results indicate that the benefit of audio-visual joint reasoning cannot be explained by the union of single-modality information gains.

Table 6: **Paired modality analysis on *reasoning over video*.** Each row compares joint audio-visual accuracy (AV) with audio-only (A), video-only (V), and a single-modality oracle (A∪V) that credits a question when either single-modality run answers it correctly, over the same paired questions. Joint inputs exceed the oracle in every setting, indicating that the gains of joint evidence cannot be explained by the union of single-modality successes.

Model	**Type**	**AV**	**V**	**A**	**A∪V oracle**	**AV − oracle**
Qwen3.5-Omni-Plus	MCQ	42.67	14.93	22.67	32.27	+10.40
Qwen3.5-Omni-Plus	OE	41.87	6.93	9.87	16.27	+25.60
Gemini-3.1-Pro	MCQ	51.73	22.40	26.13	41.87	+9.87
Gemini-3.1-Pro	OE	46.93	7.47	14.67	20.80	+26.13
Gemini-3.5-Flash	MCQ	53.33	19.73	20.53	37.07	+16.27
Gemini-3.5-Flash	OE	46.93	9.07	12.27	20.27	+26.67

A.5 OMNIVIDEOBENCH BREAKDOWN

Table 4 reports OmniVideoBench accuracy by audio type and video duration. OmniReasoning-30B-A3B improves over its base model in every audio-type and duration group.

B MATHEMATICAL PROPERTIES OF MFSD

We derive the information-theoretic interpretation of MFSD and its structural guarantees, following the evidence-ratio analysis of RLSD (Yang et al., 2026). All four views retain the original audio-video input x and the same sampled prefix $y_{<t}$; only the privileged text clues r_A, r_V, and $r_{AV} = (r_A, r_V)$ vary.

B.1 BAYESIAN INTERPRETATION OF THE INTERACTION SCORE

Assumption 1 (Joint consistency). *For fixed x, a single joint distribution $P(Y, R_A, R_V \mid x)$ satisfies*

$$\pi_\theta(y_t \mid x, r_M, y_{<t}) = P(y_t \mid x, r_M, y_{<t}), \qquad M \in \{\varnothing, A, V, AV\}, \tag{9}$$

and every conditioning event used in the derivations below has positive probability under $P(\cdot \mid x)$, so that all likelihood ratios are well defined.

Proposition 1 (Interaction as a conditional PMI increment). *Under Assumption 1, define the conditional pointwise mutual information of the realized clue pair by*

$$\mathcal{I}_t = \log \frac{P(r_A, r_V \mid x, y_{\le t})}{P(r_A \mid x, y_{\le t}) P(r_V \mid x, y_{\le t})}, \tag{10}$$

with $y_{\le 0} = \varnothing$. The interaction score in Equation 2 satisfies

$$S_t = \mathcal{I}_t - \mathcal{I}_{t-1}, \qquad \sum_{t=1}^{T} S_t = \mathcal{I}_T - \mathcal{I}_0. \tag{11}$$

Proof. For each $M \in \{A, V, AV\}$, Bayes' rule gives

$$P(y_t \mid x, r_M, y_{<t}) = \frac{P(r_M \mid x, y_{\le t}) P(y_t \mid x, y_{<t})}{P(r_M \mid x, y_{<t})}. \tag{12}$$

Dividing by $P(y_t \mid x, y_{<t})$, taking logarithms, and applying Assumption 1 yields

$$\Delta_t^M = \log \frac{\pi_\theta(y_t \mid x, r_M, y_{<t})}{\pi_\theta(y_t \mid x, y_{<t})} = \log \frac{P(r_M \mid x, y_{\le t})}{P(r_M \mid x, y_{<t})}. \tag{13}$$

Stop-gradient leaves this numerical identity unchanged. Substituting the three gains into S_t and collecting terms at each prefix,

$$\begin{aligned} S_t &= \Delta_t^{AV} - \Delta_t^A - \Delta_t^V \\ &= \log \frac{P(r_A, r_V \mid x, y_{\le t})}{P(r_A \mid x, y_{\le t}) P(r_V \mid x, y_{\le t})} - \log \frac{P(r_A, r_V \mid x, y_{<t})}{P(r_A \mid x, y_{<t}) P(r_V \mid x, y_{<t})} \\ &= \mathcal{I}_t - \mathcal{I}_{t-1}. \end{aligned} \tag{14}$$

Summing over t cancels all intermediate terms, giving $\sum_{t=1}^{T} S_t = \mathcal{I}_T - \mathcal{I}_0$. □

Interpretation. The sign of S_t compares joint support with the two individual gains:

$$\exp(S_t) = \frac{\pi_\theta(y_t \mid x, r_A, r_V, y_{<t})\, \pi_\theta(y_t \mid x, y_{<t})}{\pi_\theta(y_t \mid x, r_A, y_{<t})\, \pi_\theta(y_t \mid x, r_V, y_{<t})}. \tag{15}$$

Thus $S_t > 0$ indicates joint likelihood support beyond the product of the individual likelihood ratios; under Assumption 1, it increases the clue pair's conditional PMI. We stress that this is a non-additivity comparison in log-likelihood, not a statement of positive joint support: $\Delta_t^{AV} < 0$ and $S_t > 0$ can hold simultaneously when both individual gains are more negative than the joint gain. Positive interaction, positive joint support, and reasoning correctness are therefore distinct notions; MFSD keeps the joint-clue gain Δ_t^{AV} and the interaction S_t as separate branches in Equation 5 precisely so that each captures its own notion. In contrast, if $\Delta_t^{AV} = \Delta_t^A$ and $\Delta_t^V = 0$, then $S_t = 0$ despite positive audio-clue support. This motivates interaction-based credit assignment.

B.2 SEQUENCE CENTERING

Proposition 2 (Zero-sum and offset-invariant interaction). *Let $m_t \in \{0,1\}$, $N = \sum_t m_t > 0$, and define*

$$\bar{S} = \frac{1}{N}\sum_t m_t S_t, \qquad \widetilde{S}_t = m_t(S_t - \bar{S}). \tag{16}$$

Then $\sum_t \widetilde{S}_t = 0$. Adding a constant b to every S_t leaves $\widetilde{S}_t$ unchanged, and centering preserves the ordering of active-token scores.

Proof. By the definition of $\bar{S}$,

$$\sum_t \widetilde{S}_t = \sum_t m_t S_t - N\bar{S} = 0. \tag{17}$$

For $S'_t = S_t + b$, the mean is $\bar{S}' = \bar{S} + b$, hence $m_t(S'_t - \bar{S}') = m_t(S_t - \bar{S})$. For active positions t, u, $\widetilde{S}_t - \widetilde{S}_u = S_t - S_u$, proving order preservation. □

Consequently, centering removes response-wide additive offsets while retaining relative token credit. Since β_i and κ are constant within each response, Equation 5 gives

$$\sum_t m_t g_t = (1-\beta_i)\sum_t m_t \Delta_t^{AV}. \tag{18}$$

The interaction therefore contributes zero to the mean log-credit signal. This zero-sum property holds for the linear score g_t only: the exponentiation, clipping, and interpolation in Equation 6 are nonlinear and do not preserve it, so centering does not conserve the total advantage mass $\sum_t m_t \widehat{A}_t$. Centering also removes any response-wide constant component of S_t: a uniformly positive interaction score is fully absorbed, and a token with $S_t < 0$ still receives positive $\widetilde{S}_t$ when it lies above the response mean. MFSD thus optimizes relative interaction priority within each response rather than rewarding every token with positive raw interaction.

B.3 BOUNDED REWEIGHTING AND GRADIENT STRUCTURE

Write Equation 6 as $\widehat{A}_t = Ac_t$, where the detached multiplier is

$$c_t = (1-\lambda) + \lambda \,\mathrm{clip}\Big(e^{\mathrm{sign}(A)g_t}, 1-\epsilon_w, 1+\epsilon_w\Big). \tag{19}$$

Proposition 3 (Bounded and sign-preserving credit). *For $0 \le \lambda \le 1$ and $0 \le \epsilon_w < 1$,*

$$1 - \lambda\epsilon_w \le c_t \le 1 + \lambda\epsilon_w, \qquad \mathrm{sign}(\widehat{A}_t) = \mathrm{sign}(A), \qquad |\widehat{A}_t - A| \le \lambda\epsilon_w|A|. \tag{20}$$

Proof. The clipped exponential lies in $[1-\epsilon_w, 1+\epsilon_w]$. Multiplying by λ and adding $1-\lambda$ gives the stated bounds on c_t. Since $1 - \lambda\epsilon_w > 0$, multiplication by c_t preserves the sign of A, and $|c_t - 1| \le \lambda\epsilon_w$ gives the deviation bound. □

Policy-gradient structure. For the clipped surrogate in Appendix A.1, define

$$\begin{aligned} \rho_t &= \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\mathrm{old}}(y_t \mid x, y_{<t})}, \\ j_t(A) &= \min\{\rho_t A, \mathrm{clip}(\rho_t, 1-\epsilon_p, 1+\epsilon_p)A\}. \end{aligned} \tag{21}$$

Positivity and stop-gradient imply

$$j_t(\widehat{A}_t) = c_t j_t(A), \qquad \nabla_\theta j_t(\widehat{A}_t) = c_t \nabla_\theta j_t(A), \tag{22}$$

where the derivative exists. In an unclipped branch, the token contribution is

$$c_t A \rho_t \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}). \tag{23}$$

Thus privileged clues rescale each student-token gradient without reversing it or introducing derivatives through clue-conditioned scores.

C Reward and Evaluation Protocols

All RL rewards are binary over the 18,991 human-validated questions of OmniReasoning-RL-19K. Multiple-choice answers are verified by exact option-letter match (unparseable responses score 0); open-numeric answers are judged by a DeepSeek-V4-Flash model. The same judge and prompt are used for RL rewards and for all open-ended evaluation.

D QUALITATIVE COMPARISON OF MODEL RESPONSES

We compare complete recorded responses from Qwen3-Omni-30B-A3B-Thinking and OmniReasoning-30B-A3B on OmniVideoBench, LVOmniBench, Video-MME-v2, OmniReasoningBench, and Video-MMMU. The cases cover audio-visual temporal alignment, speaker and object reference, motion comparison, and application of video-derived knowledge to new inputs. Paired responses use identical questions and answer options. Each card gives the question and reference answer, followed by both original responses, including their full thinking text and final answers. Boldface highlights the evidence and decisions discussed in the case.

D.1 OMNIVIDEOBENCH: DIFFERENT DESCRIPTIONS OF THE SCENE AND INSCRIPTIONS

Question and reference answer

Question. After simulating the lunar landing mission, what is the content of letters behind them?

Options. A. United States B. Omega and United Statese C. Omega and Apollo D. Apollo and United States

Reference answer: A.

The base response describes a scene around 05:30 and reports two inscriptions, selecting D. The recorded post-trained response describes a later scene and reports only UNITED STATES, selecting the reference answer A. These are claims made in the responses; the timestamps and inscriptions have not been independently verified in this comparison.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user is asking for the specific content of letters visible behind characters after simulating the lunar landing mission in the provided video, requiring selection from the given multiple-choice options. Then, analyzing the video content reveals **the critical moment occurs around 05:30 during the mission simulation scene where astronauts are in a lunar module simulation**; at 05:33, the camera shows a control room background with **two distinct signs: one reading 'APOLLO' (clearly visible above 'UNITED STATES') and another displaying 'UNITED STATES'** (with the Apollo mission name prominently featured on the same sign), confirming both letters appear together as 'APOLLO' and 'UNITED STATES'. Next, evaluating the options shows A incorrectly lists only 'United States', B includes 'Omega' which appears later but not in this scene, C incorrectly states 'Omega and Apollo', while D accurately matches 'Apollo and United States' based on the verified visual evidence. Finally, the correct answer is D. </think>

D

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify the content of the letters displayed on the structure behind the astronauts after they complete their lunar landing simulation. The crux is determining which words are written on the gold truss structure behind them during the simulation scene.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the astronauts are simulating the lunar landing mission and observe the letters written on the structure behind them.

- At 06:45, the astronauts are shown in the simulator room preparing for the landing simulation.

- At 06:50, the simulation begins and the screen displays "APOLLO 11 LANDING SIMULATION" as they walk out of the simulator room.

- At 06:54, the astronauts walk past a large gold truss structure that serves as a backdrop for the mock lunar landing.

- **At 06:56, the camera shows a close-up of this gold truss structure, revealing the letters "UNITED STATES" written in red on its surface.**

- At 07:00, the astronauts stand in front of the gold structure, which is decorated with a United States flag on the left and a NASA logo on the right.

OmniReasoning-30B-A3B (continued)

- At 07:05, the astronauts pose for a photo in front of the gold truss structure, where the letters "UNITED STATES" are clearly visible behind them.

- Wait, let me re-read the question: it asks for the content of the letters behind them, and looking closely at the gold truss at 07:10, the letters spell out "UNITED STATES" in red, with no other brand names visible on this structure.

- If the letters behind them were "Omega and United States", the Omega logo would have to be painted directly on the gold truss structure behind the astronauts at 07:10, but that logo only appears on their wristwatches and in the promotional materials.

- The letters on the structure behind the astronauts are "UNITED STATES", which corresponds to option A.

Next, let me analyze the options:

- A. United States — This matches the letters painted on the gold truss behind the astronauts.

- B. Omega and United States — Omega is not written on the truss.

- C. Omega and Apollo — These words do not appear on the truss.

- **D. Apollo and United States — Apollo is not written on the truss.**

Finally, the correct option is A. </think>

A

D.2 VIDEO-MMMU: DIFFERENT STATED TEMPERATURES AND ANSWER CHOICES

Question and reference answer

Question. Based on the example question of Carnot heat engine from 9:00 to 12:00, what is the Carnot efficiency with a hot reservoir temperature of 350°C while the other information remains unchanged?

Options. A. 0.733 or 73.3% B. 0.582 or 58.2% C. 0.447 or 44.7% D. 0.545 or 54.5% E. 0.815 or 81.5% F. 0.339 or 33.9% G. 0.527 or 52.7% H. 0.705 or 70.5% I. 0.418 or 41.8% J. 0.896 or 89.6%

Reference answer: D.

The base response states a cold-reservoir temperature of 20 degrees C and selects G. The recorded post-trained response states 10 degrees C and selects D, matching the reference answer. The comparison establishes the difference between their stated assumptions and answers; it does not independently verify which temperature or timestamp appears in the video.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle this problem. First, the user is asking for the Carnot efficiency with a hot reservoir temperature of 350°C while other information remains unchanged, specifically referencing the example question from 9:00 to 12:00 in the video. Then, the video example at 9:00 shows a power plant with a hot reservoir temperature of 600°C (converted to 873 K) and a cold reservoir temperature of 20°C (converted to 293 K), resulting in a Carnot efficiency of 66% using the formula η_carnot = 1 - T_c/T_h. The problem states that the hot reservoir temperature is changed to 350°C while **the cold reservoir temperature remains unchanged at 20°C**. Next, converting the new hot temperature to Kelvin: 350°C + 273.15 = 623.15 K, and the cold temperature remains 20°C = 293.15 K. Using the Carnot efficiency formula η_carnot = 1 - (T_c / T_h), substituting the values gives η_carnot = 1 - (293.15 / 623.15) $\approx$ 1 - 0.4706 = 0.5294, or 52.94%. However, the provided options include 0.527 or 52.7%, which is likely due to using approximate conversions (e.g., 273 instead of 273.15), where **350°C = 623 K and 20°C = 293 K, giving η_carnot = 1 - (293/623) $\approx$ 0.527 or 52.7%**. Finally, comparing the calculated value of approximately 0.527 to the given options, Option G matches exactly. </think>

G

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to determine the Carnot efficiency when the hot reservoir temperature is 350°C, keeping the cold reservoir temperature at 10°C. The crux is converting the temperatures to Kelvin and applying the Carnot efficiency formula.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the Carnot heat engine example is calculated.

- At 08:48, the slide titled "Carnot Heat Engine Efficiency Example" is displayed, showing the formula for Carnot efficiency.

- At 09:30, the presenter converts the initial hot temperature of 200°C to Kelvin, calculating 200 + 273.15 = 473.15 K.

- **At 09:45, the presenter converts the cold temperature of 10°C to Kelvin, calculating 10 + 273.15 = 283.15 K.**

- At 10:14, the presenter calculates the efficiency for the initial temperatures, showing 1 - (283.15 / 473.15) = 0.402.

- If the hot temperature were 300°C instead of 350°C, the hot reservoir temperature would be 300 + 273.15 = 573.15 K, yielding an efficiency of 1 - (283.15 / 573.15) = 0.506, but the hot temperature is 350°C.

- To find the efficiency at 350°C, we convert the hot temperature: 350 + 273.15 = 623.15 K.

- We then calculate the ratio of the cold to hot temperature: 283.15 / 623.15 = 0.454521.

- **Subtracting this ratio from 1 gives the efficiency: 1 - 0.454521 = 0.545479.**

- The calculated efficiency of 0.545 corresponds to option D.

Next, let me analyze the options:

- A. 0.733 or 73.3% — incorrect.

- B. 0.582 or 58.2% — incorrect.

- C. 0.447 or 44.7% — incorrect.

- D. 0.545 or 54.5% — correct, as 1 - (283.15 / 623.15) = 0.545.

- E. 0.815 or 81.5% — incorrect.

- F. 0.339 or 33.9% — incorrect.

- G. 0.527 or 52.7% — incorrect.

- H. 0.705 or 70.5% — incorrect.

- I. 0.418 or 41.8% — incorrect.

- J. 0.896 or 89.6% — incorrect.

Finally, the correct option is D. </think>

D

D.3 VIDEO-MMMU: DIFFERENT CLAIMED DIAGRAM CORRESPONDENCES

Question and reference answer

Question. The process of spanning trees in is a key concept of graph theory. Which diagram illustrates the construction of a breadth-first spanning tree?

Options. A. a B. b C. c D. d E. e F. f G. g H. h I. i J. j

Reference answer: A.

Editorial note. The source question contains an empty referent after "in"; its wording is reproduced unchanged.

The base response claims that diagram i matches the breadth-first construction and selects I. The recorded post-trained response claims that diagram a matches and selects the reference answer A. These claims about the diagrams and their correspondence to the video are not independently verified here.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user asks to identify which diagram illustrates the construction of a breadth-first spanning tree based on the video content, with options labeled A through J. Then, the video demonstrates breadth-first search (BFS) for spanning trees: starting from vertex 'a', it connects all adjacent vertices (b, c, f) at level 1, then processes these vertices in alphabetical order to add their unvisited neighbors (d from b, g from c, e from f), forming the tree structure shown in the right panel. **The diagram labeled 'i' in the final practice section matches this BFS process exactly**, with the spanning tree built level by level starting from 'a' and expanding to 'b', 'c', 'f' first, followed by 'd', 'g', 'e', then 'h' and 'j', while the other diagrams (a-h) either show depth-first search or incorrect constructions. Next, the options A through J correspond to diagrams a-j in the video's practice sections, where diagram 'i' is explicitly labeled and visually demonstrates the breadth-first traversal order consistent with the explanation. **Finally, the correct choice is option I.** </think>

I

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify which diagram illustrates the construction of a breadth-first spanning tree based on the video. The crux is determining which of the labeled diagrams, (a) or (b), represents the step-by-step breadth-first search tree construction shown in the video.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the narrator demonstrates the step-by-step construction of a breadth-first spanning tree and identify its corresponding diagram.

- At 00:06, the narrator introduces the steps of a breadth-first search, explaining that we search horizontally before vertically.

- At 02:11, the video transitions to a practice problem titled "Breadth-First Practice" showing a graph with vertices a through g.

- At 02:28, the narrator starts at vertex "a" and connects it to its adjacent vertices "b", "c", and "f" in alphabetical order.

- At 02:51, the narrator moves to the next level and connects "d" to "b".

- At 03:08, the narrator connects "g" to "c".

- At 03:18, the narrator connects "e" to "f".

- **Wait, let me re-examine the diagram at 03:24: the completed tree is shown on the right, which matches the step-by-step progression of diagram (a) rather than diagram (b).**

- If diagram (b) were the correct representation, the steps would have to show a depth-first search where we go deep before branching, but the video explicitly demonstrates a breadth-first search where we explore all neighbors at the current depth before moving to the next level.

- The step-by-step construction shown in the video matches diagram (a), which corresponds to option A.

Next, let me analyze the options:

- A.a — This diagram correctly illustrates the breadth-first spanning tree construction shown in the video.

OmniReasoning-30B-A3B (continued)

- B.b — This diagram represents a depth-first search construction instead.
- C.c — This option is incorrect.
- D.d — This option is incorrect.
- E.e — This option is incorrect.
- F.f — This option is incorrect.
- G.g — This option is incorrect.
- H.h — This option is incorrect.
- I.i — This option is incorrect.
- J.j — This option is incorrect.

Finally, the correct option is A. </think>

A

D.4 OMNIVIDEOBENCH: ALIGNING ORCHESTRAL SOUND WITH PHYSICAL CONTACT

Question and reference answer

Question. When the singer first escalates her anger from vocal expression to physical contact with another singer, how does the orchestra musically intensify the dramatic conflict?

Options. A. The entire orchestra plays a single, extremely loud, short chord. B. The music suddenly stops completely, creating a breathtaking moment. C. The strings play a series of rapidly descending notes, conveying the character's sense of fall. D. Only the brass instruments play a single, piercing, high note.

Reference answer: A.



Video frame, 00:28

Qwen3-Omni attributes the musical change to descending strings. OmniReasoning locates the first physical contact and associates it with an orchestral chord, selecting A.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user is asking for the specific musical intensification technique used by the orchestra precisely when the singer escalates from vocal expression to physical contact with another singer in the video. Then, the video shows the singer in the black dress (the Queen) verbally escalating her anger through lines like 'Zerstörer! Todeschmerzen!' and 'Zu bist du mein! Mein! Mein!', with the physical contact occurring around 00:23 when she grabs the other singer's arms; at this exact moment (00:24-00:25), the orchestra features a dramatic shift where **the strings play a rapid, descending chromatic scale** (G#-F#-F-E#-E-D#-D-C#-C) that is fast, dissonant, and creates a falling sensation to amplify the character's emotional collapse. Next, evaluating the options: Option A describes a single loud short chord, but the video shows a descending scale rather than a single chord; Option B claims the music stops completely, but the orchestra continues playing with intensified volume; Option C states the strings play rapidly descending notes conveying a sense of fall, which matches the observed musical technique; Option D specifies only brass playing a high note, but the descending scale is performed by strings, not brass. **Therefore, the correct answer is C.** </think>

C

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify how the orchestra musically intensifies the dramatic conflict when the singer first escalates her anger from vocal expression to physical contact with another singer. The crux is determining the specific orchestral sound that accompanies this first physical escalation.

First, let me scan the video for the segments relevant to the question:

- I need to find the moment when the singer first escalates her anger from vocal expression to physical contact with another singer and observe the orchestral accompaniment.

- At 00:02-00:27, the singer in the black dress sings aggressively while pointing a sword at the singer in the white dress, but she does not make physical contact.

- **At 00:28, the singer in the black dress grabs the singer in the white dress by the neck with both hands, marking the first physical escalation of her anger.**

- **At 00:28-00:29, while she is choking the other singer, the orchestra plays a single, extremely loud, short chord.**

- At 00:30, the singer in the black dress releases the other singer and turns away.

- At 01:47-01:55, the singer in the black dress grabs the singer in the white dress by her hair and throat, which is a second physical escalation.

- At 01:56-02:06, the singer in the black dress sits on the bed while the singer in the white dress kneels before her.

- If the strings played a series of rapidly descending notes to convey the character's sense of fall, we would hear a descending scale in the violins at 00:28, but instead the orchestra plays a single, loud, short chord.

- The orchestral accompaniment at the first physical contact consists of a single, extremely loud, short chord, which corresponds to option A.

Next, let me analyze the options:

- A. The entire orchestra plays a single, extremely loud, short chord. — This matches the orchestral accompaniment at 00:28.

- B. The music suddenly stops completely, creating a breathtaking moment. — The orchestra plays a loud chord instead.

- C. The strings play a series of rapidly descending notes, conveying the character's sense of fall. — No descending scale is heard.

- D. Only the brass instruments play a single, piercing, high note. — The entire orchestra plays together.

Finally, the correct option is A. </think>

A

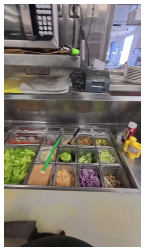
D.5 OMNIVIDEOBENCH: GROUNDING A SPOKEN PHRASE IN THE VISIBLE INGREDIENT

Question and reference answer

Question. When 'No, nada, aquí andamos.' is said, what ingredient is being handled?

Options. A. Bell pepper B. Onion C. Asparagus D. Egg

Reference answer: A.



Video frame, 00:14

Qwen3-Omni describes an ingredient absent from the options and substitutes egg. OmniReasoning links the target utterance to the bell pepper being handled and selects A.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user is asking to identify which ingredient is being handled when the phrase 'No, nada, aquí andamos.' is spoken in the video, with the provided options being bell pepper, onion, asparagus, or egg. Then, the video shows a first-person kitchen perspective where the speaker says the phrase at approximately 00:13-00:14 while handling a white, stringy ingredient being shredded or grated with a knife; this ingredient is visible in a metal container, matches the appearance of shredded cheese (possibly cheddar or a similar blend), and is being added to a mixture in a bowl alongside chopped green vegetables and diced red items. Next, evaluating the options: bell pepper is typically green or red and would be cut into strips or diced, not shredded into fine strands, so it is unlikely; onion is usually sliced or diced and while some types can be shredded, the video shows a yellowish-white shredded substance more characteristic of cheese; asparagus is a long green stalk that would be cut into pieces, not shredded into fine strands, so it is incorrect; and egg is a common ingredient in breakfast dishes like omelets or scrambles, with shredded cheese being frequently added to egg mixtures, and the sound of grating or shredding aligns with cheese preparation. Finally, **the ingredient being handled is shredded cheese, but since cheese is not an option, the next most plausible choice is egg** because it is a core component of the dish being prepared (likely a breakfast burrito or omelet), and the shredded cheese is added to the egg mixture, making egg the best answer among the given choices. </think>

D

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify which ingredient is being handled at the moment the phrase "No, nada, aquí andamos." is spoken in the video. The crux is matching the spoken Spanish phrase to the specific food item the cook is preparing at that exact timestamp.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the spoken phrase "No, nada, aquí andamos." and watch what ingredient the cook is handling on screen at that moment.

- At 00:04, the cook is holding a plate with a finished breakfast dish containing an omelet, sausage, and hash browns, and asks "Tú a cuál andas?" to inquire about someone's order.

- **At 00:13, the cook responds with "No, nada, aquí andamos." while preparing a new order, and the camera shows his gloved hands holding a yellow bell pepper on a cutting board.**

- At 00:14, the cook begins slicing the yellow bell pepper with a knife, confirming he is preparing this specific ingredient.

- If the cook were handling onions or asparagus instead, we would see him chopping red or green vegetables at 00:14, but he is clearly slicing a yellow bell pepper.

- At 00:16, the cook finishes slicing the bell pepper and moves on to the next step of the preparation.

OmniReasoning-30B-A3B (continued)

- **The ingredient being handled during the spoken line is a bell pepper, which corresponds to option A.**

Next, let me analyze the options:

- A. Bell pepper — The cook is slicing a yellow bell pepper at 00:14 while the phrase is spoken.

- B. Onion — This ingredient is not being handled at 00:14.

- C. Asparagus — This ingredient is not being handled at 00:14.

- D. Egg — This ingredient is not being handled at 00:14.

Finally, the correct option is A. </think>

A

D.6 LVOmniBench: Distinguishing background music from scene audio

Question and reference answer

Question. When the woman is having her hair trimmed in the hair salon, what is the most likely source of the background music that can be heard?

Options. A. The music playing in the salon. B. Added during video editing. C. Music playing from the woman's phone. D. Music playing on the salon's television.

Reference answer: B.



Video frame, 08:04

Qwen3-Omni relies on the expectation that salons play ambient music. OmniReasoning instead discusses the recorded audio characteristics and selects the reference answer B.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user wants to know the source of the background music playing during the hair salon scene in the video, with four specific options provided for analysis. Then, the video shows the woman getting her hair cut at a salon between approximately 00:24 to 00:33 and again around 03:05 to 03:15; during these segments, the music is a low-volume, ambient pop or electronic track that plays continuously throughout the salon scene without changing or stopping, with no visible speakers, televisions, or headphones indicating a personal source, and the sound quality matches typical salon background music that's pleasant but not intrusive. Next, evaluating the options: Option A (music playing in the salon) aligns with the visual context as salons commonly play background music for clients, and the consistent, ambient nature matches this source; Option B (added during video editing) is unlikely because the music is integrated into the scene without artificial separation and follows natural transitions, indicating it's part of the original footage; Option C (music playing from the woman's phone) is improbable since she's seated and looking at her phone or interacting with the stylist without visible headphones or a phone speaker; Option D (music playing on the salon's television) is less likely because the music lacks the typical TV broadcast characteristics like ads or dialogue, and the sound quality suggests a general background speaker system rather than a television source. Finally, **the most logical explanation is that the music is ambient background music played throughout the salon**, which is standard practice for salons to create a relaxed environment for clients, making Option A the correct choice. </think>

A

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify the most likely source of the background music playing while the woman is getting her hair trimmed in the salon. The crux is determining whether the audio is an environmental recording of the salon or a post-production music track added during editing.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the woman is getting her hair trimmed and analyze the background audio.

- At 07:53, the woman arrives at the hair salon, showing the front window with "THE LIGHT SALON" written on it.

- At 07:58, she is sitting in the salon chair while a stylist prepares to cut her hair.

- At 08:04, the stylist is blow-drying her hair, and a light, upbeat acoustic pop song is playing in the background.

- If the music were playing from the salon itself, there would be some ambient room noise or other client chatter mixed with the track, but the audio is completely clean and isolated.

- Wait, let me re-read the options: the music is a non-diegetic track added during editing, which means it is not part of the original salon environment.

- **The clean, isolated audio track of the music indicates it was added during video editing, which corresponds to option B.**

Next, let me analyze the options:

- A. The music playing in the salon — this is unlikely because there is no ambient room noise or other client chatter mixed with the track.

- B. Added during video editing — this is highly likely because the music is a clean, isolated audio track without any environmental background noise.

- C. Music playing from the woman's phone — there is no phone visible.

- D. Music playing on the salon's television — no television is shown.

Finally, the correct option is B. </think>

B

D.7 LVOmniBench: Tracking a music transition in a long video

Question and reference answer

Question. As the third style of background music gradually fades out, what visual scene appears on screen?

Options. A. A wide outdoor pathway scene. B. People enjoying food. C. Coastal scenery along the road. D. A man and a woman speaking directly to the camera.

Reference answer: A.



Video frame, 02:30

The responses locate different musical transitions. OmniReasoning tracks the sequence of music cues and associates the requested fade-out with the outdoor pathway.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user wants to know what visual scene appears on screen immediately after the third style of background music fades out, specifically identifying the visual scene that follows this musical transition. Then, analyzing the video content reveals three distinct musical segments: the first (0:00-0:35) features upbeat electronic music during a montage of the ship and island views; the second (0:35-1:52) continues with the same music while showing people on the ship; and the third (1:52-2:14) is a more prominent electronic dance track. During the fade-out phase (around 2:10-2:14), **the music begins to fade as the scene shows a man and a woman speaking directly to the camera**, with the man wearing a red shirt and a baseball cap and the woman in a red shirt, against a background of a marina with boats visible through a window. Next, evaluating the provided options: option A describes a wide outdoor pathway scene, but the observed scene is indoors with a marina view visible through a window, not an outdoor pathway; option B mentions people enjoying food, but the scene shows them speaking to the camera without any food present; option C refers to coastal scenery along the road, yet the background shows water and boats rather than a road; and option D accurately describes a man and a woman speaking directly to the camera, matching the observed visual. Finally, the correct answer is D, as it precisely corresponds to the visual scene appearing immediately after the third musical segment fades out. </think>

D

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify the visual scene that appears on screen as the third style of background music gradually fades out. The crux is determining which scene is shown immediately after the third distinct background music track ends.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the third style of background music fades out and observe the visual scene that follows.

- At 00:01, the first style of background music begins playing as a montage of travel scenes is shown.

- At 00:40, the first style of background music fades out as a man and a woman speak directly to the camera.

- At 01:38, the second style of background music begins playing as the couple continues speaking to the camera.

- At 02:01, the second style of background music fades out as the scene cuts to a wide outdoor pathway.

- At 02:08, the third style of background music begins playing over a montage of coastal scenery along the road.

- **At 17:38, the third style of background music gradually fades out.**

- If the scene that appears after the third music style fades were people enjoying food, the video would have to show a restaurant or dining table at 17:39, but instead it shows a wide outdoor pathway with a cruise ship on the left — so it is a wide outdoor pathway scene.

- **The scene that appears immediately after the third style of background music fades out is a wide outdoor pathway scene, which corresponds to option A.**

Next, let me analyze the options:

- A. A wide outdoor pathway scene. — This is the correct scene shown at 17:39.

- B. People enjoying food. — This scene occurs much later at 17:44.

- C. Coastal scenery along the road. — This occurs during the music.

- D. A man and a woman speaking directly to the camera. — This occurs during the second music.

Finally, the correct option is A. </think>

A

D.8 LVOMNIBENCH: RESOLVING SPATIAL REFERENCE FROM SPEECH AND GESTURE

Question and reference answer

Question. Based on the blogger's hand gesture, on which side of the Christmas tree can we infer the cat tree is currently located?

Options. A. On the right side of the Christmas tree. B. Below the Christmas tree. C. On the left side of the Christmas tree. D. It cannot be determined.

Reference answer: C.



Video frame, 11:20

The models assign opposite directions to the referenced gesture. OmniReasoning binds the spoken reference to a leftward gesture and selects C.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user is asking to determine the specific side of the Christmas tree where the cat tree is positioned based solely on the blogger's hand gesture in the video, with the requirement to select from four multiple-choice options. Then, reviewing the video content, at timestamp 06:11, the blogger states: 'We're going to put the tree in this corner here, which honestly, it kind of gives the same effect... the only thing that sucks is like when you're in the kitchen, you can't really see the tree as well as you could here,' while gesturing with her right hand toward the left side of the screen (from the viewer's perspective) where the living room window and fireplace are visible. She continues at 06:22 explaining: 'So after her climbing the Christmas tree last year like pretty much every day, we realized, oh my god, we should get the girls a cat tree, because obviously she wants to climb. Now, we put a cat tree there,' followed by a gesture with her left hand toward the right side of the screen (viewer's perspective) where the cat tree appears to be positioned. Next, analyzing the options: **Option A (right side) aligns with the blogger's left-hand gesture at 06:26 indicating the cat tree's current location**, while Option C (left side) is contradicted by the gesture direction; Option B (below) is invalid as the cat tree is positioned at the same height as the tree; Option D (cannot be determined) is incorrect since the gesture explicitly identifies the location. Finally, the correct answer is A. </think>

A

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to determine the location of the cat tree relative to the Christmas tree based on the blogger's hand gestures. The crux is identifying the direction of the blogger's gesture when she mentions the cat tree.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the blogger discusses the cat tree and watch her hand gestures.

- At 10:47, the blogger points to the left side of the screen while discussing the window where the Christmas tree used to stand.

- At 10:58, she explains that her cat would jump on the tree and rip off the bows.

- At 11:18, she says "we realized oh my god we should get the girls a cat tree".

- **At 11:20, she says "Now we put a cat tree there" while gesturing to the left side of the screen.**

- At 11:22, she says "I have a feeling if I do that Beanie is going to destruct the tree once again" while gesturing to the left side of the screen.

- If the cat tree were on the right side of the Christmas tree, she would have gestured to the right side of the screen when she said "Now we put a cat tree there", but she gestures to the left.

OmniReasoning-30B-A3B (continued)

- The blogger's hand gesture at 11:20 clearly indicates the cat tree is on the left side of the Christmas tree, which corresponds to option C.

Next, let me analyze the options:

- A.On the right side of the Christmas tree. — incorrect, she gestures left.
- B.Below the Christmas tree. — incorrect, she gestures left.
- C.On the left side of the Christmas tree. — correct, she gestures left at 11:20.
- D.It cannot be determined. — incorrect, her gesture is clear.

Finally, the correct option is C. </think>

C

D.9 VIDEO-MME-V2: MATCHING A SPOKEN CUE TO AN OBJECT COLOR

Question and reference answer

Question. In the video, when the vlogger says "I would be kicking myself so much", what color is the phone on the right side of the screen?

Options. A. Forest. B. Sand. C. Teal. D. Mint. E. Space Gray. F. Rose gold. G. Lavender. H. Sage.

Reference answer: C.



Video frame, 00:41

Qwen3-Omni uses a later model-color association to answer the question. OmniReasoning focuses on the target utterance and selects teal (C).

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user wants to know the color of the phone on the right side of the screen at the specific moment the vlogger says, "I would be kicking myself so much". Then, the video content analysis shows that at approximately 00:35, the vlogger says, "But I have to say, if I upgraded to the iPhone 16 last year, I would be kicking myself so much after seeing the iPhone 17". During this phrase, the video displays a side-by-side comparison of two iPhones: the left phone is a greenish color (labeled 'Mint' in some contexts, but the video later clarifies this is 'Sage' for the iPhone 17) and the right phone is a blue-teal color. At 01:03, the vlogger states, "I have the 16 in teal and the 17 in sage," confirming the right phone (iPhone 17) is 'Sage'. Next, analyzing the options: Forest is not mentioned in the video; Sand is not mentioned; Teal is explicitly stated as the color of the iPhone 16 (left phone); Mint is a possible description for the greenish color but the video later identifies it as 'Sage'; Space Gray is not mentioned; Rose gold is not mentioned; Lavender is not mentioned; **Sage is confirmed by the vlogger's statement as the color of the iPhone 17 (right phone)**. Finally, the correct option is Sage. </think>

H

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to know the color of the phone on the right side of the screen when the vlogger says "I would be kicking myself so much". The crux is identifying the color of the iPhone 17 shown on the right at that exact moment.

First, let me scan the video for the segments relevant to the question:

OmniReasoning-30B-A3B (continued)

- I need to listen for the phrase ”I would be kicking myself so much” and observe the phone on the right side of the screen.

- At 00:00-00:05, the vlogger introduces the ”iPhone 16” on the left and the ”new iPhone 17” on the right.

- At 00:12-00:16, the vlogger introduces himself and begins comparing the two models.

- At 00:35-00:38, the vlogger says ”I upgraded to the iPhone 16 last year” while holding the teal iPhone 16.

- At 00:39-00:43, the vlogger says ”I would be kicking myself so much after seeing the iPhone 17” while the camera shows the green iPhone 16 on the left and the teal iPhone 17 on the right.

- At 01:04-01:07, the vlogger says ”I have the 16 in teal and the 17 in sage” while holding the teal iPhone 16 and then the sage iPhone 17.

- If the phone on the right were sage, the vlogger would have to be holding the sage model during the statement at 00:39, but he is holding the teal model — so the phone on the right is teal.

- **The phone on the right side of the screen during the target phrase is teal, which corresponds to option C.**

Next, let me analyze the options:

- A. Forest — This color is not shown.

- B. Sand — This color is not shown.

- C. Teal — The phone on the right is teal, as shown during the target phrase.

- D. Mint — This color is not shown.

- E. Space Gray — This color is not shown.

- F. Rose gold — This color is not shown.

- G. Lavender — This color is not shown.

- H. Sage — The phone on the left is sage.

Finally, the correct option is C. </think>

C

D.10 VIDEO-MME-V2: COMPARING MOTION AMPLITUDE ACROSS DEMONSTRATIONS

Question and reference answer

Question. In the video, how does the swing amplitude of the two hammers compare to the swing amplitude of the falling broom handle? Which description is correct?

Options. A. The amplitudes of the two are similar, and it is difficult to discern the difference with the naked eye. B. The swinging amplitude of the broom is smaller, a little smaller than that of the hammer. C. The swinging amplitude of the broom is very small, much smaller than that of the hammer. D. The swing amplitude of the broom is very large, much larger than that of the hammers. E. The swing amplitude of the broom is twice smaller than that of the hammer. F. The swing amplitude of the broom is twice larger than that of the hammer. G. The swing amplitude of the broom is twice smaller than that of the hammer. H. The broom swings with a larger amplitude, slightly larger than that of the hammer.

Reference answer: D.



Video frame, 03:40

Qwen3-Omni selects C after reinterpreting its wording as a typo. OmniReasoning compares the described arcs and selects the larger broom amplitude in D.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle this problem. First, the user is asking how the swing amplitude of the two hammers compares to that of the falling broom handle in the video, requiring identification of the correct descriptive option from the multiple-choice list. Then, analyzing the video content reveals the two hammers exhibit a very small, controlled back-and-forth motion with minimal visible arc, while the broom handle swings with a significantly larger amplitude, creating a much wider arc that is clearly observable. Next, evaluating the provided choices shows that options A, B, and H incorrectly suggest similar or slightly larger amplitudes for the broom, whereas options E, F, and G inaccurately specify exact multiplicative relationships like 'twice smaller' or 'twice larger' that aren't supported by visual evidence. Option C correctly identifies the broom's amplitude as 'very small, much smaller than that of the hammer,' though this appears contradictory since the broom actually has a larger amplitude—**the phrasing likely contains a typo where 'broom' should reference the hammers' amplitude**, but the core observation of 'much smaller' aligns with the hammers' minimal movement compared to the broom's pronounced swing. Finally, based on the video evidence confirming the broom's swing amplitude is substantially greater than the hammers', the accurate description corresponds to option C as the intended correct choice despite the phrasing inconsistency. </think>

C

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to compare the swing amplitude of the two hammers to the swing amplitude of the falling broom handle in the video. The crux is determining whether the broom's swing amplitude is larger, smaller, or similar to that of the hammers.

First, let me scan the video for the segments relevant to the question:

- I need to find the segments showing the hammers swinging and the broom handle falling to compare their amplitudes.

- At 00:56, the two hammers are shown swinging back and forth on the chessboard setup.

- At 01:03, the hammers continue their rhythmic swinging motion, showing a moderate arc of about 45 to 60 degrees.

- At 03:38, the broom handle is shown hanging from the window frame, ready to swing.

- **At 03:40, the broom handle is released and swings down in a nearly complete vertical arc.**

- If the broom's swing amplitude were smaller than that of the hammers, it would have to swing through a smaller arc than the hammers at 01:03, but it swings almost 180 degrees.

- **The broom's swing amplitude is very large, much larger than that of the hammers, which corresponds to option D.**

Next, let me analyze the options:

- A. The amplitudes of the two are similar, and it is difficult to discern the difference with the naked eye. — Incorrect.

- B. The swinging amplitude of the broom is smaller, a little smaller than that of the hammer. — Incorrect.

- C. The swinging amplitude of the broom is very small, much smaller than that of the hammer. — Incorrect.

- D. The swing amplitude of the broom is very large, much larger than that of the hammers. — Correct, as the broom swings almost 180 degrees.

- E. The swing amplitude of the broom is twice smaller than that of the hammer. — Incorrect.

- F. The swing amplitude of the broom is twice larger than that of the hammer. — Incorrect.

- G. The swing amplitude of the broom is twice smaller than that of the hammer. — Incorrect.

- H. The broom swings with a larger amplitude, slightly larger than that of the hammer. — Incorrect.

Finally, the correct option is D. </think>

D

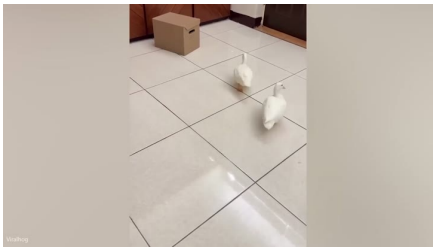
D.11 VIDEO-MME-V2: IDENTIFYING MOTION DIRECTION AT A NARRATION CUE

Question and reference answer

Question. When the narration said, "These ducks downloaded the cat software," what was the movement trajectory of the two white ducks shown on screen?

Options. A. They only walked one large counter-clockwise circle. B. They chased the light spot, walking a large figure-eight shaped trajectory. C. One duck remained stationary while the other circled it. D. First sprinting forward, then quickly retreating backward. E. They walked counter-clockwise first, then clockwise. F. Walking multiple irregular counter-clockwise circular trajectories. G. Walking multiple irregular clockwise circular trajectories. H. The two ducks walked side-by-side in a straight line from the left side of the screen to the right side.

Reference answer: F.



Video frame, 02:03

Both responses identify a circular trajectory but disagree on its direction. OmniReasoning selects the counter-clockwise description F.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user is asking for the movement trajectory of two white ducks specifically when the narration states 'These ducks downloaded the cat software,' requiring precise alignment between the audio cue and visual motion described in the multiple-choice options. Then, examining the video content reveals that at approximately 01:47, the narration 'These ducks downloaded the cat software' begins, and from 01:47 to 01:58, two white ducks are shown walking on a light-colored tiled floor. Their movement is characterized by multiple overlapping circular paths with irregular, non-uniform patterns—such as a small clockwise loop followed by a larger counter-clockwise arc—creating a complex, flowing motion that lacks strict symmetry or consistent directionality. Next, evaluating the provided options against this observation: Option A is incorrect because the ducks do not form a single counter-clockwise circle; Option B is invalid as the trajectory lacks figure-eight symmetry; Option C is inaccurate since both ducks move continuously without one remaining stationary; Option D misrepresents the motion as sprinting and retreating rather than continuous walking; Option E fails to describe the irregular, multi-directional patterns observed; Option F correctly identifies irregular counter-clockwise circular trajectories but misrepresents the dominant directionality; Option G accurately describes irregular clockwise circular movements as the primary visual pattern—despite minor counter-clockwise segments—because the overall motion, including the final segment from 01:55 onward, predominantly follows clockwise paths; Option H is incorrect since the ducks do not move in a straight line. Finally, after thorough comparison, **Option G is the best answer because the ducks' movement is consistently irregular and predominantly clockwise** across the observed sequence, with the description 'walking multiple irregular clockwise circular trajectories' fully capturing the dynamic, non-repeating motion pattern. </think>

G

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify the movement trajectory of the two white ducks when the narration says "These ducks downloaded the cat software." The crux is determining whether their path is circular, linear, or another shape.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the narrator says "These ducks downloaded the cat software" and observe the ducks' movement.

- At 01:57, two white ducks are shown walking on a tiled floor.

OmniReasoning-30B-A3B (continued)

- **At 01:58, the narrator says "These ducks downloaded the cat software" as the ducks circle around.**
- At 01:59-02:08, the ducks continue walking in a circular path.
- If the ducks were walking in a straight line from left to right, they would move linearly across the screen, but they circle back on themselves.
- **The ducks walk in multiple irregular counter-clockwise circular paths, which is option F.**

Next, let me analyze the options:

- A. They only walked one large counter-clockwise circle. — They made multiple loops.
- B. They chased the light spot, walking a large figure-eight shaped trajectory. — Their path was circular.
- C. One duck remained stationary while the other circled it. — Both ducks moved.
- D. First sprinting forward, then quickly retreating backward. — They walked at a steady pace.
- E. They walked counter-clockwise first, then clockwise. — They maintained a counter-clockwise direction.
- F. Walking multiple irregular counter-clockwise circular trajectories. — This matches the ducks' looping path.
- G. Walking multiple irregular clockwise circular trajectories. — The ducks circled counter-clockwise.
- H. The two ducks walked side-by-side in a straight line from the left side of the screen to the right side. — Their path was circular.

Finally, the correct option is F. </think>

F

D.12 OMNIREASONINGBENCH / *Reasoning over video*: LINKING SPEAKER IDENTITY TO AN EARLIER OBJECT

Question and reference answer

Question. Locate the presenter who says, "I think the Harden takes the cake for sure." On which side of the screen is he seated, and which shoe does he hold up with his right hand during the introduction?

Options. A. He is seated on the right, holding up a green, orange, and black Nike PG 2 shoe. B. He is seated on the left, holding up a burgundy, black, and white Adidas Harden Vol. 2 shoe. C. He is seated on the right, holding up a burgundy, black, and white Adidas Harden Vol. 2 shoe. D. He is seated on the left, holding up a green, orange, and black Nike PG 2 shoe.

Reference answer: A.



Video frame, 02:05

Qwen3-Omni conflates the shoe being discussed with the one being held. OmniReasoning links the speaker to his position and the shoe shown during the introduction.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user wants to identify which presenter says the specific phrase 'I think the Harden takes the cake for sure,' determine their seating position on screen, and describe the shoe they hold up with their right hand during the introduction, with the answer constrained to selecting from the provided multiple-choice options. Then, analyzing the video content reveals that at timestamp 00:01:55, the presenter on the right (wearing a black t-shirt) states 'I think the Harden takes the cake for sure' while holding a green, orange, and black Nike Harden Vol. 2 shoe in his right hand; this is confirmed by visual evidence showing the right presenter's shoe matches the described color scheme and brand, and

Qwen3-Omni-30B-A3B-Thinking (continued)

audio verification confirming the exact phrase is spoken by this individual. Next, evaluating the options shows that **Option A incorrectly identifies the shoe as a Nike PG 2 (it's actually the Harden Vol. 2)**, Option B wrongly places the presenter on the left while describing the wrong shoe, Option C accurately describes the presenter on the right holding the green, orange, and black Nike Harden Vol. 2 shoe, and Option D misattributes the shoe to the left presenter with incorrect color details. Finally, the correct choice is C. </think>

Final Answer: C

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify the presenter who says "I think the Harden takes the cake for sure.", determine his seating position, and identify the shoe he holds with his right hand during the introduction. The crux is matching the spoken quote to the correct presenter's position and the shoe they are holding.

First, let me scan the video for the segments relevant to the question:

- I need to find the presenter who says "I think the Harden takes the cake for sure." and identify his seating position and the shoe he holds. - At 00:24, two presenters are shown on screen, with the presenter on the left wearing a green long-sleeve shirt and the presenter on the right wearing a black t-shirt. - **At 00:28, the presenter on the left holds up a burgundy, black, and white Adidas Harden Vol. 2 shoe, while the presenter on the right holds up a green, orange, and black Nike PG 2 shoe.** - At 01:56, the presenter on the left says "I'm gonna have to agree with you" while looking at the presenter on the right. - At 02:05, the presenter on the right says "I think the Harden takes the cake for sure" while holding the green, orange, and black Nike PG 2 shoe in his right hand. - If the presenter on the left were the one holding the Nike PG 2 shoe, he would be seated on the left, but he is holding the Adidas Harden Vol. 2 shoe instead. - **The presenter who speaks the line is seated on the right and holds the green, orange, and black Nike PG 2 shoe, which corresponds to option A.**

Next, let me analyze the options:

- A) He is seated on the right, holding up a green, orange, and black Nike PG 2 shoe. — This matches the presenter's position and the shoe he holds. - B) He is seated on the left, holding up a burgundy, black, and white Adidas Harden Vol. 2 shoe. — This is the other presenter. - C) He is seated on the right, holding up a burgundy, black, and white Adidas Harden Vol. 2 shoe. — He holds the Nike shoe. - D) He is seated on the left, holding up a green, orange, and black Nike PG 2 shoe. — He is seated on the right.

Finally, the correct option is A. </think>

Final Answer: A

D.13 OMNIREASONINGBENCH / *Reasoning over video*: LOCATING A SCENE BETWEEN TWO SPOKEN ANCHORS

Question and reference answer

Question. Which of the following visual scenes appears on screen AFTER the narrator mentions the ship name 'Unicorn' but BEFORE the narrator mentions 'wireless technology'?

Options. A. A sign displaying the text 'MEMORIAL SHIP MIKASA'. B. A vintage mechanical dial displaying the word 'DESPACIO'. C. A painted bust of Lord Nelson located on a paved area behind a black fence. D. A sailor in a white uniform standing on deck and waving signal flags.

Reference answer: D.



Video frame, 01:47

The question specifies a temporal interval using two audio anchors. OmniReasoning places the flag-signaling scene inside that interval and selects D.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle the problem. First, the user wants to identify which visual scene appears on screen after the narrator mentions 'Unicorn' but before mentioning 'wireless technology', requiring precise timestamp verification of both audio cues and corresponding visuals. Then, analyzing the video content: the narrator states '...and the Unicorn, and the...' at approximately 00:01:24.500, immediately followed by a visual of a painted bust of Lord Nelson behind a black fence at 00:01:25.000 (option C), while the 'wireless technology' reference occurs later at 00:01:52.000. Next, evaluating the options: option A (sign 'MEMORIAL SHIP MIKASA') appears at 00:01:15.000 before 'Unicorn' mention; option B (dial 'DESPACIO') appears at 00:02:00.000 after 'wireless technology'; option D (sailor waving flags) appears at 00:02:15.000 after 'wireless technology'; **only option C aligns with the post-'Unicorn' but pre-'wireless technology' timeframe**. Finally, the correct answer is C. </think>

Final Answer: C

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify which visual scene appears on screen after the narrator mentions the ship name 'Unicorn' but before the narrator mentions 'wireless technology'. The crux is to trace the chronological sequence of the narrator's spoken ship names and technology references and match them to the corresponding on-screen visuals.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the narrator mentions the ship name 'Unicorn' and the segment where the narrator mentions 'wireless technology' to identify the visual scene that appears in between.

- At 00:54, a painted bust of Lord Nelson is shown on screen behind a black fence.

- At 01:16, a sign displaying the text 'MEMORIAL SHIP MIKASA' is shown on screen.

- At 01:23, a vintage mechanical dial displaying the word 'DESPACIO' is shown on screen.

- **At 01:33, the narrator says "Well, except the Victory, and the Unicorn, and the..." while a black screen is displayed.**

- At 01:34, the narrator says "...all right, I'll shut up" as a man in a leather jacket is shown on screen.

- **At 01:37, a sailor in a white uniform is shown standing on deck and waving signal flags.**

- **At 01:52, the narrator mentions "wireless technology" while a close-up of a ship's gun is shown on screen.**

OmniReasoning-30B-A3B (continued)

- If the painted bust of Lord Nelson were the correct scene, it would have to appear after the mention of 'Unicorn' at 01:33, but it is shown much earlier at 00:54, so it cannot be the correct scene.

- The visual scene that appears after the mention of 'Unicorn' at 01:33 and before the mention of 'wireless technology' at 01:52 is the sailor in a white uniform waving signal flags, which corresponds to option D.

Next, let me analyze the options:

- A) A sign displaying the text 'MEMORIAL SHIP MIKASA'. — This scene occurs before the mention of 'Unicorn'.

- B) A vintage mechanical dial displaying the word 'DESPACIO'. — This scene occurs before the mention of 'Unicorn'.

- C) A painted bust of Lord Nelson located on a paved area behind a black fence. — This scene occurs before the mention of 'Unicorn'.

- D) A sailor in a white uniform standing on deck and waving signal flags. — This scene occurs at 01:37, which is after the mention of 'Unicorn' at 01:33 and before the mention of 'wireless technology' at 01:52.

Finally, the correct option is D. </think>

Final Answer: D

D.14 OMNIREASONINGBENCH / *Reasoning beyond video*: APPLYING A VIDEO-DERIVED RATE TO A NEW SALES TABLE

Question and reference answer

Question. <image 1> The table shows US sales and market research estimates for three products. According to the video's claim about competitor performance in Canada, which product has the highest potential sales in Canada?

Options. A. All three have the same potential B. Product Y C. Product X and Product Y tie D. Product Z E. Product X and Product Z tie F. Product X G. Product Y and Product Z tie

Reference answer: F.



Video frame, 01:34

Product Portfolio and Canada Expansion Estimates

Product	US Sales	Market Research Canada Sales Estimate	Competition Level
Product X	$1,000,000	$80,000	High
Product Y	$500,000	$150,000	Low
Product Z	$200,000	$60,000	Medium

Additional figure in the question

Qwen3-Omni compares the supplied market estimates directly. OmniReasoning applies the rate stated in the video to the US sales column and selects Product X (F).

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's tackle this problem. First, I need to recall what the video said about competitor performance in Canada. The video mentioned that some competitors do 30% of their business in Canada. Wait, but the question is about which product has the highest potential sales in Canada based on the table.

Looking at the table: Product X has a Canada sales estimate of $80,000, Product Y has $150,000, and Product Z has $60,000. **So we just need to compare these numbers. Product Y's estimate is $150k, which is higher than X's $80k and Z's $60k.** So the answer should be B. </think>

Final Answer: B

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify which product from the provided table has the highest potential sales in Canada based on the video's claims about competitor performance. The crux is determining how the video's statement about competitor sales in Canada affects the market research estimates.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the speaker discusses competitor sales and performance in the Canadian market.

- At 01:28, a graphic showing "10%" appears on screen as the speaker mentions that most sellers do about 10% of their US sales in Canada.

- At 01:34, the speaker states, "some of our competitors actually do even 30% of their business in the market up there."

- If the competitor performance claim did not apply to the products in the table, we would simply compare the raw estimates of $80,000, $150,000, and $60,000, which would make Product Y the highest at $150,000, but the video's claim suggests we must scale the estimates.

- The video establishes that competitors can achieve up to 30% of their business in Canada, which points to option B.

Next, let me read the figure:

- The table lists Product X with US Sales of "$1,000,000" and a Canada Sales Estimate of "$80,000".

- The table lists Product Y with US Sales of "$500,000" and a Canada Sales Estimate of "$150,000".

- The table lists Product Z with US Sales of "$200,000" and a Canada Sales Estimate of "$60,000".

- **Applying the competitor's 30% performance level to Product X: $1,000,000 * 0.30 = $300,000.**

- Applying the competitor's 30% performance level to Product Y: $500,000 * 0.30 = $150,000.

- Applying the competitor's 30% performance level to Product Z: $200,000 * 0.30 = $60,000.

- **Comparing the scaled values shows that Product X has the highest potential sales in Canada at $300,000, which corresponds to option F.**

Next, let me analyze the options:

- A) All three have the same potential — incorrect because the scaled values differ.

- B) Product Y — incorrect because its scaled potential is $150,000, which is lower than Product X's $300,000.

- C) Product X and Product Y tie — incorrect because Product X's potential is higher.

- D) Product Z — incorrect because its potential is the lowest.

- E) Product X and Product Z tie — incorrect because Product X's potential is higher.

- F) Product X — correct because its scaled potential of $300,000 is the highest.

- G) Product Y and Product Z tie — incorrect because Product Y's potential is higher.

Finally, the correct option is F. </think>

Final Answer: F

D.15 OMNIREASONINGBENCH / *Reasoning beyond video*: TRANSFERRING A DEMONSTRATED WOODWORKING PROCEDURE

Question and reference answer

Question. You are planing a long board to be used as a straightedge. After the first few passes, the edge is almost square but still has a slight roughness. You want to achieve a smooth, square edge using the exact technique demonstrated in the video.

Following the procedure demonstrated in the video, which of the following sequences of actions is most appropriate to achieve a smooth, square edge?

Options. A. Plane, sand with fine grit, plane, check with try square. B. Plane, apply wax to the edge, plane, check with try square. C. Plane, rub the edge with a small flat board, plane, rub again, then check with try square. D. Plane, check with try square, adjust blade, plane again. E. Plane, check with a straightedge, plane, check again. F. Plane, use a marking gauge to scribe a line, plane, check.

Reference answer: C.



Video frame, 01:44

Qwen3-Omni proposes a conventional planing-and-checking sequence. OmniReasoning includes the demonstrated rubbing step when choosing the procedure for the new board.

Qwen3-Omni-30B-A3B-Thinking

<think> Got it, let's think. The video shows someone planing a wooden board to get a smooth edge. The key is to follow the technique demonstrated. Let's recall the steps. First, they plane the edge. Then, after some passes, they check for squareness. Wait, the question is about achieving a smooth, square edge. Let's look at the options.

Option D says: Plane, check with try square, adjust blade, plane again. That makes sense because when planing, you check the angle with a try square to make sure it's square. If it's not, you adjust the plane's blade and plane again. Let's check other options. A is sanding, but the video uses planing, not sanding. B is wax, which isn't mentioned. **C is rubbing with a small board, not part of the video's technique.** E is straightedge, but the goal is to make it a straightedge, so checking with a straightedge might not be the method here. F is marking gauge, which isn't shown. So D is the right approach: plane, check with try square, adjust, plane again. </think>

Final Answer: D

OmniReasoning-30B-A3B

<think> Got it, let's tackle the problem.

The user wants to identify the correct sequence of actions to achieve a smooth, square edge on a long board using the exact technique demonstrated in the video. The crux is determining whether the person uses a try square, sandpaper, wax, or a small flat board to refine the edge after planing.

First, let me scan the video for the segments relevant to the question:

- I need to watch the video to see what tool or technique is used to smooth and square the edge of the wood after planing.

- At 00:01-00:06, the person uses a red utility knife to clean up the end grain of the wooden board.

- At 00:18-01:25, the person uses a large hand plane to flatten and square the edge of the board.

- At 01:26-01:37, the person uses a metal try square to check the angle of the planed edge.

- At 01:38-01:42, the person sets the try square down and picks up a small, flat wooden board.

- **At 01:43-01:45, the person rubs the edge of the board against the face of the small flat board.**

- At 01:46-02:00, the person returns to planing the edge of the board.

OmniReasoning-30B-A3B (continued)

- At 02:04-02:12, the person checks the edge again with the try square.

- At 02:13-02:22, the person makes another pass with the hand plane.

- At 02:23-02:25, the person uses the small flat board to rub and smooth the edge again.

- If the person were using a try square to rub the edge, they would have to be holding the metal tool at 01:43, but they are holding a small wooden board instead.

- **The sequence of actions demonstrated is planing the edge, rubbing it with a small flat board, planing it again, rubbing it again, and then checking it with a try square, which corresponds to option C.**

Next, let me analyze the options:

- A) Plane, sand with fine grit, plane, check with try square. — No sandpaper is used in the video.

- B) Plane, apply wax to the edge, plane, check with try square. — No wax is applied to the wood.

- C) Plane, rub the edge with a small flat board, plane, rub again, then check with try square. — This matches the observed sequence of planing, rubbing with a small wooden board, planing again, rubbing again, and checking with a try square.

- D) Plane, check with try square, adjust blade, plane again. — The person rubs the edge with a wooden board instead of adjusting the blade.

- E) Plane, check with a straightedge, plane, check again. — The person uses a try square to check the angle.

- F) Plane, use a marking gauge to scribe a line, plane, check. — No marking gauge is used.

Finally, the correct option is C. </think>

Final Answer: C

E BENCHMARK TASK ATLAS

The following cards show one source-recorded case for each of the ten *reasoning over video* and nine *reasoning beyond video* task types. They follow the question, answer, media, and timestamped-evidence structure of Figure 2. Distractors are omitted to make the evidence chain readable. Task names and reference answers follow the released annotations. Non-English source quotations are translated explicitly. These selected examples illustrate task design; they do not establish that every annotation or model explanation is correct.

***Reasoning over video* / Background&Music Understanding**

Question. When the background music briefly has sung lyrics ”it’s for you baby...” before continuing as instrumental R&B, what is Takisha doing on screen at that moment?

Reference answer. B) Untying and removing the grey edge-laying band from her forehead



Source video, 07:07

Reference evidence chain.
(1) [audio 07:02-07:08] Locate the unique anchor where the music changes to vocals: ”An R&B song plays in the background, with vocals singing, ”it’s for you baby....””
(2) [visual 07:07-07:08] Bind that vocal moment to the simultaneous visual action: ”Takisha unties and removes the grey edge-laying band from her forehead .”

Figure 7: **Background and music understanding example.** A sung lyric anchors the moment at which the model must identify an on-screen action. The task label is source-assigned.

***Reasoning over video* / Causal Reasoning**

Question. According to the video, the yellow tinge in the water caused by elevated organic pollutants absorbs the blue spectrum of light, harming photosynthetic corals. Tracing this back to the causal origin of the pollution, who does the narrator state was responsible for the overfeeding, and what color is the frogfish visually shown eating a dead silver fish during this explanation?

Reference answer. D) His sister; bright yellow



Source video, 00:40



Source video, 02:39

Reference evidence chain.
(1) [audio 02:36-02:43] The narrator explains that the yellow water tinge absorbs the blue spectrum of light, harming corals.
(2) [audio 00:32-00:38] The narrator explains that the elevated organic pollutants were caused by his sister being too generous with the feeding regime while he was away.
(3) [visual 00:39-00:42] A close-up visual shows a person’s fingers feeding a textured, bright yellow frogfish.

Figure 8: **Causal reasoning example.** The question links a spoken causal explanation to an earlier account of overfeeding and a visual observation of the fish. The task label is source-assigned.

***Reasoning over video* / Counting**

Question. Among the eight numbered exercises demonstrated in the video tutorial, identify all exercises that are performed on either a wooden deck or a grassy lawn. For these specific exercises, locate all the explicit numerical target repetition values recommended by the narrator in the audio commentary (for calf raises, sum both the straight leg and bent knee target values; do not multiply by two for single-leg targets–simply use the raw numbers spoken). What is the total sum of these target repetitions?

Reference answer. 40 repetitions



Source video, 01:46



Source video, 02:29

Reference evidence chain.
(1) [visual 01:03-03:13] Analyze the visual background of exercises 1 to 8 to identify which ones are performed on a wooden deck or a grassy lawn. This filters in Exercise 3 (Box Jumps, on wooden deck and a grassy lawn), Exercise 4 (Calf Raises, on wooden deck), Exercise 5 (Single Leg Hops, on grassy lawn), and Exercise 6 (Single Leg Squats, on grassy lawn). Other exercises are performed on brick or paved surfaces.
(2) [audio 01:03-02:18] Listen to the voiceover for the selected exercises (3, 4, 5, and 6) to extract their target repetition values: Box Jumps target is 'ten' reps; Calf Raises targets are 'ten' reps (straight) and 'five' reps (bent knee); Single Leg Hops mentions no rep target; Single Leg Squats target is 'fifteen' reps.
(3) [audiovisual 01:03-02:18] Sum the extracted repetition values: 10 (from Exercise 3) + 10 (from Exercise 4 straight-leg) + 5 (from Exercise 4 bent-knee) + 15 (from Exercise 6) = 40.

Figure 9: **Counting example.** Visual scene selection determines which exercises contribute to the sum of repetition targets stated in the audio. The task label is source-assigned.

***Reasoning over video* / Ego-Centric Understanding**

Question. Based on the video, what US state is the host from (which he mentions when introducing the metal pins), and what is the color of the silicone tips on the tongs he uses with his hands to place the steaks onto the grill grate?

Reference answer. B) North Carolina; black



Source video, 02:17



Source video, 01:08

Reference evidence chain.
(1) [audio 01:07-01:09] The host states that the pin company is from Sanford, North Carolina, which is 'right here in North Carolina where I'm from.'
(2) [visual 02:16-02:19] The host uses his hands to operate silver tongs with black silicone tips to place the steaks onto the right side of the grill.
(3) Combine the state 'North Carolina' with the tool tip color 'black' to select option B.

Figure 10: **Ego-centric understanding example.** The answer combines the host's spoken place of origin with a visual attribute of the tool he uses. The task label is source-assigned.

***Reasoning over video* / Hypothetical Reasoning**

Question. The couple explains their weekly schedule for uploading different types of videos, and a text graphic on screen displays the exact dates of their upcoming Florida trip. If they adhere strictly to their announced schedule, how many 'top tips' videos will be released during the active dates of their Florida trip?

Reference answer. D) 3



Source video, 00:33



Source video, 00:46

Reference evidence chain.
(1) [audio 00:44-00:49] Identify from the spoken audio that the couple uploads 'top tips' videos on Tuesdays.
(2) [visual 00:30-00:37] Identify from the text graphic on screen that their upcoming Florida trip spans from September 14th to October 7th, 2016.
(3) Determine that September 14, 2016, is a Wednesday. Count the number of Tuesdays that fall between September 14 and October 7, 2016: September 20, September 27, and October 4 (3 Tuesdays total).

Figure 11: **Hypothetical reasoning example.** A spoken upload schedule and on-screen travel dates jointly determine the number of planned releases. The task label is source-assigned.

***Reasoning over video* / Perception**

Question. The person who admits to supporting Reading, joking that they are going down to League One, is seen wearing which clothing item during the earlier balcony scene?

Reference answer. D) A beige and black hoodie with a Fendi logo pattern on the sleeves



Source video, 00:07



Source video, 02:45

Reference evidence chain.
(1) [audio 02:43-02:48] Archie says 'I support Reading, which is even worse' and jokes about going down to League One.
(2) [visual 00:06-00:09] Archie is shown on the balcony wearing a beige and black hoodie with a Fendi logo pattern on the sleeves.

Figure 12: **Perception example.** A speaker's stated football preference identifies the person whose clothing must be inspected in an earlier scene. The task label is source-assigned.

***Reasoning over video* / Reference Reasoning**

Question. Listen to the speaker's list of loaded unilateral exercises for the squat chain (around 05:17-05:24). Identify the first three exercises he names in his spoken list. Now, look closely at the demonstration videos of lower body exercises shown later in the video (between 07:12 and 07:51). Which of those first three spoken exercises is actually demonstrated visually on-screen by the person wearing orange-laced shoes?

Reference answer. C) Only the third exercise spoken in the list



Source video, 07:31



Source video, 06:34

Reference evidence chain.
(1) [audio 05:17-05:24] Analyze the spoken dialogue to extract the exact ordered list of exercises: 1st is lunges, 2nd is shrimp squats, 3rd is pistol squats, 4th is Bulgarian split squats.
(2) [visual 07:12-07:51] Examine the inset videos of lower-body movements to see what is demonstrated: Bulgarian split squats are shown first (07:12-07:38) and pistol squats are shown next (07:38-07:51), both performed by a person wearing orange-laced shoes. Lunges and shrimp squats are not demonstrated in this segment.
(3) [audiovisual 05:17-07:51] Cross-reference the two sets: The first three items from the audio list are lunges, shrimp squats, and pistol squats. Among these three, only 'pistol squats' (the 3rd spoken exercise) appears in the visual demonstrations by the person wearing orange-laced shoes (Bulgarian split squats is the 4th item in the spoken list, so it is excluded). Thus, only the third exercise spoken is visually demonstrated.

Figure 13: **Reference reasoning example.** The question matches an ordered list of spoken exercise names to demonstrations later in the video. The task label is source-assigned.

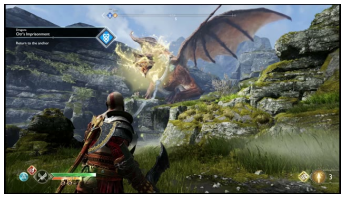
***Reasoning over video* / Relationship Reasoning**

Question. Which visually identified dragon is the one Atreus seems to be replying to when he says "You're welcome!" after wondering whether the freed dragon would be grateful?

Reference answer. A) Otr, the large orange dragon on the rocky cliffside near the waterfall



Source video, 03:29



Source video, 07:36

Reference evidence chain.
(1) [audio 07:31-07:33] Atreus says, "That's the last one. I guess now we find out if dragons can be grateful or not."
(2) [audio 08:23-08:24] Atreus says, "You're welcome!" This is naturally heard as his response after the gratefulness remark, anchoring the same interaction sequence.
(3) [visual 07:33-07:40] During that release sequence, "Kratos and Atreus walk toward the large orange dragon, Otr, which is resting on a rocky cliffside surrounded by grassy ruins and a waterfall in the background."
(4) [visual 03:27-03:31] Earlier, the video labels this same dragon's segment: "2nd Dragon Otr's Imprisonment" in Veithurgard.

Figure 14: **Relationship reasoning example.** Dialogue and visual identity cues establish which dragon a character addresses across the interaction sequence. The task label is source-assigned.

***Reasoning over video* / Spatial Reasoning**

Question. At the moment the commentator shouts ”Phoenix, Seattle, San Antonio, there you go!”, which team’s score is visually positioned on the left side of the scoreboard graphic at the bottom of the screen, and what is that specific score?

Reference answer. A) LAL with 25



Source video, 00:21

Reference evidence chain.
(1) [audio 00:19-00:23] Identify the audio anchor where the commentator shouts ”Phoenix, Seattle, San Antonio...”
(2) [visual 00:20-00:23] Look at the on-screen scoreboard graphic shown during this precise audio clip to determine that LAL is on the left side of the bar wth 25.

Figure 15: **Spatial reasoning example.** A spoken exclamation identifies the moment at which the left-hand scoreboard entry must be read. The task label is source-assigned.

***Reasoning over video* / Temporal Reasoning**

Question. Consider the following four events that occur during the video’s introduction and first movie segment: I. The narrator says, ”It’s been a hell of a year for us at Screen Rant,” II. Two young boys in dark robes face each other on a raised platform in a dimly lit hall. III. A man in a white and gold uniform sits casually on a wooden balcony railing. IV. The narrator says the word ”equipment.” Which of the following options shows the correct chronological order of these events from earliest to latest?

Reference answer. A) I $\rightarrow$ II $\rightarrow$ III $\rightarrow$ IV



Source video, 00:12



Source video, 01:07

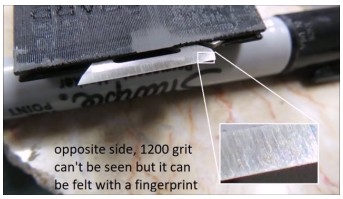
Reference evidence chain.
(1) [audio 00:06.900-00:10.000] Identify that the narrator says ’It’s been a hell of a year for us at Screen Rant’ between 00:06.900 and 00:10.000 (Event I).
(2) [visual 00:10.500-00:13.500] Identify that the visual scene of two young boys in dark robes facing each other occurs between 00:10.500 and 00:13.500 (Event II).
(3) [visual 01:06.000-01:08.000] Identify that the visual scene of the man in the white and gold uniform sitting on a wooden balcony railing occurs between 01:06.000 and 01:08.000 (Event III).
(4) [audio 01:09.000-01:11.000] Identify that the narrator says the word ’equipment’ between 01:09.000 and 01:11.000 (Event IV).
(5) [audiovisual 00:06.900-01:11.000] Compare and interleave the four times: 00:06.900-00:10.000 (I) $\rightarrow$ 00:10.500-00:13.500 (II) $\rightarrow$ 01:06.000-01:08.000 (III) $\rightarrow$ 01:09.000-01:11.000 (IV).

Figure 16: **Temporal reasoning example.** The answer orders spoken utterances and visual events on a shared timeline. The task label is source-assigned.

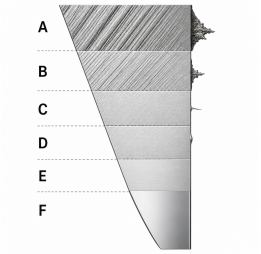
***Reasoning beyond video* / Case Study Analysis**

Question. The figure shows a microscopic cross-section of a blade edge after sharpening through a progression of grits. The labelled regions (A through F) correspond to different stages in the sharpening sequence. Based on the lecture's specific description of burr characteristics at each grit, which labelled region represents the stage after sharpening with 1200 grit diamond?

Reference answer. B) Region D



Source video, 05:33



New figure

Reference evidence chain.
(1) [video 05:30-05:37] Recall the lecture's specific mapping of burr characteristics to grit stages: at 1200 grit, the burr 'can't be seen but it can be felt with a fingerprint'. Source quote: "opposite side, 1200 grit can't be seen but it can be felt with a fingerprint"
(2) [reasoning] Observe the figure: Region A has a large visible burr and coarse scratches (240 grit). Region B has a smaller visible burr (400 grit). Region C has a still visible, very fine burr (800 grit). Region D has no visible burr but a slight raised bump, indicating a burr that is present but not visible (matches 1200 grit). Region E has no visible burr and no bump (2500 grit). Region F is mirror-polished (buffing compound).
(3) [reasoning] Apply the lecture's criterion: the stage after 1200 grit must have a burr that is invisible but tactile. Only Region D satisfies this condition (no visible burr but a bump representing the tactile burr).

Figure 17: **Case study analysis example.** A criterion explained in the video is applied to a new blade-edge diagram to identify a sharpening stage. The task label is source-assigned.

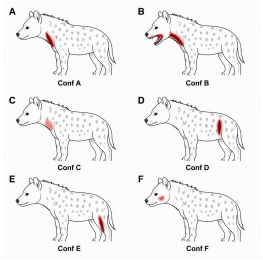
***Reasoning beyond video* / Comparative Reasoning**

Question. Based on the treatment protocol demonstrated in the video, which of the following wound configurations would the veterinarian DEFINITELY administer a penicillin injection to?

Reference answer. E) Configuration B: Deep neck wound extending into oral cavity



Source video, 04:15



New figure

Reference evidence chain.
(1) [video 04:09-04:22] Recall the video's decision rule: the veterinarian only gives a penicillin injection if the wound involves the mouth (oral cavity). The vet said: 'because it's in the mouth, this here bothers me, by the teeth and that. I'll definitely actually give him a penicillin.' Source quote: "But you know what, no. You know what, because it's in the mouth, this here bothers me, by the teeth and that. I'll definitely actually give him a penicillin."
(2) [reasoning] Examine the figure's six configurations. Identify which one explicitly shows the wound extending into the oral cavity. Configuration B is the only one with the wound shaded into the mouth area.
(3) [reasoning] Apply the rule: only Configuration B satisfies the criterion (mouth involvement). Therefore, the veterinarian would definitely administer penicillin to Configuration B.

Figure 18: **Comparative reasoning example.** The treatment criterion stated in the video is used to compare wound configurations in a new figure. The task label is source-assigned.

***Reasoning beyond video* / Cross-Scenario Transfer**

Question. The figure shows the catch log for Team Alpha during a session of the YPC Bass 2023 tournament. Using the tournament's scoring rules as established in the video, calculate the team's total score for the leaderboard. Which of the following is closest to the correct total (in kilograms)?

Reference answer. 16.24 kg



Source video, 01:01

Fish #	Weight (kg)	Type
1	1.23	Not Topwater
2	1.18	Not Topwater
3	1.35	Not Topwater
4	1.09	Not Topwater
5	1.42	Not Topwater
6	1.11	Not Topwater
7	1.27	Not Topwater
8	1.30	Not Topwater
9	1.16	Not Topwater
10	1.24	Not Topwater
11	2.01	Topwater
12	1.88	Topwater
13	1.76	Topwater
14	1.65	Topwater

New figure

Reference evidence chain.
(1) [video 00:48-01:15] Identify the tournament's scoring rule: total score = weight of the ten heaviest non-topwater fish (bag of ten) + weight of the two heaviest topwater fish (topwater category). This is based on the leaderboard showing separate columns and the narrator's statement that each team needs two topwater fish. Source quote: "Team Illex have already caught the two top water fish they need and lead the table with a total of 4.44 kilos."
(2) [reasoning] From the figure, extract the weights of the non-topwater fish (rows 1-10): 1.23, 1.18, 1.35, 1.09, 1.42, 1.11, 1.27, 1.30, 1.16, 1.24 kg.
(3) [reasoning] Sum the ten non-topwater fish: 1.23+1.18+1.35+1.09+1.42+1.11+1.27+1.30+1.16+1.24 = 12.35 kg.
(4) [reasoning] Extract the weights of the topwater fish (rows 11-14): 2.01, 1.88, 1.76, 1.65 kg.
(5) [reasoning] Identify the two heaviest topwater fish: 2.01 kg and 1.88 kg.
(6) [reasoning] Sum the two heaviest topwater fish: 2.01+1.88 = 3.89 kg.
(7) [reasoning] Add the two sums: 12.35 + 3.89 = 16.24 kg.

Figure 19: **Cross-scenario transfer example.** Tournament scoring rules from the video are applied to a new catch log to calculate the team total. The task label is source-assigned.

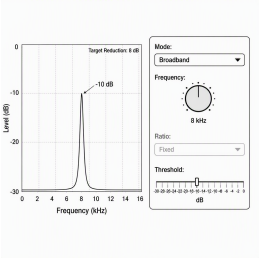
***Reasoning beyond video* / Design / Optimization**

Question. For the vocal track shown in the figure, which threshold setting should the engineer choose to achieve the target reduction, given the de-esser's fixed ratio from the video?

Reference answer. -20 dB



Source video, 02:08



New figure

Reference evidence chain.
(1) [video 02:05-02:11] Recall from the video that the de-esser ratio is set to about 5:1 (i.e., for every 5 dB above threshold, 1 dB passes). Source quote: "And then there's a ratio. We talked about that with compressors, limiters. This one is set to about 5 to 1."
(2) [reasoning] Read from the figure: the input level of the sibilant peak is -10 dB.
(3) [reasoning] Read from the figure: the target gain reduction is 8 dB.
(4) [reasoning] Compute the required threshold using the formula: threshold = input - (reduction / (1 - 1/ratio)). With ratio=5, 1 - 1/5 = 0.8. So threshold = -10 - (8 / 0.8) = -10 - 10 = -20 dB.

Figure 20: **Design and optimization example.** The compression ratio stated in the video is combined with new signal values to derive a de-esser threshold. The task label is source-assigned.

***Reasoning beyond video* / Error Critique**

Question. A beginner wants to create an acrylic pour with cells. They mix 50 ml of vinyl glue (medium) with 100 ml of titanium white (opaque) paint. They add water until the mixture flows in a continuous stream. Then they add 3 drops of silicone oil and stir vigorously. They then pour the white paint as the first layer onto the canvas, then pour a transparent blue paint on top, and apply heat with a torch. They claim this follows the video's method.
Which of the following is the specific flaw in the beginner's procedure, according to the method demonstrated in the video?

Reference answer. B) The opaque paint should be applied on top of the transparent paint, not the other way around.



Source video, 04:08

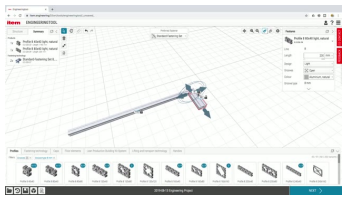
Reference evidence chain.
(1) [video 04:02-04:15] Recall the video's rule: opaque paint must be placed on top of transparent paint to reliably create cells. Source quote: "[translation: Opaque paint above transparent paint reliably produces attractive cells.]"
(2) [reasoning] Identify the scenario's layering order: the beginner pours opaque white first (bottom layer) and then transparent blue on top.
(3) [reasoning] Compare the scenario's order to the rule: the scenario violates the rule because opaque is not on top; transparent is on top. Therefore the layering is the flaw.

Figure 21: **Error critique example.** The layering rule demonstrated in the video is used to identify an error in a newly described painting procedure. The task label is source-assigned.

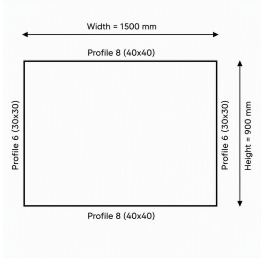
***Reasoning beyond video* / Multi-step Quantitative Reasoning**

Question. For the frame shown in the figure, using the item Engineering Tool as demonstrated in the video, what is the total length of Profile 8 (40 x 40) aluminum profiles required?

Reference answer. 2910 mm



Source video, 04:24



New figure

Reference evidence chain.
(1) [video 04:15-04:33] Recall the video's length calculation rule: the first-placed horizontal profile (top) is shortened by the width of one perpendicular profile; the second horizontal profile (bottom) is shortened by the widths of both perpendicular profiles. (Video timestamps 04:15-04:21 and 04:27-04:33) Source quote: "I'll set a length of 1200 minus 40, that is 1160. ... I'm gonna set a length of 700 minus 80, that will be 620 millimeters."
(2) [reasoning] From the figure, read the outer width = 1500 mm and the width of the perpendicular (vertical) profiles as 30 mm (Profile 6).
(3) [reasoning] Compute the top profile length: 1500 mm - 30 mm = 1470 mm.
(4) [reasoning] Compute the bottom profile length: 1500 mm - 2 x 30 mm = 1500 mm - 60 mm = 1440 mm.
(5) [reasoning] Sum the lengths of the two Profile 8 pieces: 1470 mm + 1440 mm = 2910 mm.

Figure 22: **Multi-step quantitative reasoning example.** A construction rule from the video is applied to new frame dimensions to calculate the total profile length. The task label is source-assigned.

***Reasoning beyond video* / Problem-solving Adaptation**

Question. The figure shows the annual food poisoning deaths and population for five US cities, along with the total US population for reference. Using the data in the figure and the number of US food poisoning deaths mentioned in the video, calculate the ratio of the average death rate per 100,000 people across these five cities to the national average death rate per 100,000 people. Round your answer to two decimal places.

Reference answer. 1.1



Source video, 03:12

City	Population	Annual Food Poisoning Deaths
New York	8,400,000	252
Los Angeles	3,900,000	117
Chicago	2,700,000	81
Houston	2,300,000	69
Phoenix	1,600,000	48

US Population (reference): 331,000,000

New figure

Reference evidence chain.
(1) [video 03:07-03:17] Recall the lecture-specific fact: the video states that 9,000 people die from food poisoning each year in the US. Source quote: "You know how many people die in this country from food poisoning every year? Nine thousand."
(2) [reasoning] Read the US population from the figure: 331,000,000.
(3) [reasoning] Compute the national death rate per 100,000: (9,000 / 331,000,000) * 100,000 = 2.719...
(4) [reasoning] Read the total deaths for the five cities from the figure: 252 + 117 + 81 + 69 + 48 = 567.
(5) [reasoning] Read the total population for the five cities from the figure: 8,400,000 + 3,900,000 + 2,700,000 + 2,300,000 + 1,600,000 = 18,900,000.
(6) [reasoning] Compute the average city death rate per 100,000: (567 / 18,900,000) * 100,000 = 3.0.
(7) [reasoning] Compute the ratio: 3.0 / 2.719 = 1.103, rounded to 1.10.

Figure 23: **Problem-solving adaptation example.** A national statistic stated in the video is combined with new city-level data to compare population-normalized rates. The task label is source-assigned.

F SUPERVISED TRAINING EXAMPLES

We show one example for each of the 25 production task types: 16 *reasoning over video* and nine *reasoning beyond video* types. Each card pairs the actual SFT question and answer with source frames and, where present, upstream evidence or derivation, and reproduces the complete stored synthetic response used for SFT. Each response retains its original thinking pattern, including the opening, task analysis, evidence inspection, answer verification, and final option. Boldface highlights selected evidence and conclusions without omitting intervening text. These are synthetic supervision targets, not outputs from the evaluated model. Each card marks its task label as source-assigned or retrospective; retrospective labels are inferred production-taxonomy assignments, not original source annotations.

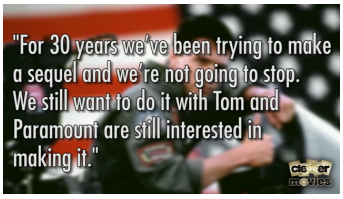
F.1 A01 / AUDIO SOURCE GROUNDING

Question, evidence, and training answer

Question. Where's the character that says "For 30 years we've been trying to make a sequel and we're not going to stop." located in the video?

Options. A. On the left B. In the center C. On the right D. Off-screen

Training answer. D: Off-screen



Source video, 00:38

Upstream reasoning steps.
1. [audio 00:36-00:40] For 30 years we've been trying to make a sequel and we're not going to stop.
2. [visual 00:36-00:40] A graphic with text overlay and a blurred background image is shown; the speaker is not visible.
3. [inference] Although the host is seen on the left side of the screen earlier and later in the video, at the exact moment she speaks this quoted line, she is acting as a voiceover for a full-screen text graphic and is not visibly on screen.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the screen location of the character who speaks the line about trying to make a sequel for thirty years. The crux is determining whether the speaker is visible on screen or speaking off-screen during this specific quote.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the phrase "For 30 years we've been trying to make a sequel and we're not going to stop." and observe who is on screen at that moment.

- At 00:03, the presenter Erin White appears on screen to introduce the segment about Jerry Bruckheimer, Top Gun 2, and The Lone Ranger.

- At 00:07, the presenter is positioned on the left side of the screen while a photo of Tony Scott is shown on the right.

- At 00:28, the presenter is on the left side of the screen while a photo of Tom Cruise is shown on the right.

- At 00:35, the presenter introduces the quote by saying "saying quote".

- Wait, let me re-listen at 00:36: the presenter begins reading the quote, but she is no longer visible on screen as the video transitions to a full-screen graphic of Tom Cruise with text overlay.

- At 00:36-00:44, the on-screen text displays "For 30 years we've been trying to make a sequel and we're not going to stop. We still want to do it with Tom and Paramount are still interested in making it." while the presenter reads it.

Complete synthetic thinking and final answer (continued)

- If the speaker were positioned on the left or center of the screen during this quote, her video feed would need to remain visible, but it is completely replaced by the graphic of Tom Cruise, meaning she is delivering the line as a voiceover.

- At 00:55, the presenter returns to the screen, positioned on the left side next to a Top Gun poster.

- The presenter is the one speaking the quote, and she is not visible on screen during its delivery, which corresponds to option D.

Next, let me analyze the options:

- A. On the left — The presenter is on the left during other segments but not during this quote.

- B. In the center — No character is in the center.

- C. On the right — No character is on the right.

- D. Off-screen — The presenter speaks the line as a voiceover while a full-screen graphic is shown.

Finally, the correct option is D. </think>

D

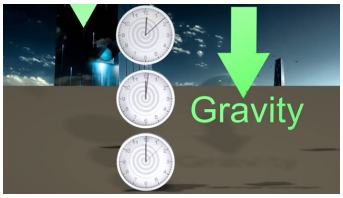
F.2 A02 / Sound Event Reasoning

Question, evidence, and training answer

Question. What extra words does the narrator say immediately after the on-screen text 'To understand how gravity affects the flow of time' appears?

Options. A. consider the following B. check out the other videos C. the difference in the rates D. the box is accelerating E. the rate at which time flows F. the time between the light pulses G. the gravitational field H. near a black hole

Training answer. A: consider the following



Source video, 00:03

Upstream reasoning steps.
1. [visual 00:00-00:06] The on-screen text shows: 'To understand how gravity affects the flow of time'.
2. [audio 00:00-00:06] The narrator says: 'To understand how gravity affects the flow of time, consider the following.'
3. [audiovisual 00:00-00:06] Comparing the visual text and the audio speech, the audio contains the extra phrase 'consider the following' that is not present in the on-screen text.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the spoken words that immediately follow a specific phrase at the beginning of the video. The crux is matching the narrator's spoken audio to the correct option among the choices provided.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the phrase "To understand how gravity affects the flow of time" and watch what is on screen while it is said.

- At 00:01, the narrator begins speaking, saying "To understand how gravity affects the flow of time" while three clocks are shown on screen.

- At 00:04, the narrator continues the sentence by saying "consider the following" as the green arrows pointing down remain on screen.

Complete synthetic thinking and final answer (continued)

- Let me verify the visual anchor: from 00:01 to 00:06 the on-screen text reads ”To understand how gravity affects the flow of time”, exactly matching the phrase quoted in the question, while the narrator speaks the same words aloud.

- Comparing the two, the narrator’s sentence continues beyond the on-screen text: the audio adds ”consider the following”, which never appears on screen.

- If the narrator had said ”check out the other videos” immediately after the opening line, that phrase would have to be audible in the first ten seconds, but it is only spoken at the very end of the video.

- At 10:45, the narrator says ”Please check out the other videos on this channel” as the closing text appears.

- The narrator’s spoken words immediately following the target phrase are ”consider the following”, which corresponds to option A.

Next, let me analyze the options:

- A. consider the following — This matches the exact words spoken by the narrator immediately after the opening phrase.

- B. check out the other videos — This is spoken at the very end.

- C. the difference in the rates — This is spoken much later.

- D. the box is accelerating — This is spoken during the acceleration explanation.

- E. the rate at which time flows — This is spoken at 04:42.

- F. the time between the light pulses — This is spoken at 02:59.

- G. the gravitational field — This is spoken during the gravity explanation.

- H. near a black hole — This is spoken at 10:15.

Finally, the correct option is A. </think>

A

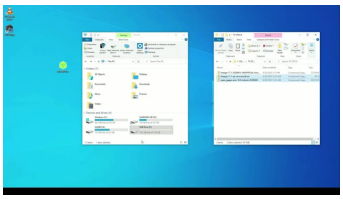
F.3 A03 / Music & Atmosphere Understanding

Question, evidence, and training answer

Question. What type of background audio is heard while the file explorer windows showing files being copied to a USB drive are visible?

Options. A. Upbeat background music from outro B. Upbeat electronic music C. Silence D. Background audio from Sling app E. Upbeat background music from TV show clips F. Faint background audio from video and app navigation G. Light pop sound effect with upbeat background tune H. Faint ambient room noise

Training answer. B: Upbeat electronic music



Source video, 02:11

Upstream reasoning steps.
1. [visual 02:00-02:22] The screen shows file explorer windows with a progress dialog indicating files being copied to a USB drive.
2. [audio 02:13-02:22] Upbeat electronic music is playing during this time interval, which overlaps with the file copying visual.
3. [audiovisual 02:13-02:22 (audio) / 02:00-02:22 (visual)] Therefore, the background audio while the file explorer windows are visible is upbeat electronic music.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the type of background audio playing while the file explorer windows show files being copied to a USB drive. The crux is determining the genre or style of the background music during this specific visual sequence.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the file explorer windows are visible showing files being copied to a USB drive.

- At 01:27, an upbeat electronic music track begins playing in the background as the narrator discusses downloading files.

- At 02:00, the screen displays the Windows desktop with two file explorer windows open, and the narrator says, "While Lineage OS is being flashed to the SD card, go ahead and copy open gaps and recovery 2 boot..."

- At 02:14, a file copying progress dialog box appears on screen showing the files being copied to the USB drive while the upbeat electronic music continues to play clearly.

- If the background audio were silent or consisted only of ambient room noise, there would be no musical melody playing behind the narrator's voice from 02:00 to 02:21, but a distinct electronic synth track is clearly audible.

- At 02:21, the file copying sequence ends and the video transitions to a new scene.

- The background track playing during this entire file-copying sequence is a light, upbeat electronic tune, which corresponds to option B.

Next, let me analyze the options:

- A. Upbeat background music from outro — This music plays during the middle, not the outro.

- B. Upbeat electronic music — This matches the synth-pop electronic track playing in the background of the file copying segment.

- C. Silence — There is clear background music playing here.

- D. Background audio from Sling app — The Sling app is not open during this segment.

- E. Upbeat background music from TV show clips — No TV show clips are playing during this demonstration.

- F. Faint background audio from video and app navigation — The audio is a prominent music track, not navigation.

- G. Light pop sound effect with upbeat background tune — No pop sound effects accompany the file copying action.

- H. Faint ambient room noise — The audio is dominated by music, not room noise.

Finally, the correct option is B. </think>

B

F.4 A04 / Cross-Modal Reference Resolution

Question, evidence, and training answer

Question. The expert whose on-screen title is 'Head of the Asia Programme at Chatham House' makes a claim about the size of the detonation. Which of the following statements did he utter?

Options. A. Given the advanced state of North Korean ballistic missile, intermediate-range ballistic missile technology, it's only a matter of time if they have developed a workable hydrogen warhead, before they miniaturize it. B. The H-bomb uses a smaller atomic bomb to trigger a larger thermonuclear explosion. C. The last time that country tested a nuclear weapon was back in 2013. D. Seoul, Tokyo, and Beijing are all in fairly easy range. E. The reported size of the detonation is relatively modest, comparable to the test that North Korea carried out in 2013. F. Experts estimate that that weapon had a yield of about seven kilotons. G. The explosion creates huge amounts of energy, triggering a seismic reaction. H. The DPRK's longer-range missiles could potentially hit much of India and Russia and even parts of the U.S. state of Alaska.

Training answer. E: The reported size of the detonation is relatively modest, comparable to the test that North Korea carried out in 2013.



Source video, 01:27

Upstream reasoning steps.
1. [visual 01:19-01:35] The visual banner displays the name 'John Nilsson-Wright' and the title 'Head of the Asia Programme at Chatham House'.
2. [audio 01:19-01:27] An audio statement is heard: 'The reported size of the detonation is relatively modest, comparable to the test that North Korea carried out in 2013.'
3. [audiovisual 01:19-01:27] The audio statement is uttered by the same person whose visual title is shown; therefore the statement belongs to the expert with that title.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the statement made by the expert with the on-screen title "Head of the Asia Programme at Chatham House" regarding the size of the detonation. The crux is matching the expert's spoken words to one of the given options.

First, let me scan the video for the segments relevant to the question:

- I need to find the expert with the on-screen title "Head of the Asia Programme at Chatham House" and listen to his statement about the detonation. - At 01:19, a man appears on screen with the title "Head of the Asia Programme at Chatham House" and the name "John Nilsson-Wright". - At 01:20, he says "The reported size of the detonation is relatively modest, it's comparable to the test that North Korea carried out in 2013". - If the expert's statement were about ballistic missile technology, the speaker would have to be Justin Bronk who appears at 02:01 with the title "Military Security Analyst", so the statement is the one about the modest size of the detonation. - The expert's spoken words match the statement about the modest size of the detonation, which is option E.

Next, let me analyze the options:

- A. Given the advanced state of North Korean ballistic missile, intermediate-range ballistic missile technology, it's only a matter of time if they have developed a workable hydrogen warhead, before they miniaturize it. — spoken by Justin Bronk. - B. The H-bomb uses a smaller atomic bomb to trigger a larger thermonuclear explosion. — spoken by the narrator. - C. The last time that country tested a nuclear weapon was back in 2013. — spoken by the anchor. - D. Seoul, Tokyo, and Beijing are all in fairly easy range. — spoken by the narrator. - E. The reported size of the detonation is relatively modest, comparable to the test that North Korea carried out in 2013. — this matches the exact words spoken by John Nilsson-Wright. - F. Experts estimate that that weapon had a yield of about seven kilotons. — spoken by the anchor. - G. The explosion creates huge amounts of energy, triggering a seismic reaction. — spoken by the narrator. - H. The DPRK's longer-range missiles could potentially hit much of India and Russia and even parts of the U.S. state of Alaska. — spoken by the narrator.

Finally, the correct option is E. </think>

E

F.5 A05 / AUDIO-ANCHORED PERCEPTION

Question, evidence, and training answer

Question. When the speaker said 'source to sink', what is the model of the machine pointed at by the person who has the same shirt color as the man sitting across from the woman with an arm cast in the meeting room?

Options. A. EOS M 400. B. EOS M 290. C. EOS M 300. D. EOS M 100.

Training answer. A: EOS M 400.



Source video, 14:33



Source video, 14:51

Upstream reasoning steps.

1. [audio 00:14:49-00:14:52] we ask them about okay source to sink
2. [visual 00:14:31-00:14:35] The man sitting across from the woman with an arm cast in the meeting room is wearing a maroon shirt.
3. [visual 00:14:50-00:14:52] The person wearing a maroon shirt points at a machine labeled 'EOS M 400'.
4. [inference] At 00:14:50, the speaker says 'source to sink' while a man in a maroon shirt (matching the shirt color of the man sitting across from the woman with a cast in the earlier meeting room scene) points to a machine clearly labeled 'EOS M 400'.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the model of the machine being pointed at when the speaker says "source to sink", based on the shirt color of the man sitting across from the woman with an arm cast in the meeting room. The crux is identifying the shirt color of that specific man in the meeting room and then finding the machine pointed at by the person wearing that same color on the factory floor when the target phrase is spoken.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the phrase "source to sink" and watch what is on screen while it is said.

- At 14:31, the video shows a meeting room where a group of people are sitting around a table, including a woman in a blue shirt with a white cast on her arm.

- At 14:32, we see a man sitting across the table from the woman with the cast, and he is wearing a maroon polo shirt.

- Wait, let me verify the shirt color of the man sitting across from the woman with the cast in the meeting room: at 14:33, a close-up shows him smiling in a dark red or maroon polo shirt, which matches the maroon polo shirt worn by the man pointing at the machine on the factory floor at 14:50.

- At 14:49, the speaker says "we ask them about okay source to sink, how do you bring in the materials" as the video cuts to a factory floor showing a man in a maroon polo shirt pointing at a large grey machine.

- At 14:50, the camera zooms in on the front panel of the grey machine being pointed at by the man in the maroon polo shirt.

- If the machine being pointed at were the EOS M 290, the label on its front panel would have to read "EOS M 290" when the man points to it at 14:50, but the text printed on the machine clearly reads "EOS M 400" — so it is the EOS M 400.

- The machine pointed at by the man in the maroon polo shirt is labeled "EOS M 400", which corresponds to option A.

Next, let me analyze the options:

- A. EOS M 400 — This matches the label on the machine pointed at by the man in the maroon polo shirt at 14:50.

- B. EOS M 290 — This model is mentioned earlier at 05:41, not pointed at here.

- C. EOS M 300 — This model is not shown in the video.

- D. EOS M 100 — This model is not shown in the video.

Finally, the correct option is A. </think>

A

F.6 A06 / COUNTING

Question, evidence, and training answer

Question. In the segment about the filter that uses a dot indicator on its strip technology to signal when it needs to be changed, how many text boxes appear in the visual list of its features?

Options. A. 15 B. 12 C. 6 D. 4 E. 3 F. 9 G. 0 H. 7

Training answer. F: 9



Source video, 01:06

Upstream reasoning steps.
1. [audio 00:53-01:02] The narrator says 'Dot time strip technology' which uniquely identifies the goldfish filter segment.
2. [visual 01:02-01:11] The visual list for the goldfish filter segment contains 9 text boxes (three sets of three, with the final set appearing at this timestamp).

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know the number of text boxes that appear in the visual list of features for the filter that uses a dot indicator on its strip technology to signal when it needs to be changed. The crux is identifying this specific filter in the video and counting the text boxes that appear on screen during its segment.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment for the filter that features strip technology to indicate when it needs to be changed and count its feature text boxes.

- At 00:39, the segment for "HOB Filter for Goldfish" begins, showing the filter on the left.

- At 00:44, the first text box appears on screen with the text "Timestrip technology tells".

- At 00:47, the second text box appears with the text "The filter needs to be changed".
- At 00:49, the third text box appears with the text "Easy access door".
- At 00:54, the fourth text box appears with the text "Makes changing the filter".
- At 00:57, the fifth text box appears with the text "Easy setup for beginners".
- At 00:59, the sixth text box appears with the text "Outflow is powerful enough".
- At 01:05, the seventh text box appears with the text "Oxygenate the water".
- At 01:07, the eighth text box appears with the text "Internal motor minimized noise".
- At 01:09, the ninth text box appears with the text "Rate of 110 gallons per hour".
- If there were only seven text boxes in this list, the count would stop at the box for "Oxygenate the water", but two more boxes appear after it to complete the list of nine features.
- The total number of text boxes that appear in this visual list is nine, which corresponds to option F.

Next, let me analyze the options:

- A. 15 — this is incorrect as only nine boxes appear.
- B. 12 — this is incorrect as only nine boxes appear.
- C. 6 — this is incorrect as only nine boxes appear.
- D. 4 — this is incorrect as only nine boxes appear.
- E. 3 — this is incorrect as only nine boxes appear.
- F. 9 — exactly nine text boxes are displayed in the goldfish filter segment.
- G. 0 — this is incorrect as only nine boxes appear.
- H. 7 — this is incorrect as only nine boxes appear.

Finally, the correct option is F. </think>

F

F.7 A07 / TEMPORAL REASONING & ORDERING

Question, evidence, and training answer

Question. At the moment the shirtless Maori warrior with facial tattoos and a red headband performing a welcoming challenge with a wooden weapon first appeared, which specific audio event was synchronized with it?

Options. A. A single warrior shouting a traditional challenge B. A group of men chanting rhythmically during an indoor performance C. The rhythmic percussive sound of poi balls accompanied by singing D. The deep sound of a traditional horn being blown

Training answer. A: A single warrior shouting a traditional challenge



Source video, 01:06

Upstream reasoning steps.
1. [visual 01:05-01:08] Shirtless Maori warrior with facial tattoos and a red headband steps forward and performs a challenge with a wooden weapon.
2. [audio 01:05-01:08] Loud, traditional shouting by a single warrior.
3. [inference] When the shirtless Maori warrior with the red headband and wooden weapon appears at 01:05, the synchronized audio is his loud, traditional shout, distinguishing it from the horn, group chant, and poi ball sounds that occur at other moments in the sequence.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the specific audio event synchronized with the first appearance of a shirtless Maori warrior with facial tattoos and a red headband performing a welcoming challenge with a wooden weapon. The crux is matching the visual of this specific warrior's first on-screen appearance with the concurrent sound.

First, let me scan the video for the segments relevant to the question:

- I need to find the first appearance of the shirtless Maori warrior with facial tattoos and a red headband performing a welcoming challenge with a wooden weapon.

- At 00:00:58, Kyle Pedley says "I've been asked to be involved in the Maori cultural welcome." as the visitors walk towards the meeting ground.

- At 00:01:01, Kyle Pedley says "It sounds both terrifying and exciting as well, so looking forward to it." while a warrior blows a traditional shell horn.

- At 00:01:05, the shirtless Maori warrior with facial tattoos and a red and white headband first appears, stepping forward with a wooden weapon and shouting a traditional challenge.

- At 00:01:08, Brad Knopp says "We got to actually learn the Haka." as the visitors watch the performance.

- At 00:01:10, a group of men is seen chanting rhythmically during an indoor Haka performance.

- At 00:01:14, women are shown performing with poi balls accompanied by singing.

- If the synchronized audio event were a group of men chanting rhythmically, the warrior would have to be shown indoors at 00:01:10, but his first appearance is outdoors at 00:01:05 — so it is a single warrior shouting.

- The first appearance of the described warrior at 00:01:05 is synchronized with a single warrior shouting a traditional challenge, which corresponds to option A.

Next, let me analyze the options:

- A. A single warrior shouting a traditional challenge — This matches the audio heard at 00:01:05 when the specific warrior first appears.

- B. A group of men chanting rhythmically during an indoor performance — This occurs later at 00:01:10 during the indoor Haka.

- C. The rhythmic percussive sound of poi balls accompanied by singing — This is heard at 00:01:14.

Complete synthetic thinking and final answer (continued)

- D. The deep sound of a traditional horn being blown — This is heard earlier at 00:01:01.

Finally, the correct option is A. </think>

A

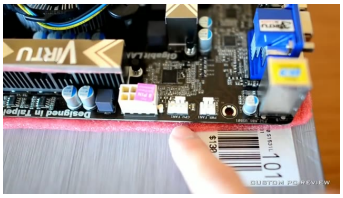
F.8 A08 / SPATIAL REASONING

Question, evidence, and training answer

Question. What is the position of the blue and pink motherboard ports relative to the CPU fan header on the screen when the narrator says, 'This is where you will plug in the four-pin connector for the CPU fan'?

Options. A. On the left. B. On the right. C. Directly above. D. Directly below.

Training answer. B: On the right.



Source video, 02:33

Upstream reasoning steps.

1. [audio 02:31-02:35] This is where you will plug in the four-pin connector for the CPU fan.
2. [visual 02:31-02:35] The camera angle shows the blue and pink rear I/O ports positioned on the right side of the screen relative to the CPU fan header.
3. [inference] The audio quote identifies the specific timeframe (02:31-02:35), during which the camera angle is reversed from its usual orientation, showing the blue and pink ports on the right side of the CPU fan header on the screen.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know the position of the blue and pink motherboard ports relative to the CPU fan header when the narrator points to it and speaks. The crux is identifying where these ports are located on the screen in the close-up shot of the CPU fan header.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the phrase "This is where you will plug in the four-pin connector for the CPU fan" and observe the position of the blue and pink ports relative to the CPU fan header on the screen.

- At 00:01, the narrator introduces himself saying, "Hey, what's up YouTube? Sam here with Custom PC Review", while showing the motherboard box labeled "ASRock" and "Z68 Pro3-M".

- At 02:26, the narrator says, "Next, find the four-pin header on your motherboard that says CPU fan", as the camera prepares to zoom in on the board.

- At 02:31, the narrator points to the header and says, "This is where you will plug in the four-pin connector for the CPU fan", showing a close-up of the board.

- Wait, let me re-read the motherboard labels in that close-up at 02:31: the white power connector to the left of the CPU fan header is labeled "8 Pin" on its pink sticker, while the header itself is labeled "CPU_FAN1" and the adjacent header is "PWR_FAN1", confirming the layout.

- If the blue and pink motherboard ports were on the left, they would have to be visible next to the white power connector with the pink sticker, but that area only shows capacitors and a heatsink — so they are on the right.

- The blue and pink motherboard ports are located on the right side of the screen relative to the CPU fan header, which corresponds to option B.

Next, let me analyze the options:

- A. On the left. — Incorrect, this side shows the white power connector.

- B. On the right. — Correct, the blue VGA and pink audio ports are visible here.

Complete synthetic thinking and final answer (continued)

- C. Directly above. — Incorrect.

- D. Directly below. — Incorrect.

Finally, the correct option is B. </think>

B

F.9 A09 / CAUSAL REASONING

Question, evidence, and training answer

Question. What defensive action does the trainer state he looks for before performing the move where he pivots on his inside foot and opens his outside hip?

Options. A. The defender switches B. The defender mirrors the ball handler C. The defender starts to sag D. The screener's defender sucks up to the frame E. The defender drops back F. The defender goes under the screen G. The defender corrals the ball handler H. The defender hedges

Training answer. D: The screener's defender sucks up to the frame



Source video, 02:17

Upstream reasoning steps.
1. [visual 02:12-02:22] The trainer is seen performing a slow-motion move where he pivots on his inside foot and opens his outside hip, which is the visual demonstration of the hip rotation.
2. [audio 02:08-02:13] The trainer says he looks for the screener's defender sucking up to the frame before performing the hip rotation.
3. [audiovisual 02:08-02:13] Combining s1 and s2, the defender's action of sucking up to the frame is the cause for the demonstrated hip rotation move.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the defensive action the trainer looks for before executing a move that involves pivoting on the inside foot and opening the outside hip. The crux is matching the trainer's spoken cue for this specific footwork to one of the listed defensive scenarios.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the trainer's description of pivoting on the inside foot and opening the outside hip, and identify the defensive action that prompts it. - At 00:11, the trainer introduces the drill as "snaking the pick and roll". - At 00:45, the trainer explains that if the screener's defender starts to "sag" or put him in a "corral situation", he will execute a step-back. - For the sag or corral reads to be correct, they would have to trigger the hip rotation, but the trainer explicitly connects them to the step-back instead. - At 01:26, the trainer explains a change of tempo called a "lift". - At 02:04, the trainer introduces the "hip rotation" option. - At 02:08, the trainer states, "the read is going to be screener's defender sucking up to our frame." - At 02:16, the trainer describes the mechanics of the hip rotation as "slightly pivoting on that inside foot, opening up that outside hip". - At 02:25, the trainer notes that defenders have a "tendency of mirroring what the offensive player does". - The trainer's explanation directly connects the inside foot pivot and outside hip opening to the read where the screener's defender sucks up to the frame, which is option D.

Next, let me analyze the options:

- A. The defender switches — This action is not mentioned. - B. The defender mirrors the ball handler — This is a general defensive tendency. - C. The defender starts to sag — This prompts the step-back. - D. The screener's defender sucks up to the frame — This is the exact read the trainer states he looks for before performing the hip rotation with the inside foot pivot. - E. The defender drops back — This is not discussed. - F. The defender goes under the screen — This is not mentioned. - G. The defender corrals the ball handler — This leads to the step-back. - H. The defender hedges — This is not mentioned.

Complete synthetic thinking and final answer (continued)

Finally, the correct option is D. </think>
D

F.10 A10 / Emotion & Sentiment Analysis

Question, evidence, and training answer

Question. How does the emotional state of the young man wearing a black cap and dark grey t-shirt change after he comes home?

Options. A. From calm to fearful. B. From calm to angry. C. From fearful to angry. D. From annoyed to fearful.

Training answer. A: From calm to fearful.



Source video, 02:09

Upstream reasoning steps.
1. [audio 02:01-02:05] The young man casually and calmly says school was good and nothing out of the ordinary happened.
2. [visual 02:06-02:12] His mother says 'Your principal called me,' and he freezes with a wide-eyed, shocked expression.
3. [audio 02:16-02:21] He stammers anxiously, trying to defend himself.
4. [inference] The young man is initially relaxed and calm when discussing his day, but his expression and tone shift to clear fear and anxiety immediately after his mother mentions that his principal called.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know how the emotional state of the young man wearing a black cap and dark grey t-shirt changes after he comes home in the video. The crux is identifying his initial emotional state when he enters the house and tracing how it shifts as the scene unfolds.

First, let me scan the video for the segments relevant to the question:

- I need to observe the emotional reactions of the young man in the black cap and dark grey t-shirt after he comes home. - At 01:54, the young man opens the door, enters the house, and calmly says "Hey mom" to the person on the couch. - At 02:00, he walks into the living room and continues speaking in a relaxed, normal tone, saying that school was "good" and "nothing out of the ordinary." - At 02:06, the mother figure on the couch tells him "Your principal called me," which immediately changes the tone of the interaction. - At 02:10, the young man stands by the open refrigerator, looking startled and muttering "shit" under his breath. - At 02:13, the mother figure confronts him in the kitchen while holding a belt, demanding to know what she told him about acting up in school. - At 02:17, the young man becomes visibly defensive and anxious, pleading "Alright mom, look. I know I got in trouble, but it wasn't me. I promise." - If he were angry rather than fearful during this confrontation at 02:17, he would be shouting back or showing aggression, but instead he is stammering, backing away, and trying to appease his mother. - At 02:22, the mother figure prepares the belt, and the young man's expression becomes increasingly terrified as he braces for a beating. - The young man goes from a relaxed, normal demeanor when entering the house to a terrified, defensive state when confronted with a belt, showing a transition from calm to fearful — that is option A.

Next, let me analyze the options:

- A. From calm to fearful. — He enters the house in a relaxed state and ends up terrified when chased with a belt. - B. From calm to angry. — He shows fear and defensiveness rather than anger. - C. From fearful to angry. — He starts calm and does not become angry. - D. From annoyed to fearful. — He is relaxed, not annoyed, when he first arrives.

Complete synthetic thinking and final answer (continued)

Finally, the correct option is A. </think>

A

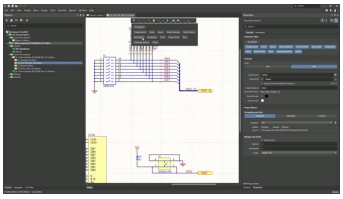
F.11 A11 / HYPOTHETICAL & PREDICTIVE REASONING

Question, evidence, and training answer

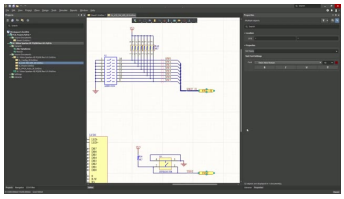
Question. If the user had not applied the pre-selection filter for ports and net labels before drawing the selection rectangle, what would most likely be the result?

Options. A. All objects except ports and net labels would be unselected. B. The software would highlight the filter settings. C. Other objects besides ports and net labels would also be selected. D. The selection rectangle would be disabled. E. Nothing would be selected. F. Only ports would be selected. G. Only net labels would be selected. H. Only ports and net labels would be selected.

Training answer. C: Other objects besides ports and net labels would also be selected.



Source video, 03:18



Source video, 03:35

Upstream reasoning steps.
1. [visual 03:15-03:22] The user applies the pre-selection filter for ports and net labels.
2. [audio 03:32-03:38] The narrator states that because of the filter, only ports and net labels are selected, despite other objects being present.
3. [audiovisual 03:32-03:38] Infer that without the filter, the other objects in the selection area would also be selected.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know the outcome of drawing a selection rectangle without first applying the pre-selection filter for ports and net labels. The crux is determining how the software behaves when no filter restricts the selection tool.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the narrator explains the function of the pre-selection filter during a rectangle selection.

- At 03:05, the narrator introduces the pre-selection filter to facilitate selecting specific schematic elements.

- At 03:15, the narrator applies the filter specifically for ports and net labels.

- At 03:25, a selection rectangle is drawn over a variety of components, wires, ports, and net labels.

- At 03:32, the narrator explains that "only ports and net labels will be selected, despite other objects within the selection area, because of our pre-selection filter."

- If the filter were not applied, the standard selection rectangle would capture all objects it touches, meaning other elements like wires and components would also be selected.

- The demonstration confirms that the filter is the sole reason other objects are ignored, so omitting it would result in selecting other objects besides ports and net labels, which is option C.

Next, let me analyze the options:

- A. All objects except ports and net labels would be unselected. — This describes the filtered outcome, not the unfiltered one.

- B. The software would highlight the filter settings. — There is no indication of this behavior in the video.

- C. Other objects besides ports and net labels would also be selected. — Without the filter, standard rectangle selection captures all objects in its path.

Complete synthetic thinking and final answer (continued)

- D. The selection rectangle would be disabled. — The tool remains fully functional without a filter.
- E. Nothing would be selected. — Standard selection would still function normally.
- F. Only ports would be selected. — This would require a port-only filter.
- G. Only net labels would be selected. — This would require a net-label-only filter.
- H. Only ports and net labels would be selected. — This is the result when the filter is active.

Finally, the correct option is C. </think>

C

F.12 A12 / Character Relationship & Interaction

Question, evidence, and training answer

Question. What is the nature of the interaction between the vlogger and the statue of Don Knotts just before the vlogger stands up and walks away?

Options. A. The vlogger says, "Thanks for the laughs, man," and gives the statue a fist bump. B. The vlogger says, "Well, take care," and gives the statue a fist bump. C. The vlogger says, "Well, take care," and shakes the statue's hand. D. The vlogger says, "Thanks for the laughs, man," and shakes the statue's hand.

Training answer. B: The vlogger says, "Well, take care," and gives the statue a fist bump.



Source video, 05:50

Upstream reasoning steps.

1. [audio—visual 05:45-05:55] The vlogger sits next to the statue, says "Well, take care," and gives the statue a fist bump before standing up.
2. [inference] The vlogger sits next to the statue, says "Well, take care," and physically gives the statue's hand a fist bump before getting up and leaving.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know the nature of the interaction between the vlogger and the statue of Don Knotts just before the vlogger stands up and walks away. The crux is identifying the exact words spoken and the physical gesture made right before he leaves.

First, let me scan the video for the segments relevant to the question: - I need to find the segment where the vlogger interacts with the Don Knotts statue and observe his final words and actions before walking away. - At 03:41, the vlogger walks toward the Metropolitan Theatre where the statue of Don Knotts is seated on a bench. - At 04:12, the camera shows a close-up of the bronze statue, depicting Don Knotts holding a script and a hat. - At 04:55, the vlogger stands next to the statue, gesturing toward it as he explains its history. - At 05:21, the camera pans over the details of the statue, showing the name "DON KNOTTS" on the base. - At 05:42, the vlogger says "Thanks for the laughs, man. Thanks for the laughs." while the camera is focused closely on the statue's face. - At 05:46, a jump cut shows the vlogger sitting on the bench directly next to the statue. - For the final dialogue to be "Thanks for the laughs, man," he would have to say those words while sitting next to the statue at the end of the scene, but he actually says them earlier during a close-up shot of the statue's face. - At 05:50, while still sitting on the bench, the vlogger says "Well, take care." and taps his closed hand against the statue's hand in a fist bump. - At 05:53, the vlogger stands up from the bench and walks out of the frame. - The vlogger says "Well, take care." and gives the statue a fist bump right before standing up, which corresponds to option B.

Next, let me analyze the options: - A. The vlogger says, "Thanks for the laughs, man," and gives the statue a fist bump. — Incorrect dialogue. - B. The vlogger says, "Well, take care," and gives the statue a fist

Complete synthetic thinking and final answer (continued)

bump. — The vlogger speaks these exact words, and he initiates a friendly fist bump. - C. The vlogger says, "Well, take care," and shakes the statue's hand. — Incorrect action. - D. The vlogger says, "Thanks for the laughs, man," and shakes the statue's hand. — Incorrect dialogue and action.

Finally, the correct option is B. </think>

B

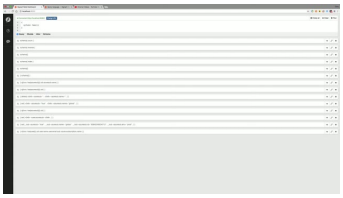
F.13 A13 / EVENT UNDERSTANDING & STATE CHANGE

Question, evidence, and training answer

Question. After the narrator announced they would begin with a simple query, what action did the user perform in the Dgraph Ratel UI?

Options. A. switched to the JSON view B. modified the query to use a filter C. typed a new query in the editor D. hovered over a node E. added a cascade directive F. clicked the Run button G. dragged a node H. changed the mutation data

Training answer. C: typed a new query in the editor



Source video, 02:24

Upstream reasoning steps.
1. [audio 02:11-02:21] The narrator says 'So let's start off with something simple.'
2. [visual 02:11-02:37] The user types a new query into the editor, as shown in the visual caption.
3. [audiovisual 02:11-02:37] Combining the audio anchor and the visual action, the user's action after the announcement is typing a new query.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know what action was performed in the Dgraph Ratel UI immediately after the narrator announced they would begin with a simple query. The crux is identifying the specific user action that followed this spoken cue.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the narrator's announcement about starting with a simple query and identify the action that immediately follows it in the UI.

- At 00:15, the user navigates to the mutate tab to enter some sample data.

- At 00:43, the user clicks the Run button to execute the mutation and add the data.

- At 00:50, the narrator says, "So let's start off with a simple uh schema query..."

- At 00:58, the user types the new query "schema {}" directly into the query editor.

- At 01:16, the user clicks the Run button to execute the schema query.

- At 01:21, the user hovers over the resulting schema nodes to inspect their attributes.

- At 01:45, the user switches to the JSON view to examine the schema values in JSON format.

- At 02:11, the narrator says, "So let's start off with something simple... we'll query for anything that has a predicate of user."

- At 02:28, the user types the new query "q(func: has(user)) { uid user.name }" into the editor.

- If the user had clicked the Run button immediately after the announcement, the query would have executed without any changes, but the user first had to input the query text — so the action is typing a new query.

- The immediate action following both announcements of a simple query is typing the query text into the editor, which corresponds to option C.

Complete synthetic thinking and final answer (continued)

Next, let me analyze the options:

- A. switched to the JSON view — occurred much later at 01:45.
- B. modified the query to use a filter — performed much later.
- C. typed a new query in the editor — immediate action following the announcement.
- D. hovered over a node — done after running the query.
- E. added a cascade directive — done near the end.
- F. clicked the Run button — done only after typing.
- G. dragged a node — performed to organize the graph.
- H. changed the mutation data — performed at the beginning.

Finally, the correct option is C. </think>

C

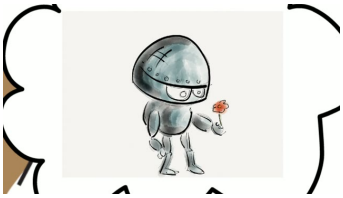
F.14 A14 / Cross-Scene Association & Long-Context

Question, evidence, and training answer

Question. The narrator mentions a prominent tech innovator during an early discussion about living in a simulation, which coincides with the drawing of a specific character type. Much later, when the narrator discusses super humans, a different version of this same character type is depicted. What distinctive physical feature does this later character possess?

Options. A. large white eyes and blue wings B. a skull mask and black hooded cloak C. a robotic right arm D. white feathered angel wings E. yellow hazmat suits and respirators F. a whistle around his neck G. glasses and brown hair H. blue overalls and a straw hat

Training answer. D: white feathered angel wings



Source video, 00:23



Source video, 03:10

This training row has no separately recorded structured upstream evidence. The source frames illustrate its question; the complete stored synthetic response follows.

Task label: retrospective.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the physical feature of a character depicted during a discussion about super humans, which is a different version of a character type drawn when a prominent tech innovator is mentioned in relation to living in a simulation. The crux is determining which character type corresponds to the tech innovator's mention and identifying the unique physical feature of the later version of this character.

First, let me scan the video for the segments relevant to the question:

- I need to find the early segment mentioning a tech innovator and a simulation to identify the character type, and then locate the later segment discussing super humans to see the updated character's features.

- At 00:00:12, a drawing of a man with glasses and brown hair appears on screen as the narrator introduces the book.

- At 00:00:21, the narrator mentions "Elon Musk living in a simulation" while a drawing of a small grey robot holding a flower is shown.

- At 00:01:15, two figures in yellow hazmat suits and respirators are shown during a discussion about Ebola.
- At 00:01:34, a cartoon farmer wearing blue overalls and a straw hat is shown holding a hoe.
- At 00:02:21, a cartoon sports coach with a whistle around his neck is drawn.

Complete synthetic thinking and final answer (continued)

- At 00:02:34, a figure representing death is shown wearing a skull mask and a black hooded cloak.
- At 00:03:08, the narrator discusses "super humans" while a blue and white robot with large white feathered angel wings is shown flying.
- At 00:03:22, a young man with a robotic right arm is shown sitting at a table.
- If the later character type were a cyborg human rather than a full robot, the character shown during the super humans discussion would have a robotic right arm instead of being a flying machine with wings.
- The flying robot shown during the super humans discussion at 00:03:08 clearly has white feathered angel wings, which corresponds to option D.

Next, let me analyze the options:

- A. large white eyes and blue wings — the robot's wings are white, not blue.
- B. a skull mask and black hooded cloak — this represents death, not the robot.
- C. a robotic right arm — this feature belongs to a cyborg shown later, not the flying robot.
- D. white feathered angel wings — the flying robot representing super humans possesses these large white wings.
- E. yellow hazmat suits and respirators — these are Ebola workers.
- F. a whistle around his neck — this is the coach.
- G. glasses and brown hair — this is the narrator.
- H. blue overalls and a straw hat — this is the farmer.

Finally, the correct option is D. </think>

D

F.15 A15 / SUMMARIZATION & SYNTHESIS

Question, evidence, and training answer

Question. What overall mood is created by the combination of the audio and visual elements in this video?

Options. A. a melancholic and sorrowful atmosphere B. a joyful and energetic atmosphere C. a dark and chaotic atmosphere D. a dramatic and epic atmosphere E. a tense and suspenseful atmosphere F. a serene and mystical atmosphere G. a calm and peaceful atmosphere H. a whimsical and fantastical atmosphere

Training answer. C: a dark and chaotic atmosphere



Source video, 00:50



Source video, 01:40

Upstream reasoning steps.

1. [audio 00:00-00:27] The audio is fast-paced death metal with heavy distorted guitars and blast beat drums, establishing an aggressive and chaotic mood.
2. [visual 00:00-02:30] The visual is a dark, surreal album cover with bone-like structures and a dark archway, establishing a dark and foreboding mood.
3. [audiovisual 00:00-02:30] Combining the aggressive audio and dark visual, the overall mood is dark and chaotic.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to determine the overall mood created by combining the audio and visual elements of the video. The crux is identifying how the aggressive musical style and the grotesque, surreal imagery on screen work together to establish a specific atmosphere.

Complete synthetic thinking and final answer (continued)

First, let me scan the video for the segments relevant to the question:

- I need to examine the album artwork on screen and listen to the musical style of the track to characterize their combined aesthetic.

- At 00:00-00:25, the video displays a static album cover featuring the band name "Ænigmatum" and the album title "Deconsecrate" written around a grotesque, fleshy, skull-like cavern with deep red and pale bone colors.

- At 00:26-01:15, the audio shifts from a dark guitar intro into frantic blast beats, heavy distorted riffs, and deep guttural vocals. If the overall mood were dramatic and epic rather than dark and chaotic, the music would need a grand, orchestral, or melodic structure instead of these dissonant, aggressive death metal elements.

- At 01:16-03:00, the relentless tempo and chaotic technical death metal riffs continue alongside the nightmarish, biomechanical visual of a spine-like pathway leading to a dark altar.

- The combination of extreme, aggressive death metal and a grotesque, fleshy, hellish landscape creates a dark and chaotic atmosphere, which corresponds to option C.

Next, let me analyze the options:

- A. a melancholic and sorrowful atmosphere — The aggressive tempo and harsh vocals are not sad.

- B. a joyful and energetic atmosphere — The dark, grotesque imagery contradicts any joyful feeling.

- C. a dark and chaotic atmosphere — The dissonant death metal and nightmarish, fleshy artwork establish this exact feeling.

- D. a dramatic and epic atmosphere — The chaotic style is too aggressive for an epic tone.

- E. a tense and suspenseful atmosphere — The music is immediately explosive rather than building suspense.

- F. a serene and mystical atmosphere — The harsh vocals and gory visuals are not peaceful.

- G. a calm and peaceful atmosphere — The fast blast beats are highly intense and aggressive.

- H. a whimsical and fantastical atmosphere — The hellish, biomechanical landscape is macabre rather than whimsical.

Finally, the correct option is C. </think>

C

F.16 A16 / Ego-Centric & First-Person Understanding

Question, evidence, and training answer

Question. When the sound of highway traffic passing underneath is heard for the first time, what gesture do I make with my left hand?

Options. A. Pointing to the right. B. Giving a thumbs up. C. Pointing to the left. D. Adjusting the bike computer.

Training answer. A: Pointing to the right.



Source video, 01:04



Source video, 00:48

Upstream reasoning steps.
1. [audio 00:48-01:08] The distinct whooshing sound of cars passing underneath is heard as the cyclist rides over the first bridge.
2. [visual 01:03-01:05] The cyclist lifts their left hand off the handlebar and points to the right with their index finger.
3. [inference] The sound of highway traffic passing below is first heard when crossing the pedestrian bridge starting at 00:48; during this time, the cyclist raises their left hand and points to the right.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the gesture made with the cyclist's left hand when the sound of highway traffic passing underneath is first heard. The crux is locating the first instance of audible highway traffic from below and observing the hand gesture that immediately follows.

First, let me scan the video for the segments relevant to the question:

- I need to find the first segment where the sound of highway traffic passing underneath is heard and identify the gesture made with the left hand. - At 00:01, the cyclist rides on a paved path through a residential neighborhood. - At 00:10, the cyclist raises their left hand briefly to adjust their glove. - At 00:18, the cyclist passes a pedestrian on the path. - At 00:48, the cyclist begins riding onto a pedestrian bridge over a highway. - At 00:50, the sound of highway traffic passing underneath becomes clearly audible for the first time. - At 01:03, the cyclist lifts their left hand off the handlebar. - At 01:04, the cyclist points to the right with their left hand, extending the index and middle fingers. - If the cyclist were giving a thumbs up at this moment, the thumb would be extended upward with the other fingers curled, but instead the index and middle fingers are extended to point toward the right side of the bridge. - At 20:53, near the end of the ride, the cyclist gives a thumbs up with their left hand. - The gesture made when the highway traffic is first heard is pointing to the right, which corresponds to option A.

Next, let me analyze the options:

- A. Pointing to the right. — This matches the gesture at 01:04, pointing toward the highway. - B. Giving a thumbs up. — observed only at 20:53. - C. Pointing to the left. — hand points right instead. - D. Adjusting the bike computer. — no adjustment occurs.

Finally, the correct option is A. </think>

A

F.17 B01 / Problem-solving Adaptation

Question, evidence, and training answer

Question. The figure shows four projects with the quality setting selected in the Blend Collage app. According to the video's classification of the save quality options, which project is incorrectly configured?

Options. A. Photo Print B. Blog Header C. Email Newsletter D. Mobile Wallpaper

Training answer. B: Blog Header



Source video, 05:25

Project Name	Quality Setting
Mobile Wallpaper	Standard
Email Newsletter	High
Blog Header	Standard
Photo Print	Maximum

New figure

Upstream reasoning steps.

1. [05:20-05:30] Recall the video's classification of the three save quality options: Standard is for phone screen, High is for email/website/blog, Maximum is for printing.
2. For each project, infer its intended use from the project name: 'Mobile Wallpaper' implies phone screen, 'Email Newsletter' implies email, 'Blog Header' implies blog, 'Photo Print' implies printing.
3. Check the quality setting against the classification: Mobile Wallpaper (Standard) → correct; Email Newsletter (High) → correct; Blog Header (Standard) → should be High, so incorrect; Photo Print (Maximum) → correct.
4. Identify the project with the mismatch: Blog Header.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

Complete synthetic thinking and final answer (continued)

The user wants to identify which project in the provided table is incorrectly configured based on the video's description of save quality options in the Blend Collage app. The crux is matching each project's quality setting in the table to the corresponding save quality option described in the video.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the presenter explains the different save quality options in the app.

- At 05:20, the presenter clicks the "SAVE" button, which brings up the "Save collage" dialog box.

- At 05:22-05:32, the dialog box shows "Standard" as "Good for gadget screens", "High" as "Best for sharing via website, blog, or email", and "Maximum" as "Ideal for printing and viewing on computer screens".

- If the "High" option were meant for gadget screens, the dialog box would describe it as such, but it actually says "Best for sharing via website, blog, or email" while "Standard" is described as "Good for gadget screens" — so "Standard" is for gadget screens and "High" is for websites, blogs, or email.

Next, let me read the figure:

- The table lists four projects: "Mobile Wallpaper", "Email Newsletter", "Blog Header", and "Photo Print".

- For "Mobile Wallpaper", the quality setting is configured as "Standard".

- For "Email Newsletter", the quality setting is configured as "High".

- For "Blog Header", the quality setting is configured as "Standard".

- For "Photo Print", the quality setting is configured as "Maximum".

- According to the video, "Standard" is for gadget screens, "High" is for websites, blogs, or emails, and "Maximum" is for printing.

- Applying these rules, "Blog Header" should be configured as "High" because it is a blog project, but it is incorrectly configured as "Standard".

- Therefore, the project "Blog Header" is incorrectly configured — this is option B.

Next, let me analyze the options:

- A. Photo Print — correctly configured as Maximum.

- B. Blog Header — incorrectly configured as Standard, as the video states blogs require High quality.

- C. Email Newsletter — correctly configured as High.

- D. Mobile Wallpaper — correctly configured as Standard.

Finally, the correct option is B. </think>

B

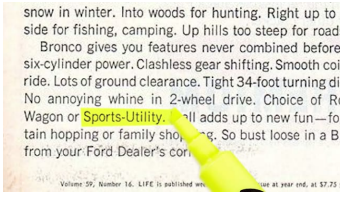
F.18 B02 / Case Study Analysis

Question, evidence, and training answer

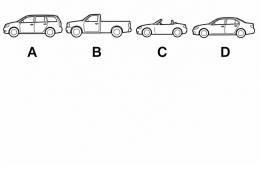
Question. Based on the lecture's classification of the original Bronco body styles, which labelled region in the figure corresponds to the model that the first Bronco brochure referred to as a 'sports utility'?

Options. A. Region A B. Region C C. Region D D. Region B

Training answer. D: Region B



Source video, 02:27



New figure

Upstream reasoning steps.

1. [00:02:24-00:02:30] Recall the lecture's statement that the first Bronco brochure referred to the pickup as a 'sports utility'.

2. Identify the labelled body styles in the figure: Region A is a wagon, Region B is a pickup, Region C is a roadster, Region D is a sedan.
3. Apply the lecture's naming: the 'sports utility' corresponds to the pickup, which is Region B.
4. Conclude that the correct answer is Region B.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify which labeled vehicle silhouette in the provided diagram represents the body style that the original Ford Bronco brochure designated as a 'sports utility'. The crux is to determine which of the three original body styles mentioned in the video corresponds to this specific term and match it to its silhouette.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment of the lecture that discusses the original body styles of the first-generation Ford Bronco and listen for the term "sports utility".

- At 02:21-02:30, the speaker explains the three original models of the Bronco, stating, "Three models were originally offered: a wagon, a pickup, which they did call a sports utility in the first brochure, and an unpopular roadster."

- Let me double-check the exact phrasing at 02:24: the speaker says "a pickup, which they did call a sports utility in the first brochure", confirming that the pickup truck body style is the one designated as a "sports utility".

- If the brochure had referred to the wagon or the roadster as the "sports utility", we would expect the speaker to associate the term with those shapes, but he explicitly links it to the pickup — so we must look for the pickup silhouette.

- The pickup truck body style is represented by the silhouette in Region B, which corresponds to option D.

Next, let me read the figure:

- Silhouette A represents a station wagon or SUV body style.

- Silhouette B represents a standard pickup truck body style with an open cargo bed.

- Silhouette C represents a two-door convertible or roadster body style.

- Silhouette D represents a standard four-door sedan body style.

- Matching the video's rule that the "pickup" was called the "sports utility", we find that silhouette B, which is the pickup truck, is the correct match.

- This means Region B is the correct region, which corresponds to option D.

Next, let me analyze the options:

- A. Region A — this represents the wagon model.

- B. Region C — this represents the roadster model, which was unpopular.

- C. Region D — this represents a sedan.

- D. Region B — this represents the pickup model, which the first brochure called a "sports utility".

Finally, the correct option is D. </think>

D

F.19 B03 / Cross-Scenario Transfer

Question, evidence, and training answer

Question. A user needs a drone that can achieve at least 26 minutes of actual flight time in real-world conditions. Based on the video's realistic flight time estimates, which drone should the user choose?

Options. A. Only Drone A B. Only Drone B C. Both drones D. Neither drone

Training answer. B: Only Drone B



Source video, 04:15

Drone	Advertised Flight Time
Drone A	27 minutes
Drone B	31 minutes

New figure

Upstream reasoning steps.
1. [04:04-04:26] Recall the video's realistic flight time estimates: Mavic Pro ˜24-25 min, Mavic 2 Pro ˜28 min.
2. Match the drones in the figure to the models: Drone A (27 min advertised) corresponds to the Mavic Pro; Drone B (31 min advertised) corresponds to the Mavic 2 Pro.
3. Apply the realistic estimates: Drone A's realistic flight time is ˜24-25 min, which is below the 26 min requirement. Drone B's realistic flight time is ˜28 min, which meets the 26 min requirement.
4. Conclude that only Drone B meets the requirement.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify which drone can achieve at least 26 minutes of actual flight time in real-world conditions based on the realistic estimates given in the video. The crux is comparing the realistic flight times of Drone A and Drone B to the 26-minute threshold.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the realistic flight time estimates of the drones and watch the flight time specifications on screen. - At 04:03, the video shows the Mavic Pro flying as the speaker says, "The original Mavic Pro has a flight time of 27 minutes, which you will not get as much as 27 minutes, that's just unrealistic. You probably get about 24, 25 minutes maybe." - At 04:17, the video shows the Mavic 2 Pro flying as the speaker says, "And the Mavic 2 Pro says you will get about 31 minutes. So you may get 28, 27 minutes with that." - If Drone A were to meet the requirement, its realistic flight time would need to be at least 26 minutes, but the video estimates it at only 24 to 25 minutes, meaning it falls short.

Next, let me read the figure:

- The table shows "Drone A" has an "Advertised Flight Time" of "27 minutes". - The table shows "Drone B" has an "Advertised Flight Time" of "31 minutes". - Based on the video, Drone A corresponds to the Mavic Pro with an advertised flight time of 27 minutes, giving a realistic flight time of 24 to 25 minutes. - Based on the video, Drone B corresponds to the Mavic 2 Pro with an advertised flight time of 31 minutes, giving a realistic flight time of 27 to 28 minutes. - Comparing the realistic flight times to the user's requirement of at least 26 minutes: Drone A (24-25 minutes) < 26 minutes, so it does not qualify. - Comparing Drone B's realistic flight time: Drone B (27-28 minutes) >= 26 minutes, so it qualifies. - Only Drone B meets the requirement, which corresponds to option B.

Next, let me analyze the options:

- A. Only Drone A — Drone A's realistic flight time is under 26 minutes. - B. Only Drone B — Drone B's realistic flight time is 27 to 28 minutes, which meets the requirement. - C. Both drones — Drone A fails the requirement. - D. Neither drone — Drone B successfully meets the requirement.

Finally, the correct option is B. </think>

B

F.20 B04 / MULTI-STEP QUANTITATIVE REASONING

Question, evidence, and training answer

Question. A retailer has the items shown in the figure. He also has 5 individual pillow covers from another source. He wants to create gift packs each containing 2 cushion covers and 1 pillow cover. How many complete gift packs can he make?

Options. A. 4 B. 7 C. 6 D. 10

Training answer. B: 7



Source video, 05:23

Item	Quantity
Cushion Set (2+5)	4
Double Bedsheet Set	1

New figure

Upstream reasoning steps.
1. [08:01-08:08] Recall from the video that a Cushion Set (2+5) contains 5 cushion covers.
2. Compute total cushion covers from the figure: 4 sets x 5 = 20 cushion covers.
3. [05:20-05:27] Recall from the video that a Double Bedsheet Set contains 2 pillow covers.
4. Compute total pillow covers from the figure: 1 set x 2 = 2 pillow covers.
5. Add the 5 individual pillow covers: total pillow covers = 2 + 5 = 7.
6. Each gift pack requires 2 cushion covers and 1 pillow cover. The number of packs is limited by the smaller of floor(20/2)=10 and floor(7/1)=7, so 7 packs.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to find the maximum number of complete gift packs that can be made using the items shown in the table and some additional pillow covers. The crux is determining the exact number of cushion covers and pillow covers contained in each set listed in the table by watching the video.

First, let me scan the video for the segments relevant to the question:

- I need to find the segments in the video where the shopkeeper explains the contents of a double bedsheet set and a cushion set (2+5).

- At 05:25-05:28, the shopkeeper explains that a double bedsheet set comes with two pillow covers and one bedsheet, saying "Double 300 ka set aayega, do pillow cover ek bedsheet."

- If a double bedsheet set only contained one pillow cover instead of two, the total number of pillow covers would be 1 + 5 = 6, which would limit the gift packs to 6, but the shopkeeper states it contains "do pillow cover".

- At 07:53-08:08, the shopkeeper describes the cushion set (2+5), stating "5 cushion aate hain, 2 load aate hain" and specifying that it contains "5 cushion ke cover aate hain" and "2 load cover aate hain."

Next, let me read the figure:

- The figure shows a table with two items: "Cushion Set (2+5)" with a quantity of 4, and "Double Bedsheet Set" with a quantity of 1.

- Wait, let me re-read the question: it states the retailer also has 5 individual pillow covers from another source, which must be added to the pillow covers obtained from the double bedsheet set.

- Total cushion covers = 4 * 5 = 20 cushion covers.

- Pillow covers from the bedsheet set = 1 * 2 = 2 pillow covers.

- Total pillow covers = 2 + 5 = 7 pillow covers.

- Packs possible from cushion covers = 20 / 2 = 10 packs.

- Packs possible from pillow covers = 7 / 1 = 7 packs.

- Complete packs = min(10, 7) = 7 packs, which corresponds to option B.

Complete synthetic thinking and final answer (continued)

Next, let me analyze the options:

- A. 4 — we can make more than this number.

- B. 7 — the 7 available pillow covers limit the total complete gift packs to this amount, as we have 20 cushion covers.

- C. 6 — this would leave one pillow cover unused.

- D. 10 — we would need 10 pillow covers, but only have 7.

Finally, the correct option is B. </think>

B

F.21 B05 / Comparative Reasoning

Question, evidence, and training answer

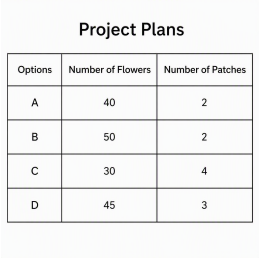
Question. For the four project plans shown in the figure, which one would have taken less total time than Judi Hu's original project (which took about 9 hours)? Use the flower-making and patch-sewing rates that Judi experienced during her project.

Options. A. Plan C B. Plan A C. Plan B D. Plan D

Training answer. B: Plan A



Source video, 04:48



New figure

Upstream reasoning steps.

1. [04:45-04:52] Recall that Judi stated each flower takes about 8 minutes to make.
2. [06:38-06:45] Recall that the on-screen text indicated each embroidered patch takes 1.5 hours (90 minutes) to sew.
3. Compute total time for Plan A: 40 flowers x 8 min = 320 min = 5.33 h; 2 patches x 90 min = 180 min = 3 h; sum = 500 min = 8.33 h (< 9 h).
4. Compute total time for Plan B: 50 x 8 = 400 min = 6.67 h; 2 x 90 = 180 min = 3 h; sum = 580 min = 9.67 h (> 9 h).
5. Compute total time for Plan C: 30 x 8 = 240 min = 4 h; 4 x 90 = 360 min = 6 h; sum = 600 min = 10 h (> 9 h).
6. Compute total time for Plan D: 45 x 8 = 360 min = 6 h; 3 x 90 = 270 min = 4.5 h; sum = 630 min = 10.5 h (> 9 h).
7. Only Plan A's total (8.33 h) is less than the original 9 h, so Plan A is correct.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know which of the four project plans shown in the figure would have taken less total time than Judi Hu's original project. The crux is calculating the total time for each plan using the flower-making and patch-sewing rates established in the video.

First, let me scan the video for the segments relevant to the question:

- I need to find the segments in the video where Judi Hu shares the time it takes to make a single flower and to sew a single embroidered patch.

- At 04:49-04:54, Judi says "I took about 8 minutes for one flower until I started getting better at it", establishing the base rate for flower-making.

Complete synthetic thinking and final answer (continued)

- Wait, let me verify the two rates against their sources: the 8-minute flower rate comes from Judi's spoken statement at 04:49-04:54, while the on-screen text at 06:34-06:36 displays "1 AND 1/2 HOURS FOR 1 PATCH", which means each patch takes 1.5 hours, or 90 minutes.

- If a patch took only 1 hour (60 minutes) instead of 1.5 hours (90 minutes) at 06:36, the original project's total time would be calculated as 45 * 8 = 360 minutes for flowers plus 2 * 60 = 120 minutes for patches, giving 480 minutes (8 hours) instead of the 9 hours mentioned in the prompt — so a patch must indeed take 90 minutes.

- At 06:30-06:33, the on-screen text shows the calculation for the original project's flowers: "15 X 3 = 45 FLOWERS" and "45 X 8 = 360 MINUTES", which equals 6 hours.

- At 06:34-06:36, the on-screen text shows "2 EMBROIDERED PATCHES" and "1 AND 1/2 HOURS FOR 1 PATCH", which means the patch-sewing time is 2 * 90 = 180 minutes, or 3 hours.

- Combining these rates, the original project's total time is 360 + 180 = 540 minutes, which is exactly 9 hours.

Next, let me read the figure:

- The figure shows a table titled "Project Plans" with columns "Options", "Number of Flowers", and "Number of Patches".

- Row "A" lists "40" flowers and "2" patches.

- Row "B" lists "50" flowers and "2" patches.

- Row "C" lists "30" flowers and "4" patches.

- Row "D" lists "45" flowers and "3" patches.

- For Plan A, the total time is (40 * 8) + (2 * 90) = 320 + 180 = 500 minutes.

- For Plan B, the total time is (50 * 8) + (2 * 90) = 400 + 180 = 580 minutes; for Plan C, it is (30 * 8) + (4 * 90) = 240 + 360 = 600 minutes; and for Plan D, it is (45 * 8) + (3 * 90) = 360 + 270 = 630 minutes.

- Comparing these to the original project's 540 minutes, only Plan A takes less total time (500 minutes), which corresponds to Option B.

Next, let me analyze the options:

- A. Plan C — this takes 600 minutes, exceeding the original.

- B. Plan A — this plan takes 500 minutes, which is less than the original project's 540 minutes.

- C. Plan B — this plan takes 580 minutes, which is more than the original project's 540 minutes.

- D. Plan D — this takes 630 minutes, exceeding the original.

Finally, the correct option is B. </think>

B

F.22 B06 / DESIGN / OPTIMIZATION

Question, evidence, and training answer

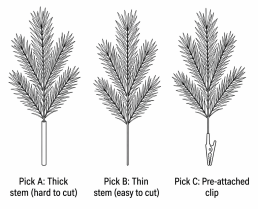
Question. For the three decorative picks shown in the figure, which one should be attached to the existing swag using the method demonstrated in the video (bending the stem back and tying it with existing branches)?

Options. A. Pick C: Pre-attached clip B. Pick B: Thin stem (easy to cut) C. Pick A: Thick stem (hard to cut)

Training answer. C: Pick A: Thick stem (hard to cut)



Source video, 00:53



New figure

Upstream reasoning steps.
1. [00:48-00:59] Recall the video's decision rule: When the stem is too thick to cut easily, use the bending method (bend the stem back, hide it, and tie with existing branches). For thinner stems, cutting is preferred.
2. From the figure, identify which pick has a thick stem that is hard to cut. The figure labels Pick A as 'Thick stem (hard to cut)'.
3. Apply the rule: Pick A is the one that should be attached using the bending-and-tying method.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify which of the three decorative picks shown in the figure should be attached to the swag using the bending and tying method shown in the video. The crux is matching the physical characteristics of the pick the woman describes and uses in the video to the descriptions of the picks in the diagram.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the woman's description of the pick's stem and watch how she manipulates and attaches it to the swag. - At 0:47-0:55, the woman holds up a pinecone pick and explains that while cutting the stem would be ideal, "this is so thick and hard to cut, you almost need bolt cutters to cut it." - If the pick had a thin stem that was easy to cut, she would have cut it as she noted that would be "ideal" at 0:50, but instead she has to bend it because of its thickness. - Wait, let me double check her exact words at 1:00-1:08: she says "So I'm just gonna bend it and attempt to hide it within my swag," and she proceeds to bend the thick green wire stem back on itself. - At 1:09-1:25, she tucks the bent stem into the swag and wraps the existing pine branches around it to secure it in place. - The video establishes that the pick used has a stem that is too thick and hard to cut, requiring it to be bent back and tied with existing branches.

Next, let me read the figure:

- The figure shows three decorative picks: "Pick A: Thick stem (hard to cut)", "Pick B: Thin stem (easy to cut)", and "Pick C: Pre-attached clip". - Pick A features a thick, solid green stem labeled "Thick stem (hard to cut)". - Pick B features a very thin wire stem labeled "Thin stem (easy to cut)". - Pick C features a clip at the base of the stem labeled "Pre-attached clip". - Applying the video's rule, the pick with the "so thick and hard to cut" stem that must be bent corresponds directly to "Pick A: Thick stem (hard to cut)", which is listed as option C.

Next, let me analyze the options:

- A. Pick C: Pre-attached clip — This pick has a clip. - B. Pick B: Thin stem (easy to cut) — This pick has a thin stem, which she says is ideal to cut. - C. Pick A: Thick stem (hard to cut) — This matches the thick stem that she bends and ties.

Finally, the correct option is C. </think>

C

F.23 B07 / Error Critique

Question, evidence, and training answer

Question. A reviewer on a tech forum claims: 'The Garmin Nuvi 2555 LMT has a glossy screen that is hard to read in bright sunlight. To fix this, you should apply a matte screen protector.' What is the specific flaw in this recommendation?

Options. A. A matte screen protector will interfere with the touchscreen's sensitivity. B. The screen is too small for a screen protector to be effective at reducing glare. C. The screen is already anti-glare, so a matte protector is unnecessary and may reduce clarity. D. The reviewer's recommendation is correct; there is no flaw.

Training answer. C: The screen is already anti-glare, so a matte protector is unnecessary and may reduce clarity.



Source video, 02:12

Upstream reasoning steps.
1. [02:08-02:16] Recall from the video that the Garmin Nuvi 2555 LMT screen is described as 'really meant for anti-glare' and compared to a Kindle screen, meaning it is matte and anti-glare by design.
2. The reviewer's claim that the screen is glossy is false based on the video-derived knowledge. Therefore, the recommendation to apply a matte screen protector to reduce glare is based on a false premise.
3. The specific flaw is that the screen is already anti-glare, so the proposed solution is unnecessary and could degrade the screen's performance.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the specific flaw in a reviewer's recommendation to apply a matte screen protector to the Garmin Nuvi 2555 LMT. The crux is determining whether the device's screen is already designed to prevent glare.

First, let me scan the video for the segments relevant to the question:

- I need to listen for the narrator's description of the device's screen surface and watch the screen as the protective film is removed.

- At 02:05, the narrator holds up the Garmin Nuvi 2555 LMT with the printed screen sticker still attached.

- At 02:09, the narrator peels off the printed screen sticker, revealing the actual screen underneath.

- If the reviewer's recommendation is correct and there is no flaw, the screen would have to be glossy and highly reflective once the sticker is removed, but the surface is matte and non-reflective.

- At 02:16-02:21, as the narrator rubs his finger across the display, he says, "you're gonna see a really nice screen, it almost looks like a Kindle screen 'cause it's really meant for anti-glare, and it's really, really nice."

- Since the device already features a built-in anti-glare screen similar to a Kindle, applying a matte screen protector is unnecessary, which points directly to option C: The screen is already anti-glare, so a matte protector is unnecessary and may reduce clarity.

Next, let me analyze the options:

- A. A matte screen protector will interfere with the touchscreen's sensitivity. — The video does not mention touch sensitivity issues.

- B. The screen is too small for a screen protector to be effective at reducing glare. — Screen size does not dictate glare reduction.

- C. The screen is already anti-glare, so a matte protector is unnecessary and may reduce clarity. — The narrator explicitly states the screen is anti-glare, making an additional matte protector redundant.

- D. The reviewer's recommendation is correct; there is no flaw. — The screen is already anti-glare, so this claim is false.

Finally, the correct option is C. </think>

C

F.24 B08 / PROCEDURE / PLANNING

Question, evidence, and training answer

Question. You have a Tashtego 2.0 toiletry roll that is stained after a camping trip. You have removed all toiletries. What is the correct next step according to the video's demonstrated cleaning procedure?

Options. A. Place the bag in a washing machine and run a normal cycle. B. Spot clean the stained area with a damp cloth. C. Hand wash the bag with mild soap and let it air dry. D. Apply a new coat of wax to the bag before washing.

Training answer. A: Place the bag in a washing machine and run a normal cycle.



Source video, 07:05

Upstream reasoning steps.
1. [07:00-07:11] Recall that the video visually demonstrates the Tashtego bag being tossed into a washing machine for cleaning.
2. Since the bag is stained and all toiletries have been removed, the correct next step is to machine wash it as shown in the video.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to know the correct next step to clean a stained Tashtego 2.0 toiletry roll based on the video. The crux is identifying the cleaning method demonstrated in the video for this specific product.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment showing the "Tashtego 2.0" and watch for any cleaning demonstrations.

- At 07:00, the video shows the toiletry roll being rinsed under running water in a sink while the narrator explains that "Last but not least, Tashtego is made of number 10 waxed cotton and coated with a water repellent wax so you can get it wet without soaking your toiletries".

- At 07:03, the video shows the toiletry roll being placed directly into a top-loading washing machine.

- Wait, is the appliance shown at 07:03 a sink or a washing machine? Looking closely, it is a top-loading washing machine with an agitator and water inside, into which the bag is dropped.

- Let me re-read the narration to see if machine washing is mentioned: the narrator says "you can get it wet without soaking your toiletries", but the visual of dropping it into the washer is what demonstrates the actual cleaning step.

- If the demonstrated cleaning procedure were hand washing the bag with mild soap, the video would have to show the bag being scrubbed by hand and left to air dry, but instead it is placed directly into a washing machine at 07:03.

- The visual demonstration of dropping the bag into the washing machine at 07:03 confirms that placing the bag in a washing machine is the correct next step, which corresponds to option A: Place the bag in a washing machine and run a normal cycle.

Next, let me analyze the options:

- A. Place the bag in a washing machine and run a normal cycle. — This matches the visual demonstration of the bag being placed in a washing machine.

- B. Spot clean the stained area with a damp cloth. — This method is not shown in the video.

- C. Hand wash the bag with mild soap and let it air dry. — Although water is run over it, the bag is ultimately placed in a washing machine.

- D. Apply a new coat of wax to the bag before washing. — No waxing process is demonstrated prior to washing.

Finally, the correct option is A. </think>

A

F.25 B09 / CROSS-DOMAIN TRANSFER

Question, evidence, and training answer

Question. A firefighter is learning to spin a long weighted hose nozzle in a forward twirl, exactly as the bo staff forward twirl taught in the video. Which hand orientation should he use to initiate the spin?

Options. A. Lead with the pinky, palm down B. Lead with the thumb, palm up (thumb dives down) C. Lead with the thumb, palm down D. Lead with the pinky, palm up

Training answer. B: Lead with the thumb, palm up (thumb dives down)



Source video, 01:25

Upstream reasoning steps.
1. [01:15-01:35] Recall the video's instruction for the forward twirl: the instructor says 'I'm going to dive my thumb across my body and down' and then 'bring that palm up'. This establishes that the forward twirl is initiated by leading with the thumb (diving it down) and having the palm face up.
2. Apply this to the firefighter's scenario: he wants to perform a forward twirl with the hose nozzle, so he must use the same hand orientation – lead with the thumb (dive it down) and keep the palm up.

Task label: source-assigned.

Complete synthetic thinking and final answer

<think> Got it, let's tackle the problem.

The user wants to identify the correct hand orientation to initiate a forward twirl with a bo staff as demonstrated in the video. The crux is determining whether the spin begins with the thumb or pinky leading, and whether the palm should be facing up or down.

First, let me scan the video for the segments relevant to the question:

- I need to find the segment where the instructor demonstrates and explains the hand orientation and movement for the forward twirl.

- At 01:13, the instructor introduces the "forward twirl" and prepares to demonstrate it.

- At 01:15, he places the center of the staff in his palm to begin the movement.

- At 01:19, he grabs the staff and explains that he is going to "dive my thumb across my body and down."

- At 01:25, he continues the motion, explaining to "bring that palm up and come across the body and down" as he completes the rotation.

- At 01:53, he demonstrates the handoff to the left hand, stating "palm up, I grab it with palm up with my left hand" before diving to the other side.

- At 02:16, he repeats the motion on both sides, showing that the thumb leads the dive on each rotation.

- Let me re-listen at 02:27: the instructor says "palm goes up, palm goes up, dive" to emphasize the handoff and initiation sequence.

- If the forward twirl were initiated by leading with the pinky and keeping the palm down, the instructor would have to say to lead with the pinky when demonstrating the forward twirl at 01:19, but he explicitly says "dive my thumb across my body and down" and reserves the pinky-lead for the backward twirl.

- At 03:26, he introduces the "backwards" twirl, instructing to "think your pinky now" and "drive up in the air."

- At 03:35, he confirms the hand orientation for the backward twirl is "palm down now."

- The instructor's breakdown shows that the forward twirl is initiated by grabbing palm up and leading with the thumb diving down, which corresponds to option B.

Next, let me analyze the options:

- A. Lead with the pinky, palm down — this orientation is used for the backward twirl.

- B. Lead with the thumb, palm up (thumb dives down) — this matches the forward twirl initiation where the hand grabs palm up and the thumb dives down.

- C. Lead with the thumb, palm down — incorrect grab hand orientation.

Complete synthetic thinking and final answer (continued)

- D. Lead with the pinky, palm up — incorrect on both counts.

Finally, the correct option is B. </think>

B