OmniJigsaw: Enhancing Omni-Modal Reasoning via Modality-Orchestrated Reordering

Yiduo Jia[1]*, **Muzhi Zhu**[1]*, **Hao Zhong**[1], **Mingyu Liu**[1], **Yuling Xi**[1]†, **Hao Chen**[1]†, **Bin Qin**[2], **Yongjie Yang**[2], **Zhenbo Luo**[2], **Chunhua Shen**[1,3]†

[1] State Key Lab of CAD & CG, Zhejiang University [2] Xiaomi Inc. [3] Ant Group

Abstract

Omni-modal reasoning is pivotal to realizing artificial general intelligence, yet its advancement is critically constrained by the limited availability of large-scale, human-annotated data for complex reasoning. Motivated by this, we propose OmniJigsaw, a self-supervised reinforcement learning framework built upon a temporal reordering proxy task. Centered on the chronological reconstruction of shuffled audio-visual clips, we successively investigate three modality orchestration strategies: (i) Joint Modality Integration (JMI), which retains the complete visual and auditory streams; (ii) Sample-level Modality Selection (SMS), which selects the dominant modality through a global decision mechanism; and (iii) Clip-level Modality Masking (CMM), which adaptively masks modalities at the clip granularity. Our analysis reveals a "bi-modal shortcut phenomenon" in JMI and demonstrates that fine-grained CMM mitigates this issue while outperforming SMS. Incorporating lightweight puzzle-quality curation and verifiable rewards, OmniJigsaw yields substantial gains across 15 video, audio, and omni-modal benchmarks, with CMM achieving the strongest overall performance, validating its effectiveness as a self-supervised paradigm for omni-modal learning. Our code is released at: https://github.com/aim-uofa/OmniJigsaw.

1 Introduction

The ultimate goal of Artificial General Intelligence (AGI) is to develop intelligent agents capable of comprehensively processing omni-modal inputs, spanning video, audio, and text, to perform complex reasoning, decision-making, and planning [1, 2]. As a pivotal stride toward this vision in the textual domain, reinforcement learning (RL) post-training [3, 4, 5] has recently driven remarkable breakthroughs in large language models (LLMs), empowering them with strong reasoning capabilities to solve complex mathematical problems [6, 7, 8] and generate high-quality, functional code [9, 10]. Despite these advancements, how to effectively explore and enhance the reasoning capabilities of omni-modal models [11, 12] remains an open and challenging problem [13, 14, 15].

A primary bottleneck impeding the extension of these RL-driven successes to omni-modal reasoning lies in acquiring massive, high-quality annotated data and providing effective supervisory signals. In text-only domains such as mathematics or coding, it is relatively straightforward to generate large-scale problem instances and provide verifiable, deterministic feedback for RL optimization [16]. In contrast, for omni-modal understanding [17, 18, 19, 20], collecting an equivalent volume of omni-modal data that requires genuine collaborative cross-modal reasoning is prohibitively expensive and labor-intensive [21, 22, 23]. Driven by these challenges, we explore a central question in this work: *Can we identify a suitable proxy task that effectively leverages massive unannotated omni-modal*

*Equal contributions.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

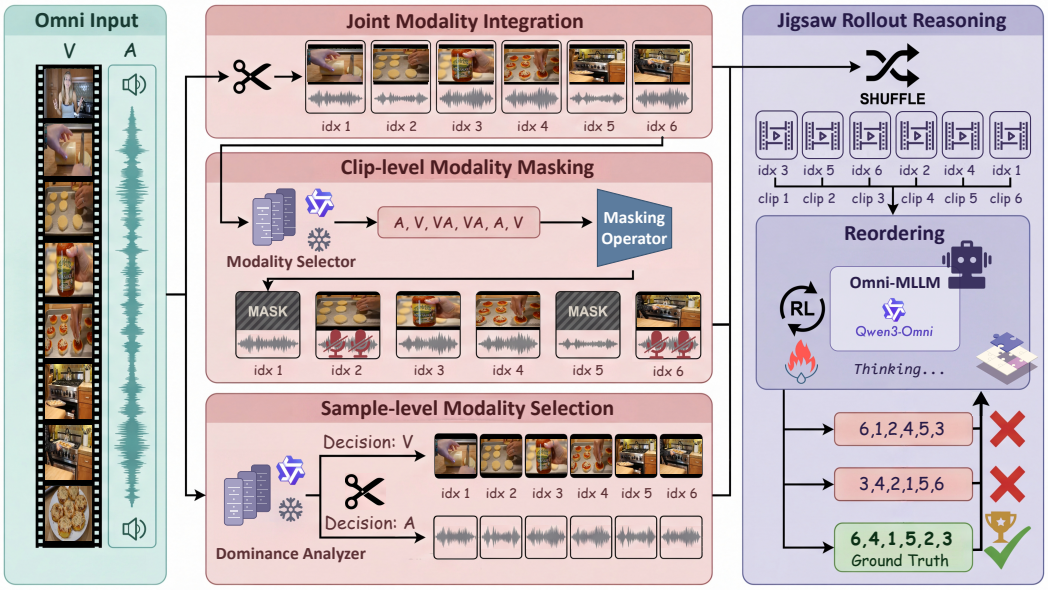


Figure 1: **OmniJigsaw framework for self-supervised omni-modal RL post-training.** Three modality orchestration strategies are illustrated: Joint Modality Integration (JMI, top) exposing both audio-visual streams; Sample-level Modality Selection (SMS, bottom) prioritizing the primary informative modality; and Clip-level Modality Masking (CMM, center) imposing a clip-wise information bottleneck. All strategies share the same temporal reordering objective.

data to enhance the versatile reasoning capabilities of omni-modal models through self-supervised post-training?

Toward this goal, we investigate temporal reordering as a self-supervised objective for omni-modal reasoning, where the model is trained to recover the original chronological order from a shuffled sequence of synchronized audio-visual clips. Adapting this objective to the omni-modal setting requires a proxy task that jointly engages visual and auditory understanding. This calls for strategic orchestration of how visual and auditory streams are exposed during reordering, thereby guiding the model to integrate cross-modal cues.

Initially, we examine a straightforward strategy, Joint Modality Integration (JMI), which provides full access to both visual and auditory streams. While JMI yields certain improvements in omni-modal reasoning, its gains counter-intuitively fall short of those obtained by uni-modal jigsaws [24] (*i.e.*, video-only or audio-only reordering). We refer to this empirical pattern as the "bi-modal shortcut phenomenon": naive joint audio-visual exposure achieves higher proxy task performance but transfers less effectively to downstream omni-modal reasoning than modality-specific reordering. We hypothesize that this occurs because modality-specific cues independently suffice to solve the task, triggering a modal short-circuit that allows the model to take the path of lower resistance, thereby hindering the robust cultivation of reasoning for the weaker modality.

Building upon this observation, we further propose two advanced modality orchestration strategies: (1) Sample-level Modality Selection (SMS), which uses the model as a global dominance analyzer to identify the primary modality and mitigate interference from less informative streams; and (2) Clip-level Modality Masking (CMM), which adaptively masks the less salient modality within each clip to construct a cross-modal information bottleneck. By advancing modality selection from sample-level arbitration to clip-level masking, CMM encourages the model to stitch together fragmented heterogeneous cues across video and audio when reconstructing the global timeline.

We further uncover that the success of such jigsaw-based proxy tasks is sensitive to puzzle quality: ill-posed clips with weak temporal directionality can provide noisy self-supervision. This motivates a lightweight curation pipeline that converts the conventionally annotation-heavy data preparation [21, 22, 23] into an automated filtering workflow and enables OmniJigsaw to adapt efficiently to massive

unannotated omni-modal data. Additionally, we design a verifiable composite reward tailored to temporal reordering, which provides stable optimization signals for RL post-training without relying on external reward models.

Extensive experiments demonstrate that OmniJigsaw yields consistent enhancements across uni-modal video reasoning, audio comprehension, and collaborative omni-modal reasoning. Notably, CMM boosts the strong Qwen3-Omni-30B-A3B-Instruct [12] baseline by absolute gains of **+4.38** on MLVU-Test [25], **+2.50** on MMAR [26], and **+1.70** on OmniVideoBench [27], with consistent gains also observed on the smaller-scale backbone. Comprehensive ablations and analyses further corroborate the efficacy of its key supporting designs.

In summary, our main contributions are:

1) We propose OmniJigsaw, a self-supervised RL post-training framework that leverages modality-orchestrated temporal reordering to enhance omni-modal reasoning without requiring human-annotated reasoning data.

2) We reveal a "bi-modal shortcut phenomenon" inherent in joint modality integration and mitigate it with two modality orchestration strategies: Sample-level Modality Selection (SMS) and the finer-grained Clip-level Modality Masking (CMM); CMM delivers substantial gains across 15 benchmarks on omni-modal models of different scales.

3) We demonstrate the sensitivity of the omni-modal jigsaw proxy task to data quality, reward mechanism, and modality orchestration granularity through targeted ablations and analyses, offering empirical insights for future research in self-supervised omni-modal learning.

2 Related Work

2.1 RL Post-Training for Omni-Modal Understanding

While RL post-training has evolved from prioritizing human intent alignment (*e.g.*, RLHF [3], DPO [4]) to comprehensively strengthening complex reasoning (*e.g.*, RLVR [5]) across textual [5, 6, 7, 8, 28, 9, 10] and visual domains (*e.g.*, VideoChat-R1 [29], RLHF-V [30], Visual-RFT [31], Diffusion-DPO [32]), its potential to simultaneously enhance video and audio reasoning capabilities in omni-modal models remains insufficiently explored [13, 14, 15]. Although specialized video [17, 18, 33] and audio [19, 34, 20] understanding, along with integrated architectures like Qwen-Omni [11, 12], have advanced rapidly with the surging demand for holistic perception in embodied AI [1, 2], current audio-visual enhancements predominantly rely on computationally intensive supervised training with meticulously annotated data (*e.g.*, Video-CoT [21], CoTasks [22], VIDEOP2R [23]), complex auxiliary objectives leveraging external reward models (*e.g.*, Video-VER [35], OmniVideo-R1 [13]), or elaborate multi-stage RL pipelines like Omni-R1 [36]. To bridge this gap without requiring costly manual annotation or architectural complexity, our OmniJigsaw framework introduces a lightweight and verifiable self-supervised proxy task that strategically orchestrates synchronized video and audio streams to concurrently bolster intrinsic omni-modal comprehension and collaborative reasoning capabilities.

2.2 Jigsaw as Self-Supervised Proxy Task

With the evolution of self-supervised learning, identifying effective proxy tasks that distill reliable supervisory signals directly from the inherent structure of data without manual annotation remains a central challenge. In this context, jigsaw-style tasks have gained increasing prominence due to their concise objectives, computational efficiency, and independence from auxiliary generative models. Pioneered by Noroozi and Favaro [37] to compel the learning of object parts and spatial layouts in the static visual domain, this permutation-based philosophy has transcended diverse modalities, extending to temporal order verification for capturing motion dynamics in videos [38], rearranged voxel reconstruction for bolstering spatial reasoning on 3D point clouds [39], permutation language modeling for bidirectional context dependencies in NLP [40], and multimodal puzzles for robust cross-modal alignment in medical imaging [41]. While recent efforts have demonstrated the efficacy of visual-only jigsaw tasks [24], OmniJigsaw surpasses these uni-modal boundaries by introducing fine-grained modality orchestration strategies, designed to further unlock the potential of structural ordering tasks within the generalized video-audio domain.

2.3 Modality Imbalance in Multimodal Learning

Modality imbalance and dominance have been extensively studied in multimodal learning [42, 43]. To mitigate overreliance on a single modality, AMCo [44] adaptively masks features based on modal dependence, while OPM/OGM [45] modulate features and gradients, respectively, according to the discriminative discrepancy between modalities. For MLLMs, MokA [46] combines modality-specific low-rank adaptation with explicit cross-modal interaction to better utilize multimodal information. Departing from these approaches in both the modality dominance mechanism and the intervention level, we focus on self-supervised RL post-training for omni-modal LLMs, identify a bi-modal shortcut induced by the proxy's reward-equivalent solution paths, and propose CMM, which changes clip-level observability according to content and temporal role to reduce the availability of uninterrupted uni-modal paths and strengthen downstream transfer.

3 Method

3.1 OmniJigsaw Formulation & Joint Modality Integration Strategy

We formulate OmniJigsaw as a permutation prediction problem over synchronized audio-visual clips. Let $\mathcal{X} = (\mathcal{V}, \mathcal{A})$ denote a video sample, where $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$ is the visual stream and $\mathcal{A} \in \mathbb{R}^{L}$ is the raw audio waveform. We uniformly divide $\mathcal{X}$ into N non-overlapping clips and remove short boundary spans with $\mathcal{T}_{trim}$ to reduce trivial continuity cues, yielding an ordered sequence $\mathcal{S} = [s_1, s_2, \ldots, s_N]$, where $s_i = (v_i, a_i)$ contains synchronized visual and acoustic signals. Given a random permutation π, the shuffled sequence is $\tilde{\mathcal{S}} = [\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_N]$, with $\tilde{s}_j = s_{\pi^{-1}(j)}$. The model $\mathcal{M}_\theta$ is trained to recover the original chronological order by predicting an index sequence $\hat{\mathbf{y}}$ aligned with the ground-truth permutation $\mathbf{y} = [\pi(1), \pi(2), \ldots, \pi(N)]$:

$$\hat{\mathbf{y}} = \mathcal{M}_\theta(\Phi(\tilde{\mathcal{S}}); \mathcal{I}_{prompt}), \quad \text{s.t.} \quad \hat{\mathbf{y}} \to \mathbf{y}, \tag{1}$$

where $\mathcal{I}_{prompt}$ denotes the task instruction and $\Phi(\cdot)$ specifies how each modality is exposed to the model. Different instantiations of $\Phi(\cdot)$ correspond to distinct modality orchestration strategies under a unified reordering objective. Figure 1 illustrates the overall pipeline of OmniJigsaw and the three modality orchestration strategies. The simplest choice is Joint Modality Integration (JMI), a straightforward strategy that retains both streams for each shuffled clip. With a temporal downsampling operator $\mathcal{D}_T$ applied to the visual stream, JMI is defined as

$$\Phi_{jmi}(\tilde{s}_j) = \mathcal{D}_T(\tilde{s}_j) = (\mathcal{D}_T(v_{\pi^{-1}(j)}), a_{\pi^{-1}(j)}), \tag{2}$$

and the rollout is given by $\mathcal{M}_\theta(\Phi_{jmi}(\tilde{\mathcal{S}}); \mathcal{I}_{jmi})$.

3.2 Sample-level Modality Selection Strategy

Sample-level Modality Selection (SMS) provides coarse access control over modalities at the sample level. Instead of exposing both modalities throughout the puzzle, SMS first selects a single modality for the entire sample and then performs reordering using only that stream. Concretely, the model estimates whether the visual or auditory stream carries richer information and stronger temporal evidence:

$$d = \arg\max_{m\in\{V,A\}} p_\theta(m \mid (\mathcal{D}_T(\mathcal{V}), \mathcal{A}); \mathcal{I}_{judge}). \tag{3}$$

The selected modality is then applied uniformly to all shuffled clips:

$$\Phi_{sms}(\tilde{s}_j|d) = \begin{cases} (\mathcal{D}_T(v_{\pi^{-1}(j)}), \mathbf{0}), & \text{if } d = V, \\ (\mathbf{0}, a_{\pi^{-1}(j)}), & \text{if } d = A. \end{cases} \tag{4}$$

The corresponding rollout is $\mathcal{M}_\theta(\Phi_{sms}(\tilde{\mathcal{S}}|d); \mathcal{I}_{sms})$. Compared with JMI, SMS removes full-modality access which may be redundant, but its sample-level decision assumes that one modality remains sufficient throughout the sequence.

3.3 Clip-level Modality Masking Strategy

As illustrated in Fig. 1 (center), Clip-level Modality Masking (CMM) refines modality access control from the sample level to the clip level through a two-stage process. This design discourages reliance

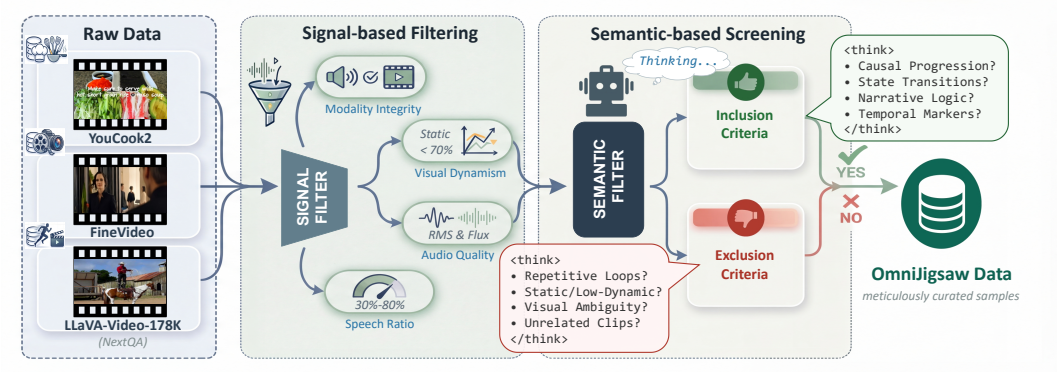


Figure 2: **Puzzle-quality curation pipeline for OmniJigsaw.** Raw videos are first filtered for valid video-audio streams and sufficient low-level dynamics, then screened for semantically recoverable chronological progression.

on any single dominant modality and instead compels the model to exploit temporally distributed cues across clips. In the first stage, a prompted pass of the same model $\mathcal{M}_\theta$ evaluates the information richness and temporal role of each clip in the ordered sequence $\mathcal{S}$, producing a modality selection vector $\mathbf{m} = [m_1, \ldots, m_N]$, where $m_i \in \{V, A, VA\}$ denotes retaining only video, retaining only audio, or retaining both modalities. In the second stage, a masking operator $\mathcal{M}_{cmm}$ keeps the selected modality and replaces the unselected one with a null tensor $\mathbf{0}$. The orchestration function Φ_{cmm} applies this operator to each shuffled clip $\tilde{s}_j$ according to its original clip-level decision $m_{\pi^{-1}(j)}$. The rollout is $\mathcal{M}_\theta(\Phi_{cmm}(\tilde{\mathcal{S}}|\mathbf{m}); \mathcal{I}_{cmm})$, where

$$\Phi_{cmm}(\tilde{s}_j|m_{\pi^{-1}(j)}) = \mathcal{M}_{cmm}(\mathcal{D}_T(\tilde{s}_j)|m_{\pi^{-1}(j)}). \tag{5}$$

By varying modality availability across clips, CMM makes it harder to solve the full puzzle through a single cue stream. The model must instead stitch together local visual and acoustic evidence to recover the target permutation. Details of the modality selection procedure are provided in Appendix C.1.

3.4 Puzzle Curation and Verifiable Reward

We find temporal reordering provides a useful self-supervised signal only when the constructed puzzle is well posed: the clips should contain valid audio-visual streams and a recoverable chronological progression. Therefore, we apply a lightweight curation procedure before RL post-training. As shown in Fig. 2, we first remove degenerate samples through low-level signal checks on modal integrity, visual dynamics, and acoustic activity, and then perform semantic directionality screening to retain videos with directionally identifiable temporal evolution while filtering out repetitive, static, or weakly ordered content. Detailed thresholds, retained-data statistics, and the screening prompt are provided in Appendix C.2, C.3.1, and F.1, respectively.

Given a curated puzzle, the reward is derived directly from the ground-truth temporal order and does not rely on external reward models. We use a compact composite reward that combines task accuracy with lightweight regularization terms:

$$R_{tot} = R_{rep} + R_{fmt} + \lambda(\text{acc}) \cdot (w_{pos} \cdot R_{pos} + w_{cont} \cdot R_{cont}). \tag{6}$$

The main supervision comes from two verifiable components. The positional term evaluates whether each clip is placed at the correct location, while the continuity term rewards correctly ordered adjacent clip pairs regardless of their absolute positions:

$$R_{pos} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i), \quad R_{cont} = \frac{1}{N-1}\sum_{i=1}^{N-1} \mathbb{I}\big[(\hat{y}_i, \hat{y}_{i+1}) \in \{(y_j, y_{j+1})\}_{j=1}^{N-1}\big]. \tag{7}$$

The auxiliary terms R_{rep} and R_{fmt} are used to discourage degenerate generations and ensure parseable outputs, while $\lambda(\text{acc})$ emphasizes fully correct chronological reconstruction during optimization. We keep these implementation choices fixed across all modality orchestration strategies and provide concrete details in Appendix C.3.

Table 1: **Performance comparison on video reasoning.** Results are reported on the Qwen2.5-Omni-3B [11] and Qwen3-Omni-30B [12] backbones under two inference modes ($w/$ and w/o audio). **Bold**: best within each backbone group; green: gains over the corresponding baseline (gray row). Related methods: Video-R1 [47], Omni-R1 [36], HumanOmniV2 [48]; benchmarks: AoTBench [49], TUNA-Bench [50], TempCompass [51], Video-TT [52], Video-Holmes [53], MLVU/MLVU-Test [25], Video-MME [54].

Method	AoT Bench	TUNA -Bench	Temp Compass	Video -TT	Video -Holmes	MLVU -Test	Video -MME	MLVU
Infer w/ Audio								
Omni-R1	52.09	53.84	63.10	36.80	40.72	52.59	63.20	65.69
HumanOmniV2	48.58	49.16	63.86	40.30	42.90	49.00	65.60	66.70
Qwen2.5-Omni-3B	50.20	49.86	62.85	32.30	36.04	44.82	56.70	57.50
+VideoJigsaw	52.60 ↑2.40	51.05 ↑1.19	64.18 ↑1.33	33.40 ↑1.10	37.56 ↑1.52	46.41 ↑1.59	58.30 ↑1.60	59.61 ↑2.11
+OmniJigsaw (JMI)	51.93 ↑1.73	48.95 ↓0.91	63.80 ↑0.95	32.50 ↑0.20	36.69 ↑0.65	44.42 ↓0.40	58.10 ↑1.40	58.51 ↑1.01
+OmniJigsaw (SMS)	52.76 ↑2.56	51.26 ↑1.40	65.44 ↑2.59	33.40 ↑1.10	38.81 ↑2.77	45.62 ↑0.80	58.30 ↑1.60	59.61 ↑2.11
+OmniJigsaw (CMM)	**53.66** ↑3.46	**51.40** ↑1.54	**66.46** ↑3.61	**36.20** ↑3.90	**39.74** ↑3.70	**47.81** ↑2.99	**58.90** ↑2.20	**61.50** ↑4.00
Qwen3-Omni-30B	64.88	62.57	70.63	44.30	50.84	57.97	72.10	70.01
+VideoJigsaw	67.45 ↑2.57	63.41 ↑0.84	**72.28** ↑1.65	44.90 ↑0.60	51.99 ↑1.15	60.16 ↑2.19	72.90 ↑0.80	71.90 ↑1.89
+OmniJigsaw (JMI)	66.83 ↑1.95	62.78 ↑0.21	71.08 ↑0.45	44.70 ↑0.40	50.24 ↓0.60	58.76 ↑0.79	72.90 ↑0.80	71.39 ↑1.38
+OmniJigsaw (SMS)	68.12 ↑3.24	65.15 ↑2.58	72.03 ↑1.40	45.80 ↑1.50	52.26 ↑1.42	61.75 ↑3.78	72.90 ↑0.80	**72.63** ↑2.62
+OmniJigsaw (CMM)	**68.90** ↑4.02	**65.29** ↑2.72	72.03 ↑1.40	**46.10** ↑1.80	**52.53** ↑1.69	**62.35** ↑4.38	**73.10** ↑1.00	72.26 ↑2.25
Infer w/o Audio								
Video-R1	52.60	55.94	70.00	40.60	42.13	52.39	62.80	67.07
Omni-R1	51.03	52.65	62.78	37.00	38.76	51.00	60.30	64.86
HumanOmniV2	47.35	49.44	63.10	38.20	38.87	48.21	61.20	66.65
Qwen2.5-Omni-3B	47.96	47.56	63.86	32.10	32.44	44.82	52.40	58.14
+VideoJigsaw	50.75 ↑2.79	51.12 ↑3.56	64.81 ↑0.95	32.70 ↑0.60	34.40 ↑1.96	46.02 ↑1.20	54.00 ↑1.60	59.98 ↑1.84
+OmniJigsaw (JMI)	48.97 ↑1.01	49.51 ↑1.95	64.68 ↑0.82	32.20 ↑0.10	33.10 ↑0.66	44.62 ↓0.20	53.80 ↑1.40	58.97 ↑0.83
+OmniJigsaw (SMS)	50.42 ↑2.46	50.91 ↑3.35	65.13 ↑1.27	32.80 ↑0.70	34.89 ↑2.45	45.62 ↑0.80	54.70 ↑2.30	59.75 ↑1.61
+OmniJigsaw (CMM)	**50.92** ↑2.96	**51.54** ↑3.98	**65.44** ↑1.58	**33.50** ↑1.40	**36.80** ↑4.36	**46.41** ↑1.59	**55.30** ↑2.90	**61.91** ↑3.77
Qwen3-Omni-30B	63.32	63.62	70.70	43.80	46.60	58.76	67.90	70.98
+VideoJigsaw	66.22 ↑2.90	64.39 ↑0.77	71.52 ↑0.82	44.90 ↑1.10	47.90 ↑1.30	61.55 ↑2.79	68.90 ↑1.00	73.14 ↑2.16
+OmniJigsaw (JMI)	65.83 ↑2.51	63.90 ↑0.28	71.39 ↑0.69	44.80 ↑1.00	47.47 ↑0.87	59.76 ↑1.00	68.80 ↑0.90	72.45 ↑1.47
+OmniJigsaw (SMS)	66.39 ↑3.07	64.73 ↑1.11	71.52 ↑0.82	45.90 ↑2.10	47.96 ↑1.36	60.76 ↑2.00	69.00 ↑1.10	72.17 ↑1.19
+OmniJigsaw (CMM)	**66.83** ↑3.51	**66.20** ↑2.58	**72.34** ↑1.64	**46.50** ↑2.70	**48.29** ↑1.69	**62.75** ↑3.99	**69.30** ↑1.40	**73.46** ↑2.48

4 Experiments

4.1 Implementation Details

We instantiate OmniJigsaw on two representative open-source omni-modal backbones of distinct parameter scales, Qwen2.5-Omni-3B [11] and Qwen3-Omni-30B-A3B-Instruct [12], using GRPO [55] for post-training. The training dataset, denoted as OmniJigsaw-8K, is constructed from YouCook2 [56], FineVideo [57], and LLaVA-Video-178K [58] using the curation pipeline introduced in Sec. 3.4, with Qwen2.5-VL-7B-Instruct [59] as the semantic assessor. All modality orchestration strategies use the same training recipe and curated data unless explicitly ablated. Training is run for

Table 2: **Performance comparison on audio and omni-modal collaborative reasoning.** Results are reported on the Qwen2.5-Omni-3B [11] and Qwen3-Omni-30B [12] backbones. **Bold**: best within each backbone group; green: gains over the corresponding baseline (gray row). Related methods: Omni-R1 [36], HumanOmniV2 [48]; benchmarks: MMAU-Pro [60], MMAU-test-mini [61], MMSU [62], MMAR [26], Daily-Omni [63], IntentBench [48], OmniVideoBench [27].

	Audio				Omni-Modal		
Method	**MMAU -Pro**	**MMAU -test-mini**	**MMSU**	**MMAR**	**Daily -Omni**	**Intent Bench**	**Omni Video Bench**
Omni-R1	52.36	77.70	61.87	59.70	54.14	64.18	32.40
HumanOmniV2	53.49	75.90	60.83	61.80	58.48	68.21	33.50
Qwen2.5-Omni-3B	46.20	73.40	55.00	53.80	55.05	59.30	31.20
+AudioJigsaw	50.05 ↑3.85	73.00 ↓0.40	57.95 ↑2.95	54.90 ↑1.10	–	–	–
+OmniJigsaw (JMI)	48.01 ↑1.81	72.80 ↓0.60	57.83 ↑2.83	53.40 ↓0.40	56.64 ↑1.59	60.19 ↑0.89	31.50 ↑0.30
+OmniJigsaw (SMS)	49.77 ↑3.57	73.20 ↓0.20	58.31 ↑3.31	56.90 ↑3.10	59.73 ↑4.68	61.38 ↑2.08	33.40 ↑2.20
+OmniJigsaw (CMM)	**51.06** ↑4.86	**73.50** ↑0.10	**59.39** ↑4.39	**57.60** ↑3.80	**61.90** ↑6.85	**62.82** ↑3.52	**34.50** ↑3.30
Qwen3-Omni-30B	56.61	74.40	70.16	68.50	69.92	67.40	38.80
+AudioJigsaw	57.67 ↑1.06	75.40 ↑1.00	70.30 ↑0.14	70.70 ↑2.20	–	–	–
+OmniJigsaw (JMI)	58.33 ↑1.72	74.50 ↑0.10	69.80 ↓0.36	69.10 ↑0.60	70.26 ↑0.34	67.95 ↑0.55	40.10 ↑1.30
+OmniJigsaw (SMS)	58.46 ↑1.85	75.80 ↑1.40	70.48 ↑0.32	69.50 ↑1.00	70.26 ↑0.34	68.21 ↑0.81	40.20 ↑1.40
+OmniJigsaw (CMM)	**58.59** ↑1.98	**76.30** ↑1.90	**70.70** ↑0.54	**71.00** ↑2.50	**71.09** ↑1.17	**68.89** ↑1.49	**40.50** ↑1.70

1,000 steps with 6 clips per sample, and the vision tower, audio tower, and internal feed-forward gating modules are frozen for efficiency. We also include uni-modal VideoJigsaw and AudioJigsaw baselines, which reorder shuffled clips using only visual or auditory cues. Evaluation uses greedy decoding. Uni-modal jigsaw formulations, detailed hyperparameters, data partitions, and prompts are provided in Appendix B, C, and F.

4.2 Main Results

Video Reasoning. We evaluate video reasoning on eight benchmarks spanning foundational visual perception to high-level downstream tasks under both w/ and w/o audio inference modes. As shown in Table 1, CMM consistently outperforms the corresponding baseline across both model scales (**+4.00** on MLVU [25] for Qwen2.5-Omni-3B [11]; **+4.38** on MLVU-Test [25] for Qwen3-Omni-30B [12]), with gains remaining visible even under the w/o audio mode. Notably, a clear ranking CMM $\succ$ VideoJigsaw $\succ$ JMI emerges across nearly all settings, indicating that simply exposing both modalities during post-training is less effective than controlling modality access.

Audio Reasoning. Table 2 shows that OmniJigsaw also improves audio reasoning (**+4.86** on MMAU-Pro [60] for Qwen2.5-Omni-3B [11]; **+2.50** on MMAR [26] for Qwen3-Omni-30B [12]). CMM achieves the strongest overall results, suggesting that clip-level audio-visual reordering does not merely improve visual reasoning but also provides a useful training signal for auditory tasks.

Omni-Modal Collaborative Reasoning. We further evaluate collaborative audio-visual reasoning on Daily-Omni [63], IntentBench [48], and OmniVideoBench [27]. As shown in Table 2, CMM again obtains the best results within both backbone groups (**+6.85** on Daily-Omni for Qwen2.5-Omni-3B [11]; **+1.70** on OmniVideoBench for Qwen3-Omni-30B [12]), while JMI yields smaller gains. These empirical findings reinforce that omni-modal collaborative reasoning benefits from structured modality access rather than naive joint exposure. Multi-seed results, evaluations on additional benchmarks, and qualitative examples are provided in Appendix D.3, D.4, and D.9, respectively.

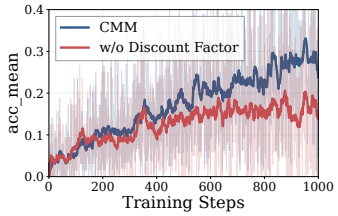


Figure 3: Training accuracy dynamics **w/ and w/o Discount Factor.**

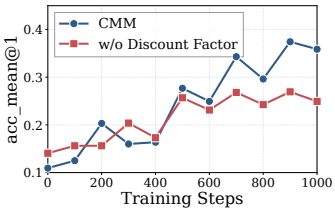


Figure 4: Validation accuracy dynamics **w/ and w/o Discount Factor.**

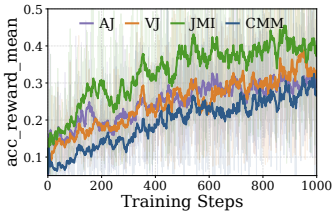


Figure 5: **Task reward (w/o R_{rep}, R_{fmt}) dynamics of JMI, CMM, VideoJigsaw and AudioJigsaw.**

4.3 Ablations and Analysis

Ablations of Puzzle Curation and Reward Design. Here, we primarily ablate the puzzle curation procedure and the accuracy-dependent discount factor $\lambda(\text{acc})$ introduced in Sec. 3.4, and analyze their individual contributions to CMM. Ablations of other reward components, semantic screening, clip count, and trimming are deferred to Appendix D.5. To assess the value of curation, we construct a control training set that matches OmniJigsaw-8K (Sec. 4.1) in source composition, mixing ratios, and total scale but uses uniform random sampling instead of the curation procedure, with the training recipe held identical. As reported in Table 3, removing curation consistently degrades CMM (**-3.99** on MLVU-Test [25], **-2.10** on MMAR [26], **-2.12** on IntentBench [48]): the reordering objective is well-posed only when adjacent clips exhibit identifiable state evolution, whereas videos with weak temporal dynamics or limited directionality produce mutually indistinguishable clips that yield no usable temporal supervision. We further ablate the discount factor by fixing $\lambda(\text{acc}) = 1$ while keeping all other reward components and the CMM recipe unchanged. Table 3 reports a uniform performance drop (**-2.09** on TUNA-Bench [50], **-1.70** on MMAR [26], **-1.33** on Daily-Omni [63]), and the optimization curves in Figs. 3 and 4 show that the variant without $\lambda(\text{acc})$ saturates early at a sub-optimal plateau, whereas the full design sustains a steady upward trajectory on both training and validation. Mechanistically, by suppressing the reward for imperfect rollouts, $\lambda(\text{acc})$ artificially amplifies the value disparity between sub-optimal and optimal solutions, driving the policy to explore the optimization landscape more aggressively in later training stages and effectively circumventing entrapment in non-optimal plateaus.

Table 3: **Ablations of puzzle curation and reward discount factor** across video, audio, and omni-modal benchmarks under CMM. Benchmarks: AoT [49], TUNA [50], TempC [51], V-TT [52], V-Holmes [53], MLVU/MLVU-T [25], V-MME [54], MMAU-P [60], MMAU-TM [61], MMSU [62], MMAR [26], D-Omni [63], Intent [48], OVB [27].

	Video								Audio				Omni-Modal		
Variant	*AoT*	*TUNA*	*TempC*	*V-TT*	*V-Holmes*	*MLVU-T*	*V-MME*	*MLVU*	*MMAU-P*	*MMAU-TM*	*MMSU*	*MMAR*	*D-Omni*	*Intent*	*OVB*
CMM	**66.83**	**66.20**	**72.34**	**46.50**	**48.29**	**62.75**	**69.30**	**73.46**	**58.59**	**76.30**	**70.70**	**71.00**	**71.09**	**68.89**	**40.50**
w/o Curation	64.94 ↓1.89	65.29 ↓0.91	71.14 ↓1.20	45.40 ↓1.10	47.41 ↓0.88	58.76 ↓3.99	68.60 ↓0.70	72.59 ↓0.87	57.67 ↓0.92	76.00 ↓0.30	70.40 ↓0.30	68.90 ↓2.10	70.01 ↓1.08	66.77 ↓2.12	40.10 ↓0.40
w/o DF	66.00 ↓0.83	64.11 ↓2.09	71.27 ↓1.07	44.70 ↓1.80	47.96 ↓0.33	61.95 ↓0.80	68.60 ↓0.70	72.36 ↓1.10	57.81 ↓0.78	75.90 ↓0.40	70.48 ↓0.22	69.30 ↓1.70	69.76 ↓1.33	67.66 ↓1.23	39.50 ↓1.00

The Bi-modal Shortcut Phenomenon. As indicated in Tables 1 and 2, integrating a second modality through naive joint exposure does not automatically help, and JMI in fact underperforms both uni-modal baselines. An additional visualization of the performance comparison among JMI, CMM, and uni-modal jigsaws is provided in Appendix D.6. We denote this as the "bi-modal shortcut phenomenon": when either modality alone suffices for reordering, the RL signal does not actually need the model to exploit the weaker modality. As shown in Figs. 5 and 6, although bi-modal cues

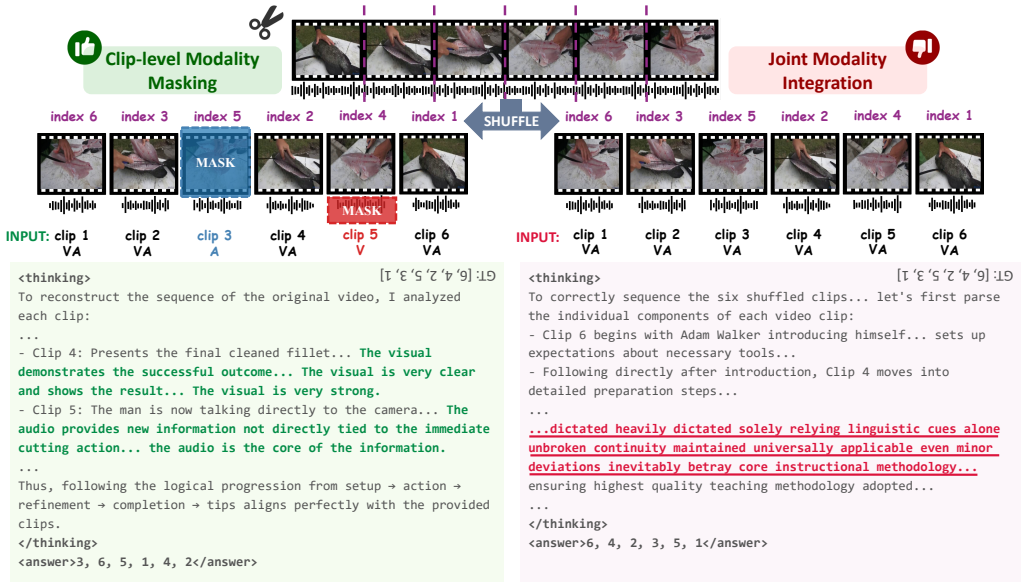


Figure 6: **Illustrative CoT comparison between CMM and JMI at training step 800.** CMM (left) guides the model to jointly analyze visual and auditory cues by masking less salient modalities (dashed boxes), while JMI (right) exhibits a bi-modal shortcut by **"solely relying [on] linguistic cues"** (underlined text) and bypassing the necessary visual analysis.

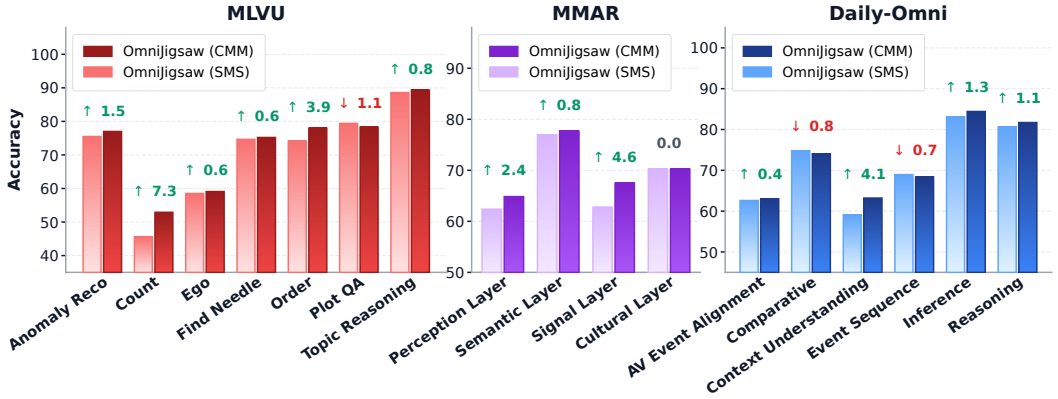


Figure 7: **Sub-capability performance comparison between CMM and SMS.** CMM's predominant superiority over SMS highlights the efficacy of clip-level orchestration in capturing temporally non-uniform audio-visual cues, whereas sample-level arbitration often misses local high-value modal information.

under JMI reduce task difficulty (evidenced by the markedly higher `acc_reward_mean` and the shortcut reasoning patterns within the CoT), this facile victory, which may attenuate the necessity to mine cues from the weaker modality, leads to weaker downstream transfer. CMM mitigates this by imposing clip-level modality access constraints, blocking the possibility of completing tasks via single-modality reliance. This strategy encourages the model to perform deep switching and integration between video and audio cues, transforming inter-modal "short-circuiting" into "mutual synergy", and thereby achieving performance surpassing uni-modal baselines. Further analyses are presented in Appendix D.1 and D.2.

Why Clip-level Granularity Matters. Figure 7 provides a detailed comparison of CMM and SMS across fine-grained sub-capabilities on representative video, audio, and omni-modal benchmarks (MLVU [25], MMAR [26], and Daily-Omni [63]). CMM outperforms SMS across most dimensions, and this performance disparity highlights the importance of granularity in data-adaptive modality

orchestration. In real-world videos, the dominant modality typically alternates non-uniformly along the temporal axis. SMS's sample-level arbitration prioritizes the modality with higher overall salience and may therefore overlook local high-value cues. CMM, in contrast, performs fine-grained local modality selection that conforms to the dynamic flow of audio-visual information, prompting the model to perform cross-modal semantic stitching across fragmented heterogeneous cues and deepening its analytical insight in complex multimodal scenarios.

5 Conclusion

We present OmniJigsaw, a self-supervised RL post-training framework that uses temporal reordering as a verifiable proxy task for omni-modal reasoning. Experiments across 15 benchmarks indicate that clip-level modality masking provides the strongest transfer among the studied modality orchestration strategies. Our empirical analysis suggests that naive joint access can induce a bi-modal shortcut, whereas controlled modality access makes the proxy task more useful for downstream reasoning. These results highlight modality orchestration as an important factor in designing self-supervised post-training objectives for omni-modal models.

Acknowledgments and Disclosure of Funding

This work was in part supported by the Hangzhou National AI Application Pilot Base (Embodied Intelligence).

References

[1] Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied AI. *arXiv preprint arXiv:2510.16732*, 2025.

[2] Jingtong Yue, Ziqi Huang, Zhaoxi Chen, Xintao Wang, Pengfei Wan, and Ziwei Liu. Simulating the visual world with artificial intelligence: A roadmap. *arXiv preprint arXiv:2511.08585*, 2025.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[5] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[7] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[9] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code Llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

[10] Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. DeepSeek-Coder-V2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*, 2024.

[11] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[12] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[13] Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, et al. OmniVideo-R1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*, 2026.

[14] Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced MLLM: A survey on RL-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025.

[15] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025.

[16] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *Advances in Neural Information Processing Systems*, 38:122721–122764, 2025.

[17] Kaibin Wang and Mingbao Lin. Test-time temporal sampling for efficient MLLM video understanding. *arXiv preprint arXiv:2511.17945*, 2025.

[18] Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. FAVOR-Bench: A comprehensive benchmark for fine-grained video motion understanding. *Advances in Neural Information Processing Systems*, 38, 2025.

[19] Enis Berk Çoban, Michael I Mandel, and Johanna Devaney. What do MLLMs hear? Examining reasoning with text and sound components in multimodal large language models. *arXiv preprint arXiv:2406.04615*, 2024.

[20] Yiwen Guan, Viet Anh Trinh, Vivek Voleti, and Jacob Whitehill. MLLM-based speech recognition: When and how is multimodality beneficial? *IEEE Transactions on Multimedia*, 2026.

[21] Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. Video-CoT: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12745–12752, 2025.

[22] Yanan Wang, Julio Vizcarra, Zhi Li, Hao Niu, and Mori Kurokawa. CoTasks: Chain-of-thought based video instruction tuning tasks. *arXiv preprint arXiv:2507.13609*, 2025.

[23] Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. VIDEOP2R: Video understanding from perception to reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pages 8303–8313, 2026.

[24] Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves MLLMs. In *International Conference on Learning Representations*, volume 2026, pages 93221–93245, 2026.

[25] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701, 2025.

[26] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38, 2025.

[27] Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. OmniVideoBench: Towards audio-visual understanding evaluation for omni MLLMs. In *International Conference on Learning Representations*, volume 2026, pages 138214–138236, 2026.

[28] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.

[29] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

[30] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

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

[32] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[33] Jasmine Yang, Poppy Zhang, and Shawndra Hill. MLLM-VADStory: Domain knowledge-driven multimodal LLMs for video ad storyline insights. *arXiv preprint arXiv:2601.07850*, 2026.

[34] Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. Omni-Emotion: Extending video MLLM with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502*, 2025.

[35] Romy Luo, Zihui Sherry Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. *Advances in Neural Information Processing Systems*, 38:83696–83727, 2025.

[36] Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-R1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *Advances in Neural Information Processing Systems*, 38:45266–45297, 2025.

[37] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[38] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[39] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32, 2019.

[41] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. In *International Conference on Information Processing in Medical Imaging*, pages 661–673. Springer, 2021.

[42] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12692–12702, 2020.

[43] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022.

[44] Ying Zhou, Xuefeng Liang, Shiquan Zheng, Huijun Xuan, and Takatsune Kumada. Adaptive mask co-optimization for modal dependence in multimodal learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[45] Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 469–485, 2025.

[46] Yake Wei, Yu Miao, Dongzhan Zhou, and Di Hu. MokA: Multimodal low-rank adaptation for MLLMs. *Advances in Neural Information Processing Systems*, 38:137470–137492, 2025.

[47] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2025.

[48] Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

[49] Zihui Sherry Xue, Romy Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *Advances in Neural Information Processing Systems*, 38:90925–90955, 2025.

[50] Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, Victoria W., and Fuzheng Zhang. TUNA: Comprehensive fine-grained temporal understanding evaluation on dense dynamic videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1810–1839, 2025.

[51] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, 2024.

[52] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards Video Thinking Test: A holistic benchmark for advanced video reasoning and understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20626–20636. IEEE, 2025.

[53] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like Holmes for complex video reasoning? In *European Conference on Computer Vision*, pages 539–556. Springer, 2026.

[54] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118, 2025.

[55] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[56] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[57] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. FineVideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

[58] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[59] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[60] Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22688–22697, 2026.

[61] Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pages 84929–84964, 2025.

[62] Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pages 31374–31410, 2026.

[63] 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.

[64] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[65] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *International Conference on Learning Representations*, volume 2026, pages 52423–52443, 2026.

[66] Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. AV-Odyssey Bench: Can your multimodal LLMs really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

[67] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10632–10643, 2025.

A Appendix Overview

The appendix is organized as follows:

- ► **Section B: Uni-Modal Jigsaw Formulation**
  Defines VideoJigsaw and AudioJigsaw as uni-modal references.
- ► **Section C: Additional Implementation Details**
  - **Section C.1: Details on CMM Modality Selection**
    Specifies the CMM modality selection procedure and parsing rules.
  - **Section C.2: Details on Puzzle Curation Pipeline**
    Describes both the heuristic signal-based filtering algorithm and the MLLM-based semantic screening configurations.
  - **Section C.3: Details on Training**
    Describes training data preparation, preprocessing, and hyperparameter settings for OmniJigsaw post-training.
  - **Section C.4: Details on Evaluation**
    Describes the benchmark suite, inference input preprocessing, and evaluation parameter settings.
- ► **Section D: More Results**
  - **Section D.1: Comparisons with VideoJigsaw and Random-CMM**
    Demonstrates CMM's value beyond generic temporal reordering RL and arbitrary modality masking.
  - **Section D.2: Theoretical Explanation and Cross-Modal Behavior**
    Explains reward-equivalent uni-modal paths and quantifies cross-modal behavior.
  - **Section D.3: Statistical Reliability**
    Reports independent training runs with three random seeds.
  - **Section D.4: Evaluation on More Benchmarks**
    Evaluates downstream transfer on WorldSense, AV-Odyssey, and VSI-Bench.
  - **Section D.5: Additional Ablation Studies**
    Ablates rewards, semantic screening, clip count, and trimming.
  - **Section D.6: Strategy Comparison Visualization**
    Visualizes the performance comparison among JMI, CMM, and uni-modal jigsaws.
  - **Section D.7: Fine-grained Evaluation across Sub-capabilities**
    Reports sub-capability evaluation results on representative video, audio, and omni-modal benchmarks.
  - **Section D.8: Cases of Semantic-based Data Filtering**
    Provides representative semantically rejected examples that illustrate why additional screening is necessary.
  - **Section D.9: Qualitative Examples**
    Presents qualitative comparisons between the baseline and its OmniJigsaw (CMM)-post-trained variant.
- ► **Section E: Limitations and Future Work**
  Discusses the current limitations of OmniJigsaw and possible future directions.
- ► **Section F: Prompts**
  - **Section F.1: Semantic Screening Prompt**
    Presents the prompt used for MLLM-based semantic screening.
  - **Section F.2: Training Prompts**
    Lists the prompts used for OmniJigsaw under different modality orchestration strategies.
  - **Section F.3: Evaluation Prompts**
    Provides the prompt templates used for benchmark evaluation.

B Uni-Modal Jigsaw Formulation

To comprehensively investigate the modality-specific enhancements yielded by OmniJigsaw, we establish two uni-modal references: VideoJigsaw and AudioJigsaw. Consistent with the notation defined in Sec. 3.1, we denote an omni-modal sample as $\mathcal{X} = (\mathcal{V}, \mathcal{A})$, which is segmented into N non-overlapping clips $\mathcal{S} = [s_1, s_2, \ldots, s_N]$, where each clip $s_i = (v_i, a_i)$ contains synchronized visual and acoustic streams. Let π be a random permutation sampled from the set of all possible bijective mappings of the index set $\{1, \ldots, N\}$ onto itself. The goal of these uni-modal tasks is to recover the ground truth chronological sequence $\mathbf{y} = [\pi(1), \pi(2), \ldots, \pi(N)]$ by observing only a single modality.

VideoJigsaw requires the model $\mathcal{M}_\theta$ to reconstruct the temporal order relying only on visual clips from the shuffled sequence, explicitly excluding any acoustic signals. This is formalized as:

$$\hat{\mathbf{y}}_v = \mathcal{M}_\theta(\tilde{\mathcal{V}}; \mathcal{I}_v), \quad \text{where } \tilde{\mathcal{V}} = [v_{\pi^{-1}(1)}, v_{\pi^{-1}(2)}, \ldots, v_{\pi^{-1}(N)}]. \tag{8}$$

The input $\tilde{\mathcal{V}}$ consists of N visual clips, where the j-th disordered clip corresponds to the $\pi^{-1}(j)$-th clip in the original chronological sequence.

Similarly, AudioJigsaw requires the model $\mathcal{M}_\theta$ to reorder shuffled audio clips without any visual assistance. The formulation is correspondingly defined as:

$$\hat{\mathbf{y}}_a = \mathcal{M}_\theta(\tilde{\mathcal{A}}; \mathcal{I}_a), \quad \text{where } \tilde{\mathcal{A}} = [a_{\pi^{-1}(1)}, a_{\pi^{-1}(2)}, \ldots, a_{\pi^{-1}(N)}]. \tag{9}$$

By decoupling the omni-modal inputs into these modality-specific variants, we can comprehensively assess and quantify the respective gains of each modality achieved through our proposed modality orchestration strategies.

C Additional Implementation Details

C.1 Details on CMM Modality Selection

CMM does not rely on an external MLLM, and both stages use the same model $\mathcal{M}_{\theta_t}$. Modality selection in CMM is a prompted pass of this model, while the prompt and execution rules remain fixed.

Given N ordered, synchronized clips, the model follows three prompt-defined steps: Content Evaluation, Modality Selection, and Sequence Verification. It selects V when visual action or state change is distinctive and audio is generic, A when speech or sound is temporally discriminative and vision is static or ambiguous, and VA when the two streams provide critical complementary cues. It then verifies that the retained evidence remains orderable and emits the modality vector used for subsequent reordering. Since information richness and temporal role are difficult to quantify objectively, we treat them as qualitative criteria specified in a fixed prompt, rather than as learned scalar quantities or outputs of a separate scorer.

Formally, the two stages are defined as follows:

$$\mathbf{m}_t = \mathrm{Parse}(\mathrm{GreedyDecode}(\mathcal{M}_{\theta_t}(\mathcal{S}; \mathcal{I}_{sel}))), \quad \hat{\mathbf{y}} = \mathcal{M}_{\theta_t}\left(\Phi_{cmm}(\tilde{\mathcal{S}} \mid \mathbf{m}_t); \mathcal{I}_{cmm}\right). \tag{10}$$

Here, $\mathcal{I}_{sel}$ and $\mathcal{I}_{cmm}$ are the selection and reordering prompts provided in Appendix F.2. Parse accepts exactly N labels in $\{V, A, VA\}$; otherwise, $\mathbf{m}_t = (VA, \ldots, VA)$.

C.2 Details on Puzzle Curation Pipeline

C.2.1 Algorithmic Implementation of Heuristic Filtering

To guarantee the fundamental solvability of the temporal puzzles, we initially develop a heuristic filtering procedure to prune ill-posed samples that lack modal integrity or irreversible transitions. The complete algorithmic workflow is systematically formalized in Algorithm 1.

First, we parse the metadata of each raw video. To ensure omni-modal integrity, samples lacking either a valid visual or audio stream are strictly discarded. Furthermore, taking computational tractability into account, samples exceeding a maximum duration of D_{max} can be further excluded. D_{max} is highly flexible and can be dynamically adjusted according to the available computational resources (we set $D_{max} = 200$ seconds).

Subsequently, we evaluate visual dynamism. To optimize processing efficiency, we uniformly sample frames at an interval of 1.0 second. Each sampled frame is spatially downsampled to 64×64 pixels and converted to grayscale. We then compute the Mean Absolute Difference (MAD) between adjacent sampled frames. A frame transition is classified as static if the MAD falls below a strict threshold of $\tau_v = 5.0$ (on a 0–255 intensity scale). Videos exhibiting a static frame ratio exceeding 70% are pruned, thereby eliminating frozen screens or visually monotonous content.

For the audio stream, we extract a mono-channel signal resampled to 16 kHz. The audio evaluation is tripartite: (1) Silence Removal: We compute the Root Mean Square (RMS) amplitude and convert it to a Decibel (dB) scale relative to the maximum amplitude. Audio segments below -40 dB are flagged as silence. Videos with a silence ratio greater than 70% are removed. (2) Dynamics Assessment: To filter out constant or monotone background noise, we calculate the onset strength envelope, representing the Spectral Flux (SF). A minimum SF variance of 0.5 is mandated, thereby excluding signals that lack sufficient acoustic dynamics. (3) Information Density Check: We employ a pretrained Silero Voice Activity Detection (VAD) model to locate human speech. We require the total

Algorithm 1 Signal-based Heuristic Filtering Pipeline

Require: Raw video dataset $\mathcal{V}_{raw}$.
Ensure: Filtered and standardized video dataset $\mathcal{V}_{filtered}$.

Hyperparameters: Duration limit $D_{max} = 200$ s (adjustable based on computational resources); Visual sample interval $\Delta t = 1.0$ s; MAD threshold $\tau_v = 5.0$; Max static ratio $R_{v_max} = 0.70$; Target sample rate $SR = 16$ kHz; RMS silence threshold $\tau_{rms} = -40$ dB; Max silence ratio $R_{s_max} = 0.70$; Min SF variance $\tau_{sf} = 0.5$; VAD ratio bounds $[R_{vad_min}, R_{vad_max}] = [0.30, 0.80]$.
$\mathcal{V}_{filtered} \leftarrow \emptyset$
for each video $V \in \mathcal{V}_{raw}$ **do**
  Extract duration D and stream metadata via `ffprobe`.
  if $D > D_{max}$ **or** V lacks audio/video streams **then**
    continue ▷ Discard invalid or overly long videos
  end if
  % Stage 1: Visual Dynamism Assessment
  Uniformly sample frame sequence $\{I_1, I_2, \ldots, I_N\}$ at interval Δt.
  $C_{static} \leftarrow 0$
  $\hat{I}_1 \leftarrow \text{Grayscale}(\text{Resize}(I_1, 64 \times 64))$
  for $i = 2$ **to** N **do**
    $\hat{I}_i \leftarrow \text{Grayscale}(\text{Resize}(I_i, 64 \times 64))$
    $\text{MAD}_i \leftarrow \frac{1}{H \times W} \sum |\hat{I}_i - \hat{I}_{i-1}|$
    if $\text{MAD}_i < \tau_v$ **then**
      $C_{static} \leftarrow C_{static} + 1$
    end if
  end for
  if $C_{static}/(N-1) > R_{v_max}$ **then**
    continue ▷ Discard visually static videos
  end if
  % Stage 2: Audio Quality and Information Density
  Extract audio $A_{16k} \leftarrow \text{Resample}(V_{audio}, SR)$
  $\text{RMS}_{dB} \leftarrow 20 \log_{10}(\text{RMS_Energy}(A_{16k}) / \max(\text{RMS_Energy}))$
  $\text{Ratio}_{silence} \leftarrow \text{Count}(\text{RMS}_{dB} < \tau_{rms}) / \text{Length}(\text{RMS}_{dB})$
  if $\text{Ratio}_{silence} > R_{s_max}$ **then continue**
  end if
  $\text{Var}_{SF} \leftarrow \text{Variance}(\text{Onset_Strength}(A_{16k}))$
  if $\text{Var}_{SF} < \tau_{sf}$ **then continue**
  end if ▷ Discard monotone noise
  $\mathcal{T}_{speech} \leftarrow \text{Silero_VAD}(A_{16k})$
  $\text{Ratio}_{speech} \leftarrow \sum(\text{durations in } \mathcal{T}_{speech})/D$
  if $\text{Ratio}_{speech} \notin [R_{vad_min}, R_{vad_max}]$ **then continue**
  end if
  % Stage 3: Standardization
  Standardize V to H.264/AAC in `.mp4` format.
  $\mathcal{V}_{filtered} \leftarrow \mathcal{V}_{filtered} \cup \{V\}$
end for
return $\mathcal{V}_{filtered}$

speech ratio to constitute between 30% and 80% of the video length. This bounded ratio deliberately ensures that the video contains rich instructional speech without overwhelmingly dominating the ambient acoustic events.

Finally, all retained candidates are standardized into the MP4 container with H.264 video and AAC audio encodings.

C.2.2 Inference Configurations for Semantic Screening

For MLLM-based semantic screening, we deploy Qwen2.5-VL-7B-Instruct [59] as the semantic assessor to guarantee temporal irreversibility and distinct state transitions within the candidate

videos. Unlike heuristic filters that operate on low-level signals, this stage focuses on high-level causal progression and logical determinism essential for OmniJigsaw. To achieve a comprehensive understanding of long-range temporal dynamics, we configure the model to process 200 uniformly downsampled frames per video, with a maximum spatial resolution of $100{,}352$ pixels per frame.

To ensure assessment reliability, the model's inference parameters are meticulously tuned, as summarized in Table 4. Specifically, we adopt a greedy decoding strategy (temperature $\tau = 0$) and a repetition penalty of 1.05 to suppress degenerative loops during long-form generation. We formally instruct the model to adhere to a Chain-of-Thought (CoT) protocol, wherein it must encapsulate its qualitative analysis, including evaluation criteria such as causal progression, visual evolution, and temporal markers, within `<think>...</think>` tags prior to emitting a definitive `YES/NO` decision within `<answer>...</answer>` tags. A maximum generation limit of $2{,}048$ tokens is allocated to accommodate exhaustive reasoning. This structured reasoning process ensures that only videos with unambiguous chronological flow and irreversible state changes are retained, effectively excluding repetitive actions or visually ambiguous content that would otherwise make the jigsaw puzzle ill-posed.

Table 4: **Inference configurations for semantic screening via Qwen2.5-VL-7B-Instruct [59].**

Category	Parameter	Value
Visual Input	Maximum Sampled Frames	200
	Maximum Resolution (pixels)	100,352
Sampling Params	Temperature (τ)	0 (Greedy Decoding)
	Top-p	1.0
	Top-k	-1 (Disabled)
	Repetition Penalty	1.05
Generation Specs	Maximum New Tokens	2,048
	Context Length Limit	32,768

C.3 Details on Training

C.3.1 Data Preparation

To construct the training dataset, denoted as OmniJigsaw-8K, we aggregate raw videos from multiple high-quality sources to ensure diversity in temporal logic: YouCook2 [56] is utilized for its procedural instructional causal chains, FineVideo [57] for coherent narrative flows, and the NextQA subset of LLaVA-Video-178K [58] for capturing dynamic physical actions. These raw data are subsequently curated through the two-stage puzzle-quality curation pipeline to eliminate ill-posed instances. The progressive filtering statistics throughout this curation process are detailed in Table 5. Ultimately, we obtain $8{,}220$ high-fidelity samples, which are randomly partitioned into a training set of $8{,}156$ instances and a validation set of 64 instances.

Table 5: **Data filtering statistics for OmniJigsaw-8K curation.** The pipeline progressively distills massive unannotated omni-modal data into high-fidelity samples suitable for the jigsaw proxy task.

Source Dataset	Raw Samples	After Stage 1	After Stage 2
YouCook2 [56]	2,000	327	327
FineVideo [57]	43,751	7,737	6,986
LLaVA-Video-178K [58] (NextQA)	3,868	982	907
Total	**49,619**	**9,046**	**8,220**

C.3.2 Data Preprocessing

Consistent with the OmniJigsaw formulation, the data preprocessing pipeline is meticulously designed to construct robust temporal reordering puzzles. Given an original video sample $\mathcal{X} = (\mathcal{V}, \mathcal{A})$, the sequence is uniformly partitioned into $N = 6$ non-overlapping clips. To prevent trivial solutions

derived from low-level boundary continuity (*e.g.*, matching identical consecutive frames across clip boundaries), the trimming operator $\mathcal{T}_{trim}$ explicitly discards 5% of the temporal duration from both the beginning and the end of each clip. Crucially, to guarantee exact multimodal synchronization and tensor uniformity, $\mathcal{T}_{trim}$ is applied to both visual and acoustic streams, forcing all clips to align with the duration of the shortest trimmed clip.

Regarding intra-clip preprocessing, the visual temporal downsampling operator $\mathcal{D}_T$ extracts frames uniformly via linearly spaced sampling at a default target rate of 2.0 FPS. To ensure computational tractability while preserving dynamic integrity, the sampled frame count per clip is strictly bounded between 2 and 12 frames. Subsequently, visual frames undergo an aspect-ratio-preserving spatial rescaling. Specifically, the target resolution is dynamically calculated to ensure that the total pixel count remains within $100{,}352$ while preserving the original aspect ratio. Concurrently, the synchronized audio waveform of each trimmed clip is resampled to 16 kHz without further truncation.

Based on these fundamental preprocessing protocols, the strategy-specific modality orchestrations, governed by the function $\Phi(\cdot)$, are implemented as follows:

- **OmniJigsaw (JMI):** As an identity mapping, Φ_{jmi} retains the complete synchronized visual and acoustic tensors $(\mathcal{D}_T(v_i), a_i)$ for each clip.
- **OmniJigsaw (SMS):** Φ_{sms} operates via a two-stage mechanism. In the first stage, a dominance analyzer (Qwen-Omni [11, 12]) evaluates the global modality dominance $d \in \{V, A\}$ by ingesting a temporally compressed global visual context (downsampled at 1.0 FPS and capped at 80 frames) alongside the full audio track. In the second stage, guided strictly by d, Φ_{sms} completely discards the unselected modality, ensuring that the jigsaw rollout is entirely uni-modal to effectively mitigate interference from the less informative stream.
- **OmniJigsaw (CMM):** Building upon JMI, $\mathcal{M}_{cmm}$ executes fine-grained clip-level modality masking according to the modality selection vector $\mathbf{m}$ (using Qwen-Omni [11, 12] as the modality selector). By replacing unselected modalities with $\mathbf{0}$, this strategy imposes a cross-modal information bottleneck that necessitates dynamic switching of attention between visual and acoustic cues.

C.3.3 Hyperparameter Settings

We instantiate OmniJigsaw on two representative open-source omni-modal backbones of different scales, Qwen2.5-Omni-3B [11] and Qwen3-Omni-30B-A3B-Instruct [12], and implement GRPO RL post-training under the proposed modality orchestration strategies. Our training pipeline is built upon the Volcano Engine Reinforcement Learning (VeRL) [64] framework, and all experiments are conducted on a cluster of $8\times$ NVIDIA H200 GPUs. Accounting for computational constraints and training efficiency, we freeze the vision tower, audio tower, and internal feed-forward gating modules across all strategies to prioritize reasoning alignment over perceptual feature learning. The training proceeds for $1{,}000$ steps with a global batch size of 8. We employ on-policy learning with a learning rate of 1×10^{-6} and a KL penalty coefficient of 1×10^{-2} estimated via the `low_var_kl` method. During the sampling process, we generate 8 responses per prompt with a decoding temperature of 0.9, a top-k of 50, and a top-p of 1.0. Furthermore, a repetition penalty of 1.05 is applied to mitigate degenerative looping and linguistic redundancy. Regarding the reward configuration, the composite function R_{tot} is parameterized with $w_{pos} = 0.5$ and $w_{cont} = 0.5$. The accuracy-dependent discount $\lambda(\text{acc})$ is set to 1.0 for perfect reordering and 0.2 otherwise. Additionally, a format reward $R_{fmt} = +0.2$ is awarded for structural adherence, while a repetition penalty $R_{rep} = -0.5$ is triggered by N-gram ($N = 20$) frequencies exceeding 3. Detailed hyperparameter configurations are summarized in Table 6. Under this setup, 1,000-step CMM training takes approximately 24 hours for Qwen2.5-Omni-3B [11] and approximately 60 hours for Qwen3-Omni-30B-A3B-Instruct [12].

C.4 Details on Evaluation

C.4.1 Notes on Evaluation Benchmarks

We evaluate OmniJigsaw across 15 diverse benchmarks covering video, audio, and omni-modal reasoning. We primarily focus on the multiple-choice QA within these benchmarks and report the top-1 accuracy as the primary performance metric. The evaluated benchmarks and their corresponding evaluation files are summarized in Table 7.

Table 6: **Hyperparameter settings for OmniJigsaw GRPO post-training implemented via VeRL [64].**

Hyperparameter	Value
Model & Infrastructure	
Base Policy Model	Qwen3-Omni-30B-A3B-Instruct [12] / Qwen2.5-Omni-3B [11]
Hardware Infrastructure	8× NVIDIA H200 GPUs
Frozen Components	Vision Tower, Audio Tower, Internal Feed-forward Gating
Training Dynamics (VeRL)	
Total Training Steps	1,000
Global Batch Size	8
PPO Mini Batch Size	8
PPO Micro Batch Size (per GPU)	4
Learning Rate	1×10^{-6}
KL Penalty Coefficient	1×10^{-2}
KL Estimator Type	`low_var_kl`
Prompt & Rollout Config	
Max Prompt Length	8,192
Max Response Length	2,048
Rollout Group Size	8
Decoding Temperature	0.9
Top-k	50
Top-p	1.0
Inference Repetition Penalty	1.05
Reward Configuration	
Positional Weight (w_{pos})	0.5
Continuity Weight (w_{cont})	0.5
Format Reward (R_{fmt})	+0.2
Discount Factor λ(acc)	1.0 (acc=1), 0.2 (acc=0)
Repetition Penalty (R_{rep})	-0.5

Table 7: **Detailed list of evaluation benchmarks and files.**

Modality	Benchmark	Evaluation File
Video	AoTBench [49]	`data_files/AoTBench_QA.json`
	TUNA-Bench [50]	`TUNA-MCQ.json`
	TempCompass [51]	`multi-choice/test-00000-of-00001.parquet`
	Video-TT [52]	`data/test-00000-of-00001.parquet`
	Video-Holmes [53]	`test_Video-Holmes.json`
	MLVU-Test [25]	`test_multi_choice_tasks.json`
	Video-MME [54]	`videomme/test-00000-of-00001.parquet`

	MLVU [25]	`MLVU/json/1_plotQA.json`
		`MLVU/json/2_needle.json`
		`MLVU/json/3_ego.json`
		`MLVU/json/4_count.json`
		`MLVU/json/5_order.json`
		`MLVU/json/6_anomaly_reco.json`
		`MLVU/json/7_topic_reasoning.json`

Audio	MMAU-Pro [60]	`test.parquet`
	MMAU-test-mini [61]	`mmau-test-mini.json`
	MMSU [62]	`question/mmsu.jsonl`
	MMAR [26]	`MMAR-meta.json`
Omni-modal	Daily-Omni [63]	`qa.json`
	IntentBench [48]	`qa.json`
	OmniVideoBench [27]	`data.parquet`

C.4.2 Inference Input Preprocessing

To ensure standardized evaluation, the inference input preprocessing protocols are dynamically instantiated based on the specific reasoning modes and modalities. The detailed video and audio processing logic for each evaluation mode is formalized as follows:

- **Video Reasoning:**
  - **Infer w/ Audio:**
    - *Video:* The visual stream is uniformly sampled to a maximum of 200 frames for Qwen3-Omni-30B-A3B-Instruct [12] and 100 frames for Qwen2.5-Omni-3B [11]. Each frame undergoes an aspect-ratio-preserving spatial rescaling to ensure that the total pixel count does not exceed 100,352 pixels.
    - *Audio:* The acoustic stream is resampled to 16 kHz with a maximum duration limit of 600 seconds for Qwen3-Omni-30B-A3B-Instruct [12] and 300 seconds for Qwen2.5-Omni-3B [11]. An adaptive audio extraction strategy is applied based on the temporal span of the sampled video frames:
      (a) Continuous Extraction: If the temporal span between the first and last sampled frames is within the maximum duration limit, a continuous audio segment centered on the midpoint of the frame sequence is extracted.
      (b) Frame-Anchored Extraction: Conversely, if the temporal span exceeds the maximum duration limit, the duration budget is equally divided by the total number of sampled frames. Short audio chunks centered at the exact timestamp of each visual frame are extracted, zero-padded if necessary, and sequentially concatenated.
  - **Infer w/o Audio:** The visual stream is processed identically to the Infer w/ Audio mode.
- **Audio Reasoning:** The raw audio waveform is resampled to 16 kHz with a maximum duration limit of 3,600 seconds for Qwen3-Omni-30B-A3B-Instruct [12] and 1,800 seconds for Qwen2.5-Omni-3B [11]. If the audio length exceeds this threshold, a central cropping operation is performed to strictly retain the middle segment within the corresponding duration budget. Missing audio streams are universally padded with zero arrays to maintain inference stability.
- **Omni-Modal Collaborative Reasoning:** Inputs are processed identically to the Infer w/ Audio mode in the Video Reasoning setting.

C.4.3 Inference Parameter Settings

To ensure the determinism and reproducibility of the model's reasoning processes, all benchmark evaluations are conducted using a greedy decoding strategy. We apply a repetition penalty of 1.05 to mitigate potential degenerative looping. The maximum number of newly generated tokens is set to 2,048 to accommodate the extensive reasoning often required for complex video, audio, and omni-modal QA tasks. The detailed configurations for the evaluation sampling parameters are summarized in Table 8.

Table 8: **Evaluation sampling parameter configurations.** These settings are uniformly applied across all benchmark evaluations to ensure deterministic and reproducible outputs.

Parameter	Value
Temperature (τ)	0 (Greedy Decoding)
Top-p	1.0
Top-k	-1 (Disabled)
Max New Tokens	2,048
Repetition Penalty	1.05

D More Results

D.1 Comparisons with VideoJigsaw and Random-CMM

D.1.1 CMM vs. VideoJigsaw

During proxy training, VideoJigsaw sees the visual stream for every clip, whereas CMM masks vision in audio-only clips and thus has no guaranteed uninterrupted visual path. Despite this constraint, CMM exceeds VideoJigsaw in 31/32 matched video settings (2 backbones $\times$ 2 inference modes $\times$ 8 benchmarks) in Table 1. Across these settings, CMM and VideoJigsaw average +2.72 and +1.62 points over the baseline, respectively.

To test whether video-only jigsaw learning accounts for audio transfer, we evaluate the Qwen2.5-Omni-3B [11] VideoJigsaw checkpoint on the four audio benchmarks in Table 2 under identical protocols.

Table 9: **Performance comparison of VideoJigsaw and CMM on audio reasoning.** Results are reported on Qwen2.5-Omni-3B. **Bold**: best result; green: gains over the baseline (gray row).

Method	MMAU-Pro	MMAU-test-mini	MMSU	MMAR
Qwen2.5-Omni-3B	46.20	73.40	55.00	53.80
+VideoJigsaw	47.22 ↑1.02	72.20 ↓1.20	56.27 ↑1.27	52.50 ↓1.30
+CMM	**51.06** ↑4.86	**73.50** ↑0.10	**59.39** ↑4.39	**57.60** ↑3.80

As shown in Table 9, VideoJigsaw yields inconsistent audio transfer, averaging -0.05 points relative to the baseline, whereas CMM improves all four benchmarks, averaging +3.29 points over the baseline and +3.34 points over VideoJigsaw. These results show that CMM is not merely a stronger video-only jigsaw, but an omni-modal proxy that jointly advances video, audio, and audio-visual reasoning.

D.1.2 CMM vs. Random-CMM

We additionally conduct a Random-CMM control experiment on Qwen2.5-Omni-3B [11], in which each example's CMM Stage-1 vector is randomly shuffled across clip positions once and then fixed throughout training. This preserves per-example $V/A/VA$ counts and the global distribution while breaking assignment–content alignment. We keep the data, Stage-2 construction, reward, GRPO [55] configuration, and training budget fixed; video evaluation uses *Infer w/o Audio*.

Table 10: **Performance comparison of Random-CMM and CMM on video reasoning.** Results are reported on Qwen2.5-Omni-3B under *Infer w/o Audio*. **Bold**: best result; green: gains over the baseline (gray row).

Method	AoT Bench	TUNA -Bench	Temp Compass	Video -IT	Video -Holmes	MLVU -Test	Video -MME	MLVU
Qwen2.5-Omni-3B	47.96	47.56	63.86	32.10	32.44	44.82	52.40	58.14
+VideoJigsaw	50.75 ↑2.79	51.12 ↑3.56	64.81 ↑0.95	32.70 ↑0.60	34.40 ↑1.96	46.02 ↑1.20	54.00 ↑1.60	59.98 ↑1.84
+JMI	48.97 ↑1.01	49.51 ↑1.95	64.68 ↑0.82	32.20 ↑0.10	33.10 ↑0.66	44.62 ↓0.20	53.80 ↑1.40	58.97 ↑0.83
+Random-CMM	49.58 ↑1.62	50.07 ↑2.51	64.94 ↑1.08	32.40 ↑0.30	33.81 ↑1.37	45.22 ↑0.40	54.20 ↑1.80	59.57 ↑1.43
+CMM	**50.92** ↑2.96	**51.54** ↑3.98	**65.44** ↑1.58	**33.50** ↑1.40	**36.80** ↑4.36	**46.41** ↑1.59	**55.30** ↑2.90	**61.91** ↑3.77

Tables 10 and 11 report the results across video, audio, and omni-modal benchmarks. Random-CMM consistently improves over JMI by 0.49/0.94/1.35 points on average over the Video/Audio/Omni groups, showing that a clip-level bottleneck itself is useful. However, on video benchmarks it remains

Table 11: **Performance comparison of Random-CMM and CMM on audio and omni-modal collaborative reasoning.** Results are reported on Qwen2.5-Omni-3B. **Bold**: best result; green: gains over the baseline (gray row).

	Audio				Omni-Modal		
Method	**MMAU-Pro**	**MMAU-test-mini**	**MMSU**	**MMAR**	**Daily-Omni**	**Intent Bench**	**Omni Video Bench**
Qwen2.5-Omni-3B	46.20	73.40	55.00	53.80	55.05	59.30	31.20
+JMI	48.01 ↑1.81	72.80 ↓0.60	57.83 ↑2.83	53.40 ↓0.40	56.64 ↑1.59	60.19 ↑0.89	31.50 ↑0.30
+Random-CMM	49.05 ↑2.85	73.20 ↓0.20	58.05 ↑3.05	55.50 ↑1.70	58.56 ↑3.51	60.82 ↑1.52	33.00 ↑1.80
+CMM	**51.06** ↑4.86	**73.50** ↑0.10	**59.39** ↑4.39	**57.60** ↑3.80	**61.90** ↑6.85	**62.82** ↑3.52	**34.50** ↑3.30

0.50 points below VideoJigsaw on average. CMM instead improves over VideoJigsaw by 1.01 points on average and over Random-CMM by 1.50/1.44/2.28 points across the three groups. Thus, Random-CMM preserves part of CMM's gain over JMI but not its advantage over VideoJigsaw; CMM's content-aware clip assignments yield the additional gains above both controls.

D.2 Theoretical Explanation and Cross-Modal Behavior

D.2.1 Reward-equivalent Uni-Modal Paths

Since the model's use of modality evidence is not directly verifiable by an objective rule-based metric, the reward in Sec. 3.4 verifies the response and predicted order without explicitly scoring modality use. When video or audio alone yields the same predicted order as the full audio-visual input, the two paths receive identical permutation rewards. JMI thus admits reward-equivalent uni-modal paths. Uni-modal jigsaws retain only one stream, requiring successful rollouts to consistently draw on that modality. Under JMI, the model can instead rely on whichever complete stream suffices for a given puzzle, which can weaken the incentive to learn from both modalities and integrate their complementary cues.

CMM changes the observation, not the reward: $\Phi_{cmm}(\tilde{\mathcal{S}} \mid \mathbf{m})$ varies modality availability across clips, reducing uninterrupted uni-modal paths. This explains why JMI's higher proxy reward (Fig. 5) need not yield stronger transfer.

D.2.2 Controlled Modality Intervention

To isolate evidence use, we apply a common input protocol to the step-1,000 JMI and CMM checkpoints of Qwen3-Omni-30B [12] on the fixed 64-instance held-out RL post-training validation set. We use strict exact-order accuracy: the complete six-clip permutation must match the ground-truth order. For each checkpoint, we first identify full-AV-correct predictions and then re-evaluate them under V-only and A-only inputs, with the puzzle, permutation, prompt, and greedy decoding fixed.

Table 12: **Controlled modality intervention on JMI and CMM.** Results are reported on Qwen3-Omni-30B using 64 held-out puzzles.

Method	**Full-AV correct**	**Either V-only or A-only**	**Neither**
JMI	31/64 (48.44%)	22/31 (70.97%)	9/31 (29.03%)
CMM	28/64 (43.75%)	12/28 (42.86%)	16/28 (57.14%)

As shown in Table 12, CMM solves nearly as many full-AV puzzles as JMI (28 vs. 31), yet substantially fewer of its successful predictions remain exact under either uni-modal condition: 42.86% vs. 70.97%. Conversely, 57.14% of CMM's full-AV successes fail under both uni-modal conditions, compared with 29.03% for JMI. This provides direct behavioral evidence that CMM induces greater

dependence on complementary audio-visual evidence, whereas JMI more often retains a successful uni-modal solution path.

D.3 Statistical Reliability

To assess the statistical reliability of the gains, we independently train Qwen2.5-Omni-3B [11] with CMM using three random seeds (1, 42, and 1234) and the same RL recipe. We evaluate Video-Holmes [53], MMAR [26], and Daily-Omni [63], representing video, audio, and omni-modal reasoning, respectively.

Table 13: **Statistical reliability across three training seeds.** Results are reported on Qwen2.5-Omni-3B; Video-Holmes uses *Infer w/o Audio*.

Run	Video-Holmes	MMAR	Daily-Omni
Qwen2.5-Omni-3B	32.44	53.80	55.05
+CMM (Seed 1)	36.80	57.60	61.90
+CMM (Seed 42)	36.31	58.20	61.49
+CMM (Seed 1234)	37.13	57.10	62.49
+CMM (Mean $\pm$ SD)	**36.75 $\pm$ 0.41**	**57.63 $\pm$ 0.55**	**61.96 $\pm$ 0.50**

As shown in Table 13, every run exceeds the baseline on every benchmark, with mean gains of +4.31/+3.83/+6.91 points and sample SDs of only 0.41/0.55/0.50, respectively. Thus, while RL stochasticity produces nonzero variation, the improvements remain reproducible across independent training runs and are not explained by single-run variance.

D.4 Evaluation on More Benchmarks

We further evaluate Qwen3-Omni-30B-A3B-Instruct [12] and its step-1,000 CMM checkpoint on WorldSense [65], AV-Odyssey [66], and VSI-Bench [67] under the inference protocol specified in Appendix C.4.

Table 14: **Evaluation in scenarios without clean temporal directionality.** Results are reported on Qwen3-Omni-30B. **Bold**: best result; green: gains over the baseline (gray row).

Method	WorldSense	AV-Odyssey	VSI-Bench
Qwen3-Omni-30B	53.00	30.49	53.74
+CMM	**53.63** ↑0.63	**32.29** ↑1.80	**55.01** ↑1.27

As shown in Table 14, CMM improves over the baseline by +0.63/+1.80/+1.27 points on WorldSense/AV-Odyssey/VSI-Bench, respectively. These consistent gains provide direct evidence that the proxy can transfer to scenarios without clean temporal directionality, including spatial reasoning.

D.5 Additional Ablation Studies

D.5.1 Reward Components

We further ablate the two verifiable task-reward terms introduced in Sec. 3.4: the continuity reward R_{cont} and the positional reward R_{pos}. Both variants use the curated OmniJigsaw-8K data, CMM modality orchestration, and the same training recipe, changing only the corresponding reward component. This comparison examines whether positional correctness, local temporal adjacency, or their combination provides more effective supervision.

Table 15: **Ablations of positional and continuity reward components** across video, audio, and omni-modal benchmarks under CMM. Benchmarks: AoT [49], TUNA [50], TempC [51], V-TT [52], V-Holmes [53], MLVU/MLVU-T [25], V-MME [54], MMAU-P [60], MMAU-TM [61], MMSU [62], MMAR [26], D-Omni [63], Intent [48], OVB [27].

	Video								Audio				Omni-Modal		
Variant	AoT	TUNA	TempC	V-TT	V-Holmes	MLVU-T	V-MME	MLVU	MMAU-P	MMAU-TM	MMSU	MMAR	D-Omni	Intent	OVB
CMM	**66.83**	**66.20**	**72.34**	**46.50**	**48.29**	**62.75**	**69.30**	**73.46**	**58.59**	**76.30**	**70.70**	**71.00**	**71.09**	**68.89**	**40.50**
w/o R_{cont}	65.88 ↓0.95	63.97 ↓2.23	71.20 ↓1.14	44.60 ↓1.90	47.80 ↓0.49	61.75 ↓1.00	68.50 ↓0.80	72.26 ↓1.20	57.74 ↓0.85	75.80 ↓0.50	70.42 ↓0.28	69.20 ↓1.80	69.67 ↓1.42	67.44 ↓1.45	39.40 ↓1.10
w/o R_{pos}	65.77 ↓1.06	64.25 ↓1.95	71.14 ↓1.20	44.50 ↓2.00	47.74 ↓0.55	61.55 ↓1.20	68.50 ↓0.80	72.22 ↓1.24	57.70 ↓0.89	75.70 ↓0.60	70.38 ↓0.32	69.30 ↓1.70	69.59 ↓1.50	67.28 ↓1.61	39.60 ↓0.90

As shown in Table 15, removing either R_{pos} or R_{cont} leads to consistent degradation across video, audio, and omni-modal benchmarks. This indicates that positional correctness and local temporal continuity provide complementary supervision for the jigsaw objective. While R_{pos} directly evaluates whether each clip is placed at the correct position, R_{cont} rewards recoverable adjacent transitions and supplies additional structural feedback for partial order recovery. The full reward performs best, suggesting that combining both terms yields a more reliable reward signal for temporal reordering.

D.5.2 Semantic Screening

To isolate the contribution of semantic screening, we conduct a Stage-1-only curation control on Qwen2.5-Omni-3B [11]. Keeping the training-set size unchanged, we uniformly sample 8,156 examples after signal checks but before semantic screening, with the source ratios and all CMM settings fixed.

Table 16: **Ablation of semantic screening.** Results are reported on Qwen2.5-Omni-3B under CMM; Video-Holmes uses *Infer w/o Audio*. Green: gains over Stage 1 only (gray row).

Curation Setting	**Video-Holmes**	**MMAR**	**Daily-Omni**
Stage 1 only	35.82	56.70	60.74
Stage 1 + Stage 2 (full)	36.80	57.60	61.90
Δ	**+0.98**	**+0.90**	**+1.16**

As shown in Table 16, Stage 2 yields consistent gains over Stage 1 alone, demonstrating the benefit of semantic screening beyond low-level signal filtering. Together with Table 3, these results support the effectiveness of the data curation pipeline.

D.5.3 Clip Count and Trimming

We also conduct ablations on the reordering proxy's clip count N and per-side trimming ratio using Qwen2.5-Omni-3B [11].

All CMM variants in Table 17 outperform the baseline. With 5% per-side trimming fixed, $N = 6$ achieves the highest average accuracy across the three representative benchmarks. $N = 4$ gives an easier, coarser proxy, whereas $N = 8$ yields shorter evidence fragments and harder optimization; 5% trimming balances boundary-shortcut suppression and evidence preservation.

D.6 Strategy Comparison Visualization

Fig. 8 provides a visualization of the pattern observed in Sec. 4.2: JMI remains consistently below the uni-modal references, whereas CMM delivers the strongest overall transfer. This reinforces that naive

Table 17: **Sensitivity to clip count and trimming.** Results are reported on Qwen2.5-Omni-3B under CMM; Video-Holmes uses *Infer w/o Audio*. **Bold**: default setting and its best scores; gray: baseline.

Method	Clip Count N	Trim per Side	Video-Holmes	MMAR	Daily-Omni
Qwen2.5-Omni-3B	–	–	32.44	53.80	55.05
+CMM	4	5%	36.15	57.00	62.24
+CMM	**6**	**5%**	**36.80**	**57.60**	61.90
+CMM	8	5%	35.87	56.40	61.15
+CMM	6	0%	35.98	56.70	60.90
+CMM	6	10%	36.25	56.30	61.40

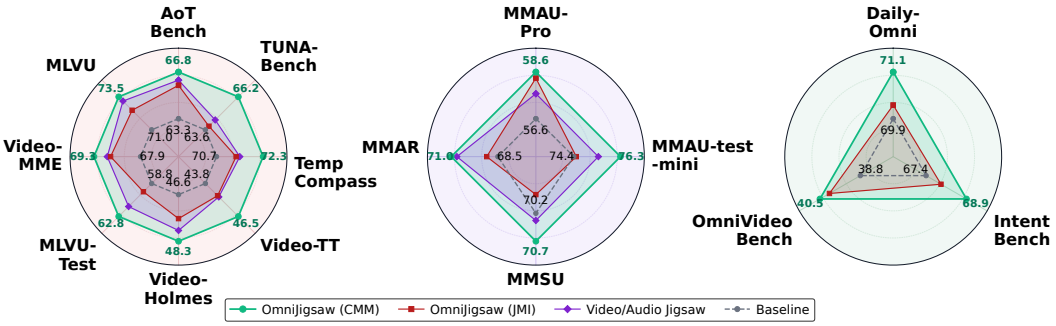


Figure 8: **Performance comparison of JMI, CMM, and uni-modal jigsaws across video, audio, and omni-modal benchmarks.** CMM's consistent superiority and JMI's performance degradation relative to uni-modal references empirically support the "bi-modal shortcut phenomenon".

joint exposure is insufficient, and that effective omni-modal jigsaw post-training requires controlled modality orchestration.

D.7 Fine-grained Evaluation across Sub-capabilities

The sub-capability evaluation results of OmniJigsaw on Video-MME [54] (video reasoning), MMAU-test-mini [61] (audio reasoning), and OmniVideoBench [27] (omni-modal collaborative reasoning) are systematically reported in Tables 18, 19, and 20, respectively. Overall, the three modality orchestration strategies exhibit distinct strengths across benchmarks and sub-capabilities. In particular, they show relatively consistent improvements on several sub-capabilities in Video-MME and MMAU-test-mini, while the gains on OmniVideoBench are more heterogeneous, reflecting the greater difficulty of omni-modal collaborative reasoning. CMM's regressions in Video-MME Temporal Perception and Spatial Reasoning appear to reflect the heterogeneous composition of QA samples within these sub-capabilities. Specifically, these drops correspond to only 2 fewer correct predictions out of 55 samples and 3 out of 56, respectively. Item-level analysis shows that Temporal Perception combines local content retrieval with temporal relations, while the Spatial Reasoning regressions concern venue identification, narrative organization, and performance themes rather than geometric relations. More direct evaluations show positive gains: Video-MME Temporal Reasoning rises from 57.1 to 62.1 (Table 18), and VSI-Bench from 53.74 to 55.01 (Table 14), suggesting localized regressions rather than broad temporal or spatial erosion.

D.8 Cases of Semantic-based Data Filtering

To further demonstrate the efficacy of the two-stage puzzle-quality curation pipeline, we provide representative qualitative examples of videos that pass initial signal-based heuristic filtering but are subsequently rejected during the semantic-based reasoning screening.

Table 18: **Sub-capability evaluation results on Video-MME [54].** **Green Bold** indicates performance surpassing the baseline.

Method	*Temporal Perception*	*Spatial Perception*	*Attribute Perception*	*Action Recognition*	*Object Recognition*	*OCR Problems*	*Counting Problem*	*Temporal Reasoning*	*Spatial Reasoning*	*Action Reasoning*	*Object Reasoning*	*Information Synopsis*
Qwen3-Omni-30B	80.0	70.4	77.0	69.0	76.6	74.1	46.3	57.1	78.6	57.2	65.9	80.5
+JMI	78.2	**74.1**	**78.8**	**71.2**	74.6	**75.5**	**47.4**	**62.7**	76.8	**58.6**	65.9	**80.8**
+SMS	80.0	**72.2**	**80.2**	**71.2**	74.6	74.1	**48.1**	**61.6**	73.2	**59.3**	**67.0**	80.5
+CMM	76.4	**75.9**	**78.8**	**73.8**	76.3	**77.0**	44.4	**62.1**	73.2	**61.8**	65.9	**80.8**

Table 19: **Sub-capability evaluation results on MMAU-test-mini [61].** **Green Bold** indicates performance surpassing the baseline.

	Task-wise			Difficulty-wise		
Method	**Sound**	**Music**	**Speech**	**Easy**	**Hard**	**Medium**
Qwen3-Omni-30B	76.88	72.75	73.57	70.09	70.76	77.78
+JMI	**77.18**	**73.65**	72.67	**70.54**	68.22	**78.89**
+SMS	**78.98**	**73.65**	**74.77**	**70.98**	68.64	**80.93**
+CMM	**79.28**	**75.15**	**74.47**	**73.66**	**72.03**	**79.26**

Table 20: **Sub-capability evaluation results on OmniVideoBench [27].** **Green Bold** indicates performance surpassing the baseline.

Method	*Attribute Comparison*	*Bkg. & Music Understanding*	*Causal Reasoning*	*Counting*	*Ego Reasoning*	*Fine-grained Perception*	*Hypothetical Reasoning*	*Reference Reasoning*	*Relationship Reasoning*	*Sentiment Analysis*	*Spatial Understanding*	*Summarization*	*Temporal Understanding*
Qwen3-Omni-30B	28.57	34.04	47.14	25.16	45.07	38.46	58.33	41.84	70.83	38.27	22.58	63.27	33.58
+JMI	**61.90**	31.91	**49.29**	21.29	**54.93**	**47.25**	45.83	**45.92**	66.67	38.27	22.58	55.10	32.85
+SMS	**38.10**	27.66	**50.00**	**26.45**	**47.89**	**41.76**	41.67	37.76	**79.17**	**44.44**	**30.65**	61.22	**34.31**
+CMM	**38.10**	29.79	46.43	24.52	43.66	**41.76**	58.33	**44.90**	70.83	**41.98**	**32.26**	55.10	**40.15**

D.8.1 Case 1: Indistinct State Changes

As illustrated in Figure 9, the video shows a person singing and gesturing in a temple-like indoor scene with consistent background elements (*e.g.*, microphone, statues, and decorations) throughout its duration. Although visually appealing and continuously dynamic, it is not suitable for the jigsaw proxy task. As indicated by the MLLM (Qwen2.5-VL-7B-Instruct [59]) output, this sample exhibits coherent performance dynamics but lacks sufficiently distinct visual states, clear environmental transitions, and reliable temporal markers for reordering. Consequently, while the video is not static in a low-level signal sense, it still lacks a directionally unambiguous chronological progression that would allow shuffled clips to be restored. This case illustrates the necessity of our semantic-based reasoning screening, which further removes videos with mild temporal variation but insufficient temporal-causal structure, ensuring that the retained samples exhibit strong and recoverable temporal reordering cues.

D.8.2 Case 2: Disjointed Narrative

Figure 10 shows another semantically rejected sample with a disjointed narrative. Although this video contains noticeable scene changes, it remains unsuitable for the jigsaw proxy task. The video jumps across visually distinct yet weakly related scenes (*e.g.*, sitting in a bathtub with books and appearing in an indoor room filled with books and artwork), rather than forming a single causal progression. The model explicitly identifies the lack of a coherent narrative flow, the absence of reliable temporal markers, and the fact that the clips could be rearranged without losing their meaning, and therefore returns a final `NO` decision. This example highlights that visual diversity alone does not imply suitability for the jigsaw proxy task. It further demonstrates the necessity of our semantic-based reasoning screening, which removes videos with apparent state changes but without directionally unambiguous chronological progression, thereby curating high-quality samples characterized by robust temporal-causal structures that are well suited for the jigsaw proxy task.

D.9 Qualitative Examples

Figures 11 and 12 show qualitative comparisons between the Qwen3-Omni-30B-A3B-Instruct [12] baseline and its OmniJigsaw (CMM)-post-trained variant on two representative downstream tasks: Sub-Scene Captioning and Video Summarization. OmniJigsaw (CMM) consistently generates responses that are more faithfully grounded in the observed video content, while reducing unsupported speculation and improving long-horizon semantic coherence. As shown in Fig. 11, the baseline drifts toward a post-incident description, focusing on actions such as "knelt down beside her" and "examined it closely", while failing to capture the decisive rescue chain required by the question. After OmniJigsaw (CMM) post-training, the model produces a more faithful interpretation of the event, explicitly identifying that the man "grabbed the woman's arm" and "prevented her from falling", which is substantially better aligned with the ground-truth explanation and the fine-grained scoring points. As illustrated in Fig. 12, the baseline produces a more speculative narrative with unsupported details such as "wealthy man", "descendant of a famous musician", and "humble background", revealing weaker grounding in the observed sequence. After OmniJigsaw (CMM) post-training, the summary becomes considerably more faithful to the video storyline, capturing a much more coherent event chain involving "a specific lineage", "a family of cotton farmers", "a contract", and "blood donation/transfusion". These qualitative examples indicate that our modality-orchestrated reordering post-training strategy enhances both fine-grained action grounding and long-horizon narrative composition, while reducing unsupported speculation and strengthening the model's ability to associate temporally distributed clues into coherent semantic abstractions.

E Limitations and Future Work

OmniJigsaw demonstrates consistent gains across video, audio, and omni-modal collaborative reasoning, but several aspects need to be further explored. First, due to resource constraints and efficiency considerations, our study is conducted under a relatively conservative training setup. Investigating its transferability across different data scales, training settings, and omni-modal architectures is a valuable next step. Second, the current puzzle curation pipeline is performed offline and therefore cannot explicitly adapt puzzle difficulty to the model's evolving capabilities during post-training. Capability-aware or curriculum-based filtering that better matches different training stages merits further exploration. Third, our proxy task relies on a relatively simple temporal reordering formulation with uniformly segmented clips. Richer puzzle designs such as variable clip lengths, overlapping clips, and hybrid spatio-temporal reordering deserve further study. Finally, beyond temporal reordering, exploring a broader family of self-supervised proxy tasks that naturally require joint video-audio reasoning may further advance omni-modal post-training for complex reasoning.

Input:



Output: Qwen2.5-VL

<think>
1. The video features a person singing into a microphone in what appears to be a temple-like setting. The individual moves around the space, gesturing expressively while singing. The background remains consistent throughout, showcasing golden statues and ornate decorations, which suggests a single location with no significant environmental changes.

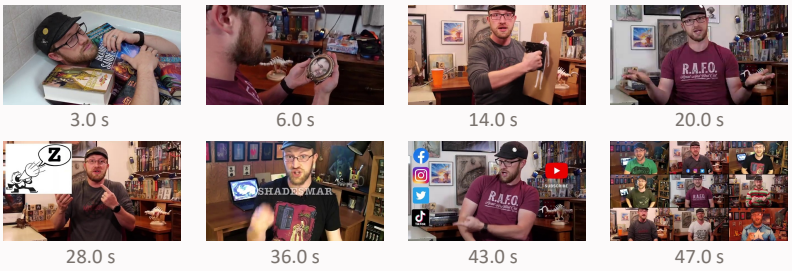
2. The Evaluation Criteria:
- **Causal Progression:** The person's actions (singing and gesturing) are continuous and directed towards the microphone, indicating a logical sequence.
- **Visual State Changes:** There are **no significant visual state changes**; the setting and **background remain constant**.
- **Logical Narrative:** The person's movements and gestures suggest a performance, maintaining a coherent narrative.
- **Temporal Markers:** The consistent setting and **lack of environmental changes** make it difficult to identify temporal markers.

3. The Exclusion Criteria:
- **Repetitive Loops:** The person does not perform repetitive actions; their movements are varied and expressive.
- **Static/Low-Dynamic Content:** While the background is static, the person's dynamic movements and expressions add dynamism to the scene.
- **Parallel/Unrelated Clips:** The video is a single, continuous shot with no unrelated highlights.
- **Visual Ambiguity:** The setting provides enough visual context to distinguish different parts of the video, but the lack of environmental changes makes it **challenging to identify specific visual cues for reordering**.
</think>

<answer>
NO
</answer>

Figure 9: **Case 1: Indistinct State Changes.**

Input:



Output: Qwen2.5-VL

<think>
1. **Causal Progression:** The video **does not follow a single, continuous causal progression** but rather jumps between different activities and settings. For example, the man is first seen reading outdoors, then in a bathtub surrounded by books, and later in a room filled with books and artwork. Each scene change is abrupt and **does not logically follow the previous one**.
2. **Visual State Changes:** There are clear visual state changes, such as ...
3. **Logical Narrative:** The video **lacks a coherent narrative flow**. The transitions between scenes do not provide a clear sequence of events that build upon each other. Instead, they seem to be standalone clips that **could be rearranged without losing their meaning**.
4. **Temporal Markers:** There are **no significant temporal markers** such as time-lapse photography, changes in lighting, or physical wear on objects that would indicate a passage of time or a sequence of events.

5. **Repetitive Loops:** The video does not contain repetitive loops ...
6. **Static/Low-Dynamic Content:** While the video includes some dynamic elements like the man's gestures and the changing settings ...
7. **Parallel/Unrelated Clips:** The video appears to be **a compilation of unrelated clips**, each capturing a different moment or activity. The lack of a clear narrative thread suggests that these clips could be rearranged without losing their individual meanings.
8. **Visual Ambiguity:** The visual content is somewhat ambiguous due to the abrupt transitions and lack of a clear sequence. The different settings and activities **do not provide enough context to infer a logical order**.
</think>

<answer>
NO
</answer>

Figure 10: **Case 2: Disjointed Narrative.**

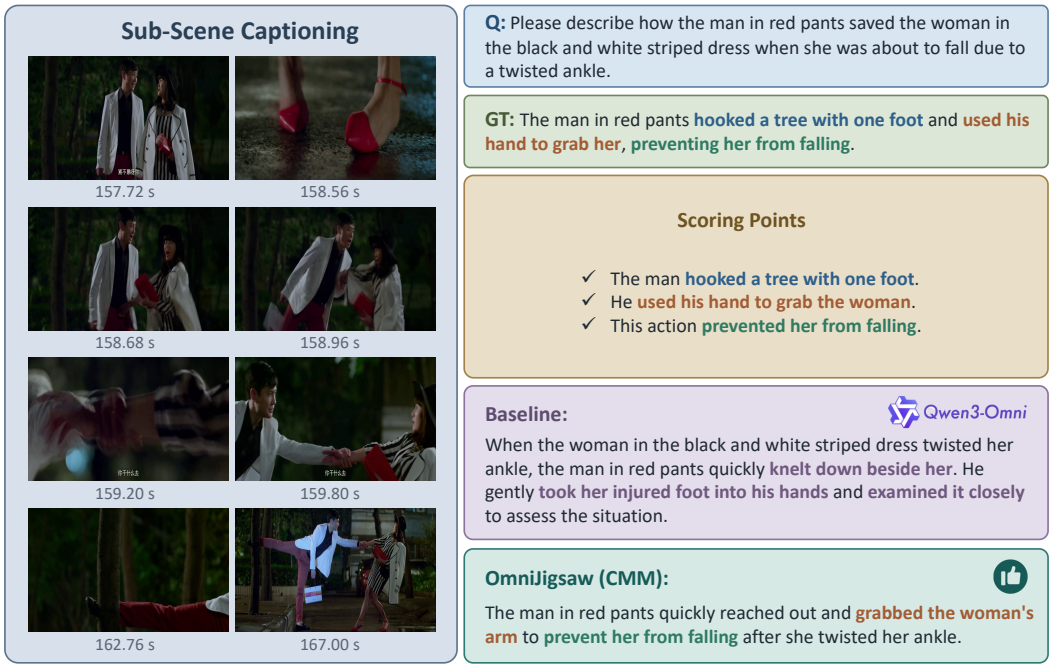


Figure 11: **Qualitative example of Sub-Scene Captioning.** Comparison between the Qwen3-Omni-30B-A3B-Instruct [12] baseline and its OmniJigsaw (CMM)-post-trained variant.

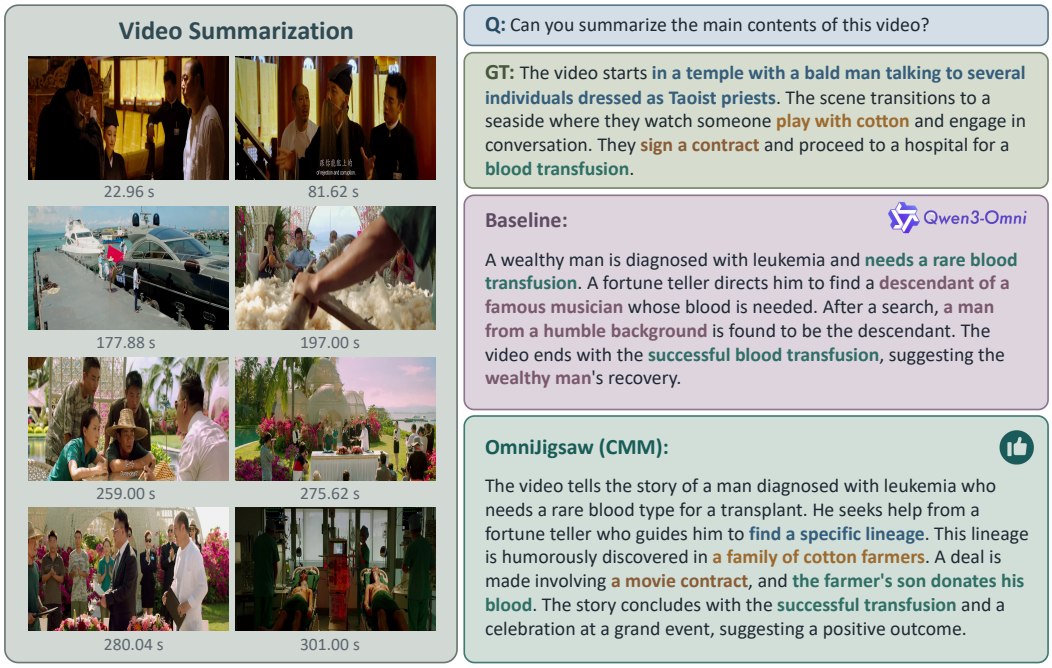


Figure 12: **Qualitative example of Video Summarization.** Comparison between the Qwen3-Omni-30B-A3B-Instruct [12] baseline and its OmniJigsaw (CMM)-post-trained variant.

F Prompts

F.1 Semantic Screening Prompt

Prompt for Semantic Screening

You are an expert video analyst tasked with determining if a video is suitable for a "Video Jigsaw" (temporal reordering) task.

Task Definition:
A video is suitable (YES) if it contains a clear, **directionally unambiguous** chronological progression with **distinct visual states**. If this video were cut into 3-6 segments and shuffled, a human should be able to restore the original order by identifying specific visual cues and logical flow.

Evaluation Criteria:
1. **Causal Progression:** Does action A lead to result B? (*e.g.*, ingredients being mixed then put into an oven).
2. **Visual State Changes:** Is there a clear evolution of an object or environment? (*e.g.*, a house being built, a drawing being completed, a plant growing).
3. **Logical Narrative:** Is there a sequential flow of events? (*e.g.*, a person walking into a building, then sitting at a desk).
4. **Temporal Markers:** Are there environmental changes like lighting (day to night) or physical wear/consumption?

Exclusion Criteria (Answer NO if):
1. **Repetitive Loops:** The person is doing the same action over and over (*e.g.*, a gym workout loop, a person just nodding/talking without changing position).
2. **Static/Low-Dynamic Content:** Talking heads with static backgrounds, or videos with minimal movement/change.
3. **Parallel/Unrelated Clips:** A compilation of unrelated highlights where order doesn't matter.
4. **Visual Ambiguity:** Different parts of the video look too similar (*e.g.*, a person walking in a desert with no landmarks), making reordering a guess.

Output Format:
You MUST follow this format:
<think>
1. Briefly describe the key actions and state changes observed.
2. Evaluate against the Evaluation and Exclusion criteria.
3. Conclude whether the video is suitable for the "Video Jigsaw" task.
</think>
<answer>YES or NO</answer>

F.2 Training Prompts

Prompt for OmniJigsaw (JMI)

You are given 6 **shuffled** video clips (with audio) that were created by slicing one original video into 6 equal-length temporal segments.

Here are the clips, each tagged with an index reflecting the current (shuffled) order in which they are shown. Please pay attention to both the visual actions and the specific content of the audio.

Clip 1: <video>
Clip 2: <video>
Clip 3: <video>
Clip 4: <video>
Clip 5: <video>
Clip 6: <video>

Task:
1. **For each clip**, describe its specific visual events and audio features (*e.g.*, distinct speech phrases or sound patterns).
2. Analyze the temporal logic by comparing these features to determine how the story or action progresses from one segment to another.
3. Based on this content-driven analysis, reassemble the original video and output the indices in correct order, separated by commas.

Answer format example:
2, 3, 1, 4, 6, 5

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer. **Ensure your reasoning is strictly based on audio-visual facts and follows a logical progression, avoiding unrelated digressions.** The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer MUST BE enclosed within <answer> </answer> tags.

Prompt for OmniJigsaw (SMS) (Dominance Analyzer)

Role
You are a Multimodal Content Analyst. Your goal is to evaluate the **visual stream** and **audio track** of the provided video to determine the optimal modality for a "Temporal Jigsaw Puzzle".

Task
You must decide whether the **Visual stream** or the **Audio stream** contains stronger, more deterministic clues for reconstructing the chronological order if this video were sliced into shuffled clips.

Analysis Guidelines
1. Evaluate Visual Suitability (For Video Jigsaw):
* **Look for:** Clear physical actions (*e.g.*, pouring water), scene transitions, camera movement, or object state changes (*e.g.*, assembling a puzzle).
* **Penalty:** Give low priority to "V" if the video is static (*e.g.*, a person sitting still with minimal movement), repetitive (loops), mostly black/blurry, or shows minimal change from beginning to end.
2. Evaluate Audio Suitability (For Audio Jigsaw):
* **Look for:** Continuous Speech (*e.g.*, narrative/dialogue with logical flow), Musical Progression (*e.g.*, verse -> chorus), or Sequential Sound Events (*e.g.*, footsteps -> door open -> slam).
* **Penalty:** Give low priority to "A" if the audio is constant background noise, a repetitive music loop without progression, or silence.
3. Comparison & Decision:
* **Select "V"** if the visual evolution provides a stricter, unambiguous timeline (*e.g.*, a silent movie with clear plot actions).
* **Select "A"** if the auditory narrative provides the primary logical thread (*e.g.*, a podcast, a speech, or a static shot of a narrator).
* *Tie-Breaker:* If both are good, choose the one that requires less ambiguous guessing for solving the "Temporal Jigsaw Puzzle".

Input Data
<video>

Output Format
Output **only** the single character decision inside valid tags.
<answer>V</answer>
or
<answer>A</answer>
You **FIRST** think about the comparison process as an internal monologue (**Analyze Visual Suitability -> Analyze Audio Suitability -> Final Decision**) and **THEN** provide the final answer. The thinking process MUST BE enclosed within <thinking> </thinking> tags. The final answer MUST BE enclosed within <answer> </answer> tags.

Prompt for VideoJigsaw & OmniJigsaw (SMS) (Video Jigsaw Rollout)

You are given 6 **shuffled** video clips that were created by slicing one original video into 6 equal-length temporal segments.

Here are the clips, each tagged with an index reflecting the current (shuffled) order in which they are shown. Please pay attention to both the visual actions and specific content.

Clip 1: <video>
Clip 2: <video>
Clip 3: <video>
Clip 4: <video>
Clip 5: <video>
Clip 6: <video>

Task:
1. **For each clip**, describe its specific visual events and features (*e.g.*, distinct event phases or evolving visual states).
2. Analyze the temporal logic by comparing these features to determine how the story or action progresses from one segment to another.
3. Based on this content-driven analysis, reassemble the original video and output the indices in correct order, separated by commas.

Answer format example:
2, 3, 1, 4, 6, 5

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer. **Ensure your reasoning is strictly based on visual facts and follows a logical progression, avoiding unrelated digressions.** The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer MUST BE enclosed within <answer> </answer> tags.

Prompt for AudioJigsaw & OmniJigsaw (SMS) (Audio Jigsaw Rollout)

You are given 6 **shuffled** audio clips that were created by slicing one original audio into 6 equal-length temporal segments.

Here are the clips, each tagged with an index reflecting the current (shuffled) order in which they are shown. Please pay attention to both the specific auditory events and content.

Clip 1: <audio>
Clip 2: <audio>
Clip 3: <audio>
Clip 4: <audio>
Clip 5: <audio>
Clip 6: <audio>

Task:
1. **For each clip**, **briefly** describe its specific auditory events and features (*e.g.*, distinct speech phrases or sound patterns).
2. Analyze the temporal logic by comparing these features to determine how the speech context, musical progression, or sound events evolve from one segment to another.
3. Based on this content-driven analysis, reassemble the original audio and output the indices in correct order, separated by commas.

Answer format example:
2, 3, 1, 4, 6, 5

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer.
Thinking Process Constraints:
1. **Be Concise:** Your entire thinking process MUST NOT exceed **800 words**.
2. **Strictly Auditory:** Ensure your reasoning is strictly based on auditory facts and follows a logical progression, avoiding unrelated digressions.
The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer MUST BE enclosed within <answer> </answer> tags.

Prompt for OmniJigsaw (CMM) (Modality Selector)

Role
You are a Multimodal Jigsaw Puzzle Expert. Your goal is to curate the modalities for a high-quality "OmniJigsaw" puzzle based on 6 sequential video clips.

What is OmniJigsaw?
OmniJigsaw challenges a solver to reconstruct the chronological timeline of shuffled clips using a curated mix of Visual (V), Audio (A), and Cross-modal (VA) clues.

Input Data
You are given 6 **ordered** video clips (each contains both video and audio). These clips were extracted from one original video in chronological order and have equal length.
Clip 1: <video> Clip 2: <video> Clip 3: <video> Clip 4: <video> Clip 5: <video> Clip 6: <video>

Task
Assign the most appropriate modality (**"V"**, **"A"**, or **"VA"**) to **each** clip based on its information richness and its role in the temporal sequence.
Analysis Steps:
1. **Content Evaluation:** For each clip, assess the richness and distinctiveness of its Visual stream vs. its Audio stream.
2. **Modality Decision Logic:**
- **Select "V":** If the **Visual** evidence is distinct (*e.g.*, scene change, physical motion) while the audio is generic, ambient, or redundant.
- **Select "A":** If the **Audio** evidence is distinct (*e.g.*, continuous sentence, distinct sound cue) while the visual is static, blurry, or less informative.
- **Select "VA":** If **BOTH** video and audio provide rich, critical clues for placing this clip in the timeline (*e.g.*, a specific character action synchronized with a plot-critical dialogue). Use VA when the clip serves as a strong "anchor" that requires full context.
3. **Sequence Verification:**
- Review your selections. Do they allow the solver to deduce the order through natural content progression? Ensure the sequence isn't rendered unsolvable by over-hiding critical cues.

Constraints
- **Avoid** outputting all "V", all "A", or all "VA".
- Strive for a natural mix. While VA is valuable, distinct V and A clips are also essential for testing specific modality understanding.

Output Format
You MUST provide the final result in valid JSON format inside <answer>...</answer> tags following this structure:
`{"modalities": ["V", "A", "VA", "V", "A", "V"]}`
You **FIRST** think about the **Content Richness and Modality Strategy** as an internal monologue (Content Evaluation -> Modality Selection -> Sequence Verification) and **THEN** provide the final design. The thinking process MUST BE enclosed within <thinking> </thinking> tags. The final design MUST BE enclosed within <answer> </answer> tags.

Prompt for OmniJigsaw (CMM) (Jigsaw Rollout)

You are given 6 **shuffled** clips that were created by slicing one original video into 6 equal-length temporal segments.

Important Note on Modalities: To test your cross-modal understanding, some clips may have missing audio (the audio is silent) or missing video (the video frames are entirely black). Use the non-missing and informative modalities in each clip to identify its content.

Here are the clips, each tagged with an index reflecting the current (shuffled) order in which they are shown. Please pay attention to both the visual facts and auditory clues.

Clip 1: <video>
Clip 2: <video>
Clip 3: <video>
Clip 4: <video>
Clip 5: <video>
Clip 6: <video>

Task:
1. **For each clip**, describe its specific visual events and audio features (*e.g.*, distinct speech phrases or sound patterns). If a modality is missing, focus on the available one. **For clip connections**, identify subtle transition clues (*e.g.*, action/state continuity, unfinished sentences, background sounds) and analyze the causal, logical, or emotional progression.
2. Analyze the temporal logic by integrating individual clip content with inter-clip relationships, explicitly chaining together both specific continuity details and thematic progression.
3. Based on this evidence-driven analysis, reassemble the original video and output the indices in correct order, separated by commas.

Answer format example:
2, 3, 1, 4, 6, 5

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer. **Ensure your reasoning is strictly based on clip content features, inter-clip relationships, and specific connection evidence, avoiding unrelated digressions.** The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer MUST BE enclosed within <answer> </answer> tags.

F.3 Evaluation Prompts

Prompt Template for Video Reasoning

Please answer the following question based on the complete video content.

<task>
Question: {question}
Candidates: {candidates}
</task>

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer (just the option letter) MUST BE enclosed within <answer> </answer> tags.

Prompt Template for Audio Reasoning

Please answer the following question based on the complete audio content.

<task>
Question: {question}
Candidates: {candidates}
</task>

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer (just the option letter) MUST BE enclosed within <answer> </answer> tags.

Prompt Template for Omni-Modal Collaborative Reasoning

Please answer the following question based on the complete video content (visual + audio).

<task>
Question: {question}
Candidates: {candidates}
</task>

You **FIRST** think about the reasoning process as an internal monologue and **THEN** provide the final answer. The reasoning process MUST BE enclosed within <thinking> </thinking> tags. The final answer (just the option letter) MUST BE enclosed within <answer> </answer> tags.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the Abstract and Sec. 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We don't present formal theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed descriptions of the training data sources, data curation pipeline, preprocessing procedures, training hyperparameters, evaluation files, inference settings, and prompts in Secs. 4.1, C.2, C.3, C.4, and F. We believe our work is fully reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and data are released at https://github.com/aim-uofa/OmniJigsaw and https://huggingface.co/datasets/zju/OmniJigsaw-8K, respectively.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: See Secs. 4.1, C.2, C.3.1, C.3.2, C.3.3, and C.4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: See Appendix D.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix C.3.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work performed does not raise relevant societal impact concerns.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We don't have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used in this work are properly credited and cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our new assets are released with accompanying documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We don't involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We don't involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The usage of LLMs is described in detail. See Sec. 4.1 and Appendix C.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.