# ICLR 2026 多模态智能与模型文明：深度研究报告与趋势分析

第十四届国际学习表示会议（ICLR 2026）被学术界视为机器学习研究从“模型构建时代”向“模型文明时代”转型的里程碑 1。在这一转型过程中，多模态智能（Multimodal Intelligence）不仅是技术演进的核心，更是模型应用与治理的先锋。本报告基于近两万篇投稿论文的统计数据、核心论文的标题与摘要整理、以及评审意见的深度分析，旨在为研究人员提供视觉多模态方向的权威参考与前瞻洞察。

## 一、 ICLR 2026 投稿数据与全球研究版图

ICLR 2026 的投稿规模呈现出爆发式增长，有效投稿总量达到 19,797 篇，较 2025 年的 11,672 篇大幅增长了约 70% 1。这一数字不仅确立了 ICLR 作为全球最大规模机器学习会议的地位，也反映了基础模型（Foundation Models）研究的全面渗透。然而，在数量激增的同时，论文的平均得分却出现了显著下滑，从 2025 年的 5.12 分降至 2026 年的 4.21 分 2。这种“分数跳水”现象引发了评审系统的巨大压力，评审周期被压缩至极限的两周，且评审专家普遍反映论文质量存在同质化严重的趋势 1。

根据论文标题、关键词及主要研究领域的聚类分析，多模态应用与基础模型占据了绝对主导地位。多模态视觉-语言-大语言模型（VLM-LLM）的研究占比已从 2023 年的 16% 攀升至 2026 年的约 40% 4。

### ICLR 2026 核心统计指标与历年对比

通过下表可以直观地观察到，ICLR 2026 在维持约 30% 接受率的前提下，对高质量研究的筛选变得更加严苛。

| **统计指标** | **ICLR 2024** | **ICLR 2025** | **ICLR 2026 (当前数据)** |
| --- | --- | --- | --- |
| 总投稿数 | 7,304 | 11,672 | 18,949 - 19,797 |
| 平均得分 | 5.11 | 5.12 | 4.21 |
| 最高得分 | 9.0 | 10.0 | 8.5 |
| 预计录取率 | 30.94% | 31.73% | ~28.18% (当前已录 5339 篇) |
| 拒绝/撤稿率 | 47.07% | 42.17% | 45.57% (拒绝) / 26.19% (撤稿) |

数据来源： 1

### 研究领域分布详情

“基础模型与前沿模型（含 LLM）”以及“计算机视觉/音频/多模态应用”是目前竞争最激烈的两个赛道，两者合计占比接近 38% 1。这表明学术界已形成共识：纯文本模型的研究已步入成熟期，未来的突破口在于跨模态的对齐、推理与物理世界交互 2。

| **排名** | **主要研究领域** | **论文数量** | **占比 (%)** |
| --- | --- | --- | --- |
| 1 | 基础模型/前沿模型 (Foundation Models) | 3,962 | 20.15% |
| 2 | 视觉/音频/多模态应用 (Multimodal Apps) | 3,458 | 17.59% |
| 3 | 生成模型 (Generative Models) | 1,841 | 9.36% |
| 4 | 对齐、公平性与安全性 (Safety & Alignment) | 1,512 | 7.69% |
| 5 | 数据集与基准测试 (Datasets & Benchmarks) | 1,496 | 7.61% |
| 6 | 强化学习 (Reinforcement Learning) | 1,291 | 6.57% |

数据来源： 1

## 二、 核心范式演进：从 Next-Token 到 Next-X Prediction

ICLR 2026 的多模态研究正在经历一场范式转移。传统的以语言为中心、将视觉视为辅助输入的模式，正逐渐被统一的“Next-X Prediction”范式取代 8。在这里，X 不仅代表文本 Token，还可以是视频帧（Frames）、连续的潜在空间表示（Continuous Latents）或多尺度的特征块 8。

### 统一多模态基础模型的架构竞争

目前，多模态领域呈现出三足鼎立的架构格局：自回归建模（Autoregressive Modeling）、预测编码器（Predictive Encoders）以及扩散生成器（Diffusion-based Generators） 8。

1. **离散自回归模型**：以 Chameleon 为代表的模型将文本和视觉信息全部离散化为 Token，利用 Transformer 的强扩展性进行联合预测 8。
2. **连续潜在预测模型**：如 VAR、MAR 和 TransFusion，这些模型直接在潜在空间中建模动力学。这种方法的优势在于能够更有效地捕捉视觉世界的高熵细节，而无需像离散 Tokenizer 那样进行由于量化带来的信息损耗 8。
3. **预测架构 (JEPA)**：联合嵌入预测架构（JEPA）代表了由 Yann LeCun 倡导的非生成路径。V-JEPA 2 等模型不通过生成像素来学习，而是通过在潜在空间中预测未来的表征来建模世界规律 9。这种架构的核心在于捕捉世界的可预测部分（如物体运动轨迹），而忽略不可预测的随机噪声 8。

## 三、 视觉多模态核心论文分析：标题与摘要整理

针对用户对视觉多模态方向论文整理的需求，本节筛选并深度分析了 ICLR 2026 投稿及相关工作中最具代表性的几篇论文。

### 1. SAM 3: Segment Anything with Concepts

**标题**：SAM 3: SEGMENT ANYTHING WITH CONCEPTS 12

**摘要概要**：SAM 3 提出了“可提示概念分割（Promptable Concept Segmentation, PCS）”任务，旨在根据文本描述或图像示例，识别并分割图像或视频中所有符合该概念的实例。SAM 3 克服了 SAM 1 和 SAM 2 仅支持点、框等视觉提示的局限。该模型由图像级检测器和基于记忆的视频追踪器组成，两者共享单一视觉主干。通过引入独特的“存在头（Presence Head）”设计，SAM 3 成功将“识别”与“定位”功能解耦，显著提升了在开放词汇任务下的检测准确度。在 SA-Co 基准测试中，SAM 3 的精度达到现有系统的两倍以上，并能通过与 MLLM 结合处理复杂的语义推理提示 12。

**深度分析与推荐理由**：SAM 3 是视觉多模态领域最具工程影响力的工作之一。它标志着分割任务从“手动指定物体”向“语义语义理解物体”的飞跃。对于从事自动化标注、医疗影像分析（如 Onco-Seg 应用 16）或工业检测的研究者，SAM 3 的解耦架构和大规模 SA-Co 数据引擎具有极高的参考价值。

### 2. Video-Thinker: Sparking “Thinking with Videos” via RL

**标题**：VIDEO-THINKER: SPARKING “THINKING WITH VIDEOS” VIA REINFORCEMENT LEARNING 17

**摘要概要**：尽管“带图思考（Thinking with Images）”在图像推理中取得了成功，但在视频领域仍具挑战。Video-Thinker 通过强化学习（特别是 GRPO 算法）激发多模态大模型的视频推理能力。模型在推理过程中不再是被动接受视频序列，而是能够主动调用内在的“定位”和“描述”能力生成中间推理线索。作者构建了包含 10,000 个推理序列的 Video-Thinker-10K 数据集。实验表明，Video-Thinker 在 Video-Holmes 等具有挑战性的视频推理基准测试中显著超越了 Video-R1 等竞争对手，并确立了 7B 规模模型的新 SOTA 17。

**深度分析与推荐理由**：该研究展示了强化学习如何将视频理解从单纯的特征聚合提升到逻辑严密的思维链推理。对于关注视频 QA、长视频理解和视觉思维链（Visual CoT）的研究人员，这篇论文提供了 SFT 与 RL 结合的实战范本。

### 3. OmniVerifier: Generative Universal Verifier

**标题**：Generative Universal Verifier as Multimodal Meta-Reasoner 18

**摘要概要**：针对多模态模型在推理过程中容易产生视觉幻觉的问题，本文提出了“全能验证器（OmniVerifier）”。这是一个用于视觉验证和精细化对齐的生成式插件。作者设计了自动化流水线构建大规模验证数据，并训练了 7B 规模的验证模型。在 ViVerBench 基准测试中，OmniVerifier 能有效检测生成的图像或推理过程中的逻辑错误，并给出修改建议。通过顺序测试时缩放（Sequential Test-Time Scaling），模型能够通过多轮验证迭代修正输出，显著提高了多模态推理的可靠性 18。

**深度分析与推荐理由**：在“模型文明”时代，如何确保生成内容的真实性是核心命题。这篇论文关于“验证器作为策略模型”的视角，为解决 VLM 幻觉和提升推理准确率提供了系统性的工具。

### 4. Mamba-3: Improved Sequence Modeling using State Space Principles

**标题**：MAMBA-3: IMPROVED SEQUENCE MODELING USING STATE SPACE PRINCIPLES 7

**摘要概要**：随着推理计算规模的扩大，Transformer 的二次计算复杂度成为瓶颈。Mamba-3 提出了三项核心方法论改进：1）更具表达力的梯形离散化（Trapezoidal Discretization）递归；2）复数值状态更新规则，增强了模型的状态追踪能力；3）多输入多输出（MIMO）架构，优化了硬件并行效率。Mamba-3 在语言建模、检索和状态追踪任务中取得了显著增益，并在固定推理预算下建立了新的帕累托前沿 7。

**深度分析与推荐理由**：虽然 Mamba 最初面向文本，但其在视频处理等长序列多模态任务中极具潜力。Mamba-3 对复杂动力学的建模能力（如数据依赖型 RoPE 的实现 7）使其成为构建下一代高效多模态架构的关键参考。

## 四、 视觉多模态方向可参考的重点论文推荐

除了上述详细整理的论文外，根据 ICLR 2026 的投稿动态和评审反馈，以下视觉多模态论文在特定子领域内具有极高的研究参考价值。

### 具身智能与视觉-语言-动作 (VLA) 模型

* **VLM4VLA: Revisiting Vision-Language-Models in VLA Models** 22：该论文深入探讨了预训练视觉编码器对机器人策略性能的影响。研究揭示了一个反直觉的结论：VLM 的通用 VQA 能力并不能直接预测其在下游具身控制任务中的表现。核心瓶颈在于视觉模块，通过注入控制相关的监督信号，即使冻结编码器也能获得增益。
* **BayesianVLA** 24：针对 VLA 模型中存在的“信息崩溃（Information Collapse）”问题（即模型过度依赖视觉观察而忽略语言指令），该研究提出了贝叶斯分解框架。通过 Latent Action Queries 区分视觉先验与语言条件后验，显著提升了模型在分布外（OOD）环境下的指令遵循能力。
* **DynamicVLA** 25：专注于动态物体操纵（如抛掷、捕捉），该工作提出了包含 20 万个合成片段的 DOM 基准，并利用紧凑的卷积视觉编码器实现了极速的多模态推理和闭环适配。

### 视觉推理与精细化感知

* **GePBench: Geometric Perception of MLLMs** 26：该研究揭示了现有模型在几何感知（如形状、相对位置、大小判断）方面的巨大短板。通过 28.5 万个合成样本的系统化评估，作者展示了训练合成几何任务如何能够迁移并提升模型在自然图像理解上的表现。
* **CURV: Enhancing Chart Understanding through Grounded Visual Reasoning** 27：该研究提出了课程学习框架，通过将图表问答（CQA）重新定义为多轮视觉推理，强制模型在逻辑推理的同时进行动态视觉定位。
* **MMRB2: Multimodal RewardBench 2** 28：建立了针对全能模型（Omni Models）的奖励建模基础，涵盖了文本到图像、图像编辑、交错生成以及“带图思考”推理任务，是进行强化学习对齐（RLHF）的重要参考基准。

### 3. 3D 视觉与生成一致性

* **Consistent3DGen** 29：针对 3D 扩散模型中存在的视角一致性问题（即 3D 物体与输入图像不完全匹配），该论文提出了一种训练无关的框架。通过将确定性重建（VGGT 算法）与生成式扩散相结合，实现了已知视角下的精确重建与未知视角下的创意补全。
* **Fracture-GS: Dynamic Fracture Simulation with Gaussian Splatting** 30：将物理引擎与高斯泼溅技术集成，用于模拟和渲染物体的动态破碎过程，展示了视觉表征与物理规律结合的新方向。

## 五、 深度洞察：视觉多模态研究的未来趋势与挑战

通过对 ICLR 2026 投稿池的全面审计，我们可以推断出未来两至三年内视觉多模态领域的三个关键趋势。

### 1. 从“缩放定律”到“对齐物理世界”

早期的多模态研究主要关注如何扩大模型参数或数据集规模。然而，ICLR 2026 的趋势表明，仅仅通过 Scaling 是无法获得真正的智能的 2。研究重点正在转向如何将模型智能“锚定”在物理世界中。这体现在对具身智能（VLA）、地理空间定位（GLEAM 6）以及时间尺度理解（Mamba-3 7）的重视上。未来的视觉多模态模型需要像人类一样感知时间和空间，而不仅仅是作为文本特征的聚合器。

### 2. 验证与推理：多模态思维链的崛起

视觉推理正在进入“思维链”时代。论文如 Video-Thinker 和 CURV 表明，让模型在给出答案前先进行“视觉思考”（定位关键区域、缩放细节、描述推理步骤）可以显著提升其在复杂任务中的准确率 17。同时，随着生成能力的增强，验证器（OmniVerifier 19）将成为模型治理的标准配置，通过闭环反馈减少模型的视觉幻觉。

### 3. 数据治理与合成数据的质量革命

随着模型训练对高质量数据的需求日益枯竭，ICLR 2026 见证了“数据工厂”模式的成熟。SAM 3 和 Gaia2 均采用了“人类在环（Human-in-the-Loop）”与“AI 验证（AI Verifier）”结合的策略来生产数以百万计的精细化标注 15。未来的竞争力将不再取决于数据的绝对数量，而取决于如何通过自动化流水线（如 GAIA 数据飞轮 32）高效筛选、修正并生成具有挑战性的负样本。

## 六、 结论与建议

总结 ICLR 2026 的榜单趋势，视觉多模态领域正处于从“单纯描述”向“复杂推理与主动交互”跨越的关键阶段。

1. **对于寻求高影响力课题的研究者**：建议关注“跨模态验证器”与“具身智能视觉编码器对齐”。这两个方向直接应对了现有大模型的短板。
2. **对于关注工程落地的研究者**：SAM 3 的解耦设计与 Mamba-3 的推理优化提供了极佳的范式，特别是在资源受限的环境下。
3. **对于基准测试研究者**：ICLR 2026 的趋势显示，静态、闭集的基准测试已逐渐边缘化，未来需要更多像 Gaia2 这样具有异步动态和实时验证能力的评估平台 31。

总体而言，ICLR 2026 不仅仅是一场学术会议，它通过 19,797 篇论文向我们展示了一个即将到来的、由全能模型（Omni Models）和自治代理（Agents）构建的文明蓝图 1。多模态智能作为这一文明的基础设施，其研究深度正随着“Next-X”范式的普及而不断拓宽。

#### 引用的著作

1. ICLR 2026 Submission Storm: 19,797 Papers, New Review Rules, and the Age of AI Convergence | CSPaper Forum, 访问时间为 二月 2, 2026， <https://forum.cspaper.org/topic/169/iclr-2026-submission-storm-19-797-papers-new-review-rules-and-the-age-of-ai-convergence>
2. ICLR 2026 Trends: Agentic AI, Multimodal Models & Data Governance - Encord, 访问时间为 二月 2, 2026， <https://encord.com/iclr-2026/>
3. ICLR 2026 Scores Released: Reviewers Call It "Crazy", DeepMind Researchers Show Comeback from Desperate Situation, 访问时间为 二月 2, 2026， <https://eu.36kr.com/en/p/3551362253731718>
4. Vision Language Models: A Survey of 26K Papers - ResearchGate, 访问时间为 二月 2, 2026， <https://www.researchgate.net/publication/396457863_Vision_Language_Models_A_Survey_of_26K_Papers>
5. ICLR 2026 Accepted Paper List, 访问时间为 二月 2, 2026， <https://papercopilot.com/paper-list/iclr-paper-list/iclr-2026-paper-list/>
6. ICLR 2026 Statistics - Paper Copilot, 访问时间为 二月 2, 2026， <https://papercopilot.com/statistics/iclr-statistics/iclr-2026-statistics/>
7. Mamba-3: Improved Sequence Modeling using State Space Principles - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/forum?id=HwCvaJOiCj>
8. ICLR 2026 Workshop on Multimodal Intelligence: Next Token Prediction and Beyond - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf/dbfd766aad3ada3875a0758013716727e704a7e3.pdf>
9. ICLR 2026 Workshop on Multimodal Intelligence: Next Token Prediction and Beyond, 访问时间为 二月 2, 2026， <https://iclr.cc/virtual/2026/workshop/10000795>
10. ICLR 2026 Workshop on Multimodal Intelligence: Next Token Prediction & Beyond, 访问时间为 二月 2, 2026， <https://mmintelligence.github.io/>
11. keon/awesome-physical-ai: A curated list of academic papers and resources on Physical AI — focusing on Vision-Language-Action (VLA) models, world models, embodied ai, and robotic foundation models. - GitHub, 访问时间为 二月 2, 2026， <https://github.com/keon/awesome-physical-ai>
12. SAM 3: Segment Anything with Concepts - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/forum?id=r35clVtGzw>
13. SAM 3 Unveiled at ICLR 2026: The Next Leap in Everything Segmentation - Empowering the Model to Grasp "Concepts" - 36氪, 访问时间为 二月 2, 2026， <https://eu.36kr.com/en/p/3507360247585920>
14. [2511.16719] SAM 3: Segment Anything with Concepts - arXiv, 访问时间为 二月 2, 2026， <https://arxiv.org/abs/2511.16719>
15. SAM 3: SEGMENT ANYTHING WITH CONCEPTS - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf/9cb68221311aa4d88167b66c0c84ef569e37122f.pdf?ref=blog.roboflow.com>
16. Onco-Seg: Adapting Promptable Concept Segmentation for Multi-Modal Medical Imaging, 访问时间为 二月 2, 2026， <https://www.medrxiv.org/content/10.64898/2026.01.11.26343874v2.full-text>
17. VIDEO-THINKER: SPARKING “THINKING WITH VIDEOS” VIA REINFORCEMENT LEARNING - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf/2434c327ac439ca9afeb57c45297d6f58cbecc77.pdf>
18. Generative Universal Verifier as Multimodal Meta-Reasoner - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/forum?id=DM0Y0oL33T>
19. GENERATIVE UNIVERSAL VERIFIER AS MULTIMODAL META-REASONER - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf/1c83185ff1b476816381b4646296a18a650e9b51.pdf>
20. MAMBA-3: IMPROVED SEQUENCE MODELING USING STATE SPACE PRINCIPLES - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf?id=HwCvaJOiCj>
21. Mamba 3 Improved Sequenc | PDF | Matrix (Mathematics) - Scribd, 访问时间为 二月 2, 2026， <https://www.scribd.com/document/937928450/13549-Mamba-3-Improved-Sequenc>
22. VLM4VLA: REVISITING VISION-LANGUAGE-MODELS IN VISION-LANGUAGE-ACTION MODELS 原创 - CSDN博客, 访问时间为 二月 2, 2026， <https://blog.csdn.net/u013250861/article/details/157097013>
23. [2601.03309] VLM4VLA: Revisiting Vision-Language-Models in Vision-Language-Action Models - arXiv, 访问时间为 二月 2, 2026， <https://arxiv.org/abs/2601.03309>
24. BayesianVLA: Bayesian Decomposition of Vision Language Action Models via Latent Action Queries - ResearchGate, 访问时间为 二月 2, 2026， <https://www.researchgate.net/publication/399985510_BayesianVLA_Bayesian_Decomposition_of_Vision_Language_Action_Models_via_Latent_Action_Queries>
25. DynamicVLA: A Vision-Language-Action Model for Dynamic Object Manipulation - ResearchGate, 访问时间为 二月 2, 2026， <https://www.researchgate.net/publication/400237222_DynamicVLA_A_Vision-Language-Action_Model_for_Dynamic_Object_Manipulation>
26. GePBench: Evaluating Fundamental Geometric Perception for Multimodal Large Language Models | OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/forum?id=UwtPOWp3NI>
27. CURV: ENHANCING CHART UNDERSTANDING THROUGH GROUNDED VISUAL REASONING - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf/0a1a20abd7cd6206131cdb87ff9961972b6e83ec.pdf>
28. Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image - arXiv, 访问时间为 二月 2, 2026， <https://arxiv.org/html/2512.16899v2>
29. CONSISTENT3DGEN: BRIDGING STOCHASTIC GEN- ERATION AND DETERMINISTIC RECONSTRUCTION FOR IMAGE-TO-3D DIFFUSION MODELS - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/pdf/1e12dbc997400bb63befd0fb2e377557c21a2df3.pdf>
30. ICLR 2026 Conference - OpenReview, 访问时间为 二月 2, 2026， <https://openreview.net/group?id=ICLR.cc/2026/Conference>
31. Gaia2: Benchmarking LLM Agents on Dynamic and Asynchronous Environments, 访问时间为 二月 2, 2026， <https://openreview.net/forum?id=9gw03JpKK4>
32. GAIA: A Data Flywheel System for Training GUI Test-Time Scaling Critic Models, 访问时间为 二月 2, 2026， <https://www.researchgate.net/publication/400083703_GAIA_A_Data_Flywheel_System_for_Training_GUI_Test-Time_Scaling_Critic_Models>
33. Snorkeling in RL environments, 访问时间为 二月 2, 2026， <https://snorkel.ai/blog/snorkeling-in-rl-environments/>