arXiv:2611.32172v2 [cs.CV] 25 Nov 2026

MediaStormBench-CameraLog：多模态模型能否识别相机对数曲线？

Ming Ma[1*†]、**Wenhao Hu**[6*]、**Qichao Wang**[2]、、**YuhaoWang**[3]、、**JunpengZhao**[4]、、**Qiyi Wang**[5]、、**Fangrui Zhang**[6]、、**Zhiqian Lai**[6]、、**Shanxin Guan**[6]、、**Haoyi Pei**[6]、、**Hongan Chen**[6]、和 **Rui Gong**[6]

[1]中国科学院大学、[2]南洋理工大学、[3]上海交通大学、[4]复旦大学、[5]南京大学、[6]MediaStorm

MediaStormBench 是一个面向真实影视制作工作流程的多模态基准测试，其各个子集均源自 MediaStorm 日常视频制作中的实际需求。本文发布其首个子集 MediaStormBench-CameraLog（CameraLogBench），聚焦于素材导入和初始调色中的一项常见基础任务：在没有元数据的情况下，确定拍摄某个片段时使用了哪条 Log 曲线。相机通过厂商定义的对数曲线，将宽范围的场景光压缩为像素代码值，从而录制 Log 素材；只有确定该曲线后，才能在后期制作中正确还原色彩。然而，该曲线仅记录在文件元数据中，在转码和团队交接过程中很容易丢失。此时，调色师只能在波形示波器上逐一尝试候选曲线，而错误的选择会改变整个片段的色调和色彩。仅从图像中还原编码颇具挑战，因为各厂商的曲线在代码值上十分接近，而且还受到未知曝光的影响。CameraLogBench 以专业拍摄且不含元数据的视频素材为基础，测试多模态模型能否仅凭图像识别相机的 Log 曲线；此外还设置了一项探测测试，通过移除正确选项来揭示模型是否凭先验猜测。由于每条候选曲线都由已发布的厂商公式严格定义，我们基于这些公式的单一实现构建了四种对照项：分级负样本、同一场景的合成帧、上下文规格摘要以及免训练公式基线。这些对照项共同将语义先验与像素的光度测量区分开来。在 11 个模型中，模型在真实视频素材上的准确率仅略高于始终猜测出现频率最高曲线的水平；移除正确选项后，模型几乎从不弃答，揭示出它们严重依赖场景与品牌先验。提供厂商规格能显著提高模型在经校准的图表上的识别能力，但对缺乏光度参考的自然图像几乎没有帮助，暴露出阅读规格与

以及测量像素。

项目页面：https://tech.ysjf.com/mediastormbench**代码**：
https://github.com/MediaStorm-Tech/MediaStormBench**数据集**：
https://huggingface.co/datasets/mediatech-tech/MediaStormBench

1. 引言

多模态模型正在进入视频制作工作流程。持续进行内容制作的团队需要模型在每个阶段承担具体任务，从前期制作规划、片场拍摄和素材导入，到剪辑、调色和交付；然而，模型能否真正集成到制作流程中，取决于其在该任务上的可靠性。现有大多数多模态基准测试围绕通用视觉理解和学术推理设计，很少借鉴真实制作工作，因此其得分无法直接回答这一问题。

因此，我们构建了 MediaStormBench。MediaStorm 是一家长期在视频平台发布原创内容的制作公司，我们将其日常制作工作流中的各个步骤转化为

*共同贡献 †项目负责人通讯作者：*Rui Gong*（*gary@ysjf.com*）。电子邮箱：*MingMa*（*mam22@ion.ac.cn*）、*Wenhao Hu*（*winstonhu@ysjf.com*）



图 1：一个真实的基准测试样本。**左侧：**一段没有元数据的 Log 片段中的三帧。**右上：**12条官方曲线；真实标签为 RED Log3G10。**右侧中部：**中间帧的亮度波形，虚线标出帧中位数 429，且 12 条曲线位于同一 10位编码值轴上。根据这些曲线，18% 灰卡对应的代码值范围为 285 到 500，而中位数 429 解码后的场景光最多相差 3.2 挡。**右下：**该帧分别按 Log3G10 正确解码和按 S-Log3 错误解码的结果。

标准化评估子集。本文发布了首个子集 CameraLogBench，它源于媒体管理和调色中一个反复出现且棘手的问题：在没有元数据的情况下，判断拍摄片段时使用了哪种 Log 曲线。

识别相机的 Log 编码，是专业后期制作中色彩复原和多机位匹配的先决条件。为了保留宽广的场景动态范围，电影摄影机、无人机、无反相机和手机会使用与宽色域配套的厂商发布的对数传递曲线压缩场景光，例如 Sony S-Log3（索尼公司，2014年）和 ARRI LogC3（ARRI，2017年）。未转换的Log视频素材看起来灰暗、褪色（图1，左侧），后期制作首先要对曲线应用逆变换。如果应用了错误的曲线，软件不会报错，但整个片段的色调和色彩都会系统性失真（图1，右下）。图1 中的真实帧，其10位代码值中位数为 429：解码为 Apple Log 时，该值比标准灰卡低 0.9 档；解码为 ARRI LogC4 时，则高出 2.3 档——场景光相差 3.2 档（约 9×）。真实制作项目也经常混用不同品牌：我们收集的 12 部作品每部都混用了 2 到 4 条曲线，而只要有一台相机被错误识别，其镜头就无法与其他镜头对齐。

然而，所使用的曲线仅记录在文件元数据中，而这些元数据在代理转码、重新封装以及团队之间交接的过程中很容易丢失。剪辑与调色软件依赖这些元数据来匹配色彩空间：Premiere Pro 文档指出，转码后的文件通常必须手动添加标签（Adobe Inc.，2026），并且无法读取元数据的素材会被视为 Rec.709（AdobeInc.，2025）。面对元数据已经丢失的灰片，调色师会观察波形示波器，寻找近中性区域，并逐一尝试候选曲线；这一过程耗时且依赖经验。要为长期积累的未标记素材库添加标签，需要一种仅凭图像就能识别 Log 曲线的方法。

本文探讨多模态模型能否在没有元数据的情况下，仅根据帧识别相机的 Log 曲线。这项任务面临两项困难。首先，候选曲线在代码值空间中高度聚集：在我们考察的 12条官方曲线中，有 9 条将黑位映射到相邻的代码值 94 和 95，有 7 条将 18% 灰卡映射到 392 至 420 的狭窄范围内。其次，镜头的曝光未知：仅凭全局亮度无法区分切换曲线与将曝光调整一挡。因此，答案取决于像素代码值的细微分布，而不是帧中显示了哪些物体。现有多模态基准主要关注高层次的语义理解和主体推理（Goyal 等人，2017年，Liu 等人，2024年，Yue 等人，2024年）、拍摄参数估计（Fang 和 Chen，2025年）或表观颜色感知（Liang 等人，2025年）；没有任何基准测试检验模型能否根据低级像素统计，区分相机的光电编码。

CameraLogBench 的核心数据集包含 330 段无元数据的专业拍摄片段，涵盖全部 12 条官方曲线，由 20 台相机机身拍摄，分布于 12 部作品和一次专门拍摄中。每段片段提供三帧给模型，并联合评估曲线识别、相机品牌识别，以及移除正确选项后的探测任务，以判断模型是在验证像素，还是根据先验猜测。由于每条候选曲线都由已发布的厂商公式严格定义，我们还基于这些公式的单一实现构建了四种对照项：“非 Log”负例（通过调色生成）、场景固定而仅改变曲线的合成帧、模型可在上下文中查阅的规格摘要，以及无需训练的解析公式基线。

对 11 个主流多模态模型的评测得出三项发现。第一，真实视频素材上的识别远未解决：没有参考材料时，最佳准确率仅为 0.20；而对每个项目都回答出现频率最高的曲线 D-Log，得分已达 0.16，11 个模型中有 7 个低于这一基线。第二，模型依赖场景和品牌捷径：移除正确选项后，它们几乎从不弃答，而是改选另一条曲线，且最常选同一品牌下编码相似的曲线。第三，提供厂商规格后，强模型在经校准的色卡上的准确率提升至 0.82，却几乎无法改善缺少光度参考的自然图像表现，这揭示了读取规格与测量像素之间的差距。

我们的贡献如下：

1. **MediaStormBench 的首个子集和一项新任务**：将后期制作中实际需要确定素材 Log 曲线的需求形式化为无元数据的相机 Log 曲线识别任务，涵盖 330 段专业拍摄片段、12 条官方曲线和 20 款相机机身。
2. **基于厂商公式构建的四种对照项**，以及两个基线，用于区分内容先验与光度测量。3. **对 11 种多模态模型的诊断**，揭示了它们在真实视频素材上的瓶颈、场景和品牌捷径，以及校准图表与自然图像之间的差距。

2. 相关工作

编码识别与摄像机来源。 剪辑与调色软件会根据文件元数据识别色彩空间（Adobe Inc.、2026、2025年）。助手工具 Eddie 会对像素进行采样，以判断未标记素材是否为 Log，但具体曲线仍取自残留元数据或摄像机品牌（EddieAI、2026）。Pearlstein 等人 （2019年）使用约 250 万张图像训练神经网络，以区分 BT.601 与 BT.709，这是一项二元分类任务。图像取证中的摄像机来源识别通过相机内痕迹（例如传感器噪声）区分设备型号（Gloe 和 Böhme，2010，Hosler 等人，2019年，Cozzolino 等人，2019年），而不是通过编码曲线：同一机身可以记录不同的曲线，不同品牌的机身也可以记录相同的曲线（第 3.2节）。LumiGrade（Guo 等人，2026）评估 Log 视频调色，但将输入

编码视为既定内容；我们考察的是它之前的步骤。

真实标签由程序或公式定义的基准测试。CLEVR（Johnson 等人，2017年）、DynaMath（Zou 等人，2025年）和 PRiSM（Imani 等人，2025年）使用程序生成图像和可验证的答案，但其测试图像是合成渲染图或通过程序绘制的色卡。DoRF（Grossberg 和 Nayar，2003）收集了 201 条胶片、传感器和数码相机的响应曲线，后续研究据此针对单张图像拟合连续响应函数（Li 和 Peers，2017年，Liu 等人，2020）。在我们的设定中，公式是相机实际执行的厂商标准，主要测试对象是真实视频素材，任务则是精确区分代码值非常接近的官方曲线；同一公式实现还可生成合成图像、免训练基线和规范参考。

多模态模型的定量视觉能力。通用多模态基准测试强调高层语义理解和主体推理（Goyal 等人，2017年，Liu 等人，2024年，Yue 等人，2024年），而 ChartQA（Masry 等人，2022年）则考察从统计图表中读取数值的能力。与我们更为接近的是 Camer- aBench（Fang 和 Chen，2025年），该基准估计照片的拍摄参数；SNAP（Kotseruba 和 Tsotsos，2025年）测量相机设置对识别的影响；PhotoBench（Qi 等人，2025年）测试摄影评析；ColorBench（Liang 等人，2025年）评估颜色感知与推理；QuantiPhy（Li 等人，2026）则表明，视频中对物理量的估计高度依赖先验。它们的证据仍然建立在可见的场景语义或几何内容之上，而在我们的任务中，答案取决于低层像素代码值的分布；探测和重新编码控制则将像素测量与基于先验的猜测区分开来（表 1）。

3. CameraLogBench

CameraLogBench以不含元数据的真实片段作为主要测试，并使用根据厂商公式推导出的四种对照项，细分模型在这些片段上的表现（图2 和表2）。

3.1. 任务定义

基准测试为每个片段设置三个问题。真实样本展示同一视频片段的三帧，不含元数据，分别取自其时长的 1/4 、 1/2 和 3/4 处，并按时间顺序作为独立图像一同发送；合成样本则是一张静态帧。

- **曲线识别，共 17 个选项**。选项包括 15 种“曲线／色域”编码，以及“非 Log 编码”和“无法确定”（附录 G）。其中 14 种编码基于 12 条具有已发布厂商公式的官方曲线（附录A）；剩下的一种 D-Log M 是 DJI 消费级机型常见的编码，没有已发布的公式，也绝不会作为真实标签。模型需要输出其观察到的证据、排名前两位的候选项以及最终选择。
- **品牌识别，11 个选项**。选项包括 10 个相机品牌以及“无法确定”。由于某一品牌的机身可以记录另一厂商的曲线（第 3.2 节），因此该问题将设备识别与曲线识别区分开来。
- **探测，16个选项**。帧和提示词保持不变，只从选项中移除了正确的编码，因此正确答案始终是“无法确定”；如果模型切换到另一条曲线，就说明它是在依赖先验进行猜测。

表1： 与相关视觉基准的比较。答案来源：如何获得真实标签。合成 / 真实：程序生成或真实的测试图像。抗猜测：用于测试依据先验进行猜测的成对扰动或探测测试。✓ 是，△ 部分， 否。

基准测试	Task	答案来源	合成	真实	抗猜测	规模
MMMU（Yue 等人，2024年）	大学水平 QA	人类	✗	✓	✗	1.15万道问题
ChartQA（Masry 等人，2022年）	图表阅读	人类 + 模型	✗	✓	✗	3.27万条问题
CLEVR（Johnson 等人，2017年）	合成场景推理	程序	✓	✗	△	10万张图像， ~1M QA
DynaMath（Zou 等人，2025年）	参数化视觉数学	程序	✓	✗	✓	501 × 10个问题
QuantiPhy（Li 等人，2026）	物理量，视频	程序，测量，人类	✓	✓	✓	3300 个问题
CameraBench（Fang andChen，2025年）	照片拍摄参数	相机元数据	✗	✓	✗	100 张照片
ColorBench（Liang 等人，2025年）	颜色感知，推理	人类 + 程序	△	✓	✓	1,448 个实例（5.8k+ QA）
LumiGrade (Guo 等人，2026)	Log 视频调色	指标 + 用户评分	✗	✓	✗	100+ 视频片段
德累斯顿（Gloe 和 Böhme，2010）	相机型号取证	采集设备	✗	✓	✗	>14k 张图像，73 台设备
CameraLogBench（我们提出的）	对数曲线识别	记录 + 厂商公式	✓	✓	✓	330 个真实片段 + 对照项

3.2. 真实视频素材

数据集与标签来源。真实集合包含330段未经调色的 Log 片段，由20种相机机身录制，涵盖全部12条官方曲线（表2，图 3a）。其中，170段来自12个真实制作项目，涵盖 D-Log（DJI，2022年）、Log3G10（RED Digital Cinema，2017年）、S-Log3（索尼公司，2014年）、N-Log（尼康公司，2018年）、Film Gen5（Blackmagic Design，2021年）和 KineLOG3（Kinefinity，2025年）。其余160段来自一次专门拍摄，涵盖 LogC3（ARRI，2017年）和 LogC4（ARRI，2025年）、C-Log2 和 C-Log3（佳能美国公司，2018年）、Apple Log（苹果公司，2023年）和 Apple Log 2（苹果公司，2025年），以及 F-Log2 C（富士胶片株式会社，2024年）。所有片段均为 3840×2160 10位 ProRes视频，时长在0.67至52.4秒之间（图 3b）；模型仅接收提取出的帧，这些帧不包含元数据。制作项目的标签取自项目记录，并通过原始录制文件的文件头或伴随文件进行交叉核验；专门拍摄的标签则读取自原始文件中保存的机内设置。

将品牌与曲线解耦的子集。在普通拍摄中，品牌与曲线紧密绑定，但这些制作项目自然包含一个解耦子集：Nikon 相机机身在内部以 R3D 格式录制了 35 个 RED Log3G10 视频片段，其中 17 个来自 ZR，18 个来自 Z8；而 Z8 和 Z6 III 还以 Nikon 自有的 N-Log 录制了另外 12 个视频片段。这使我们能够直接在真实视频素材上检验模型识别的是曲线还是品牌。



图 2： CameraLogBench 的评估流程和对照项。 **顶行：** 真实视频片段中的三帧（不含元数据）分别与曲线识别（17 个选项）、品牌识别（11 个选项）以及移除正确选项后的探测（16 个选项）配对，并在多个提示层级下提供给多模态模型；公式基线在相同项目上进行评分。 **底行：** 由已发布的曲线公式推导出的四个对照项，箭头标示了它们进入评估流程的位置：分级负样本、合成帧、规格摘要和公式基线（第 3.3节）。

代码值分布。在真实拍摄中，构图和曝光会自由变化，不能假定某一帧处于接近 18%灰的位置。各视频片段的帧中位代码值在其曲线对应的 18%灰度级附近广泛分散，并且不同曲线之间高度重叠（图 3c）；对于 D-Log，其 18%灰对应的代码值为 408，中位代码值范围为 186 至 665。每个视频片段中最暗区域的代码值都处于曲线黑位下限，或（在大多数片段中）远高于黑位下限（图 3d），因此自然拍摄的帧中没有纯黑参考。

真实色卡。 为测量画面包含已知光度参考时的识别效果，我们还使用 10 台相机机身拍摄了标准色卡的 30 帧，涵盖 12 种编码（表 2）。

3.3. 基于厂商公式推导的对照项和基线

仅凭真实视频素材上的准确率，无法判断模型究竟掌握了光度规律，还是依赖场景偏好。我们在单一公式代码库中实现了全部 12 项已发布规范（附录A），并据此推导出四个对照项和两个基线（图 2，底行）。

调色片段。 为测试模型能否将未转换的 Log 素材与还原后的素材区分开来，我们取用 KineLOG3 的全部 5 个视频片段，以及其他 12 种真实编码各自的 6 个视频片段，使用官方逆函数将其解码为 Rec.709，并进行调色：其中 38 个采用标准转换，26 个采用创意外观，13 个采用模拟 Log 素材的降低对比度处理。这 77 个负例的真实标签均为“非 Log”，其内容与原始视频片段相同，因此主题内容不会透露某个视频片段是否经过调色。

合成帧。 在真实集合中，不同曲线是在不同场景下由不同机身拍摄的。为消除这种偏差，我们使用官方公式对线性HDR图像进行正向编码。这些图像的像素值与场景光成正比：将平均亮度与18%灰对齐，把图像转换为与目标曲线配对的色域，再由公式将其映射为代码值。共有三个集合，示例见

表 2： 数据集构成及与各曲线配对的色域。真实集合包含 330 个 Log 片段和 77 个分级负样本。录制这些片段的相机机身见图 3a，30 个真实色卡帧详见附录 C （表 6）。 *Apple Log 2 和 Apple Log 采用相同的传递曲线，但配对的色域不同；Fujifilm F-Log2 C 保留 F-Log2 的传递曲线，并与更宽广的 F-Gamut C 配对，因此表中的 12 条官方曲线对应 13 种真实编码。

		Real		合成	
曲线	色域	片段	色卡	场景	色卡
Apple Log *	Rec.2020	11	1	50	3
Apple Log 2 *	Apple 广色域	30	1	–	3
ARRI LogC3	ARRI Wide Gamut 3	25	1	50	9
ARRI LogC4	ARRI Wide Gamut 4	22	15	50	3
Blackmagic Film Gen5	Blackmagic Wide Gamut	16	1	50	3
Canon C-Log2	Cinema Gamut	20	2	–	3
Canon C-Log3	Cinema Gamut	21	2	50	3
DJI D-Log	D-Gamut	54	1	50	3
Fujifilm F-Log2 C*	F-Gamut C	31	1	50	6
Kinefinity KineLOG3	Kinefinity 宽色域	5	–	50	3
Nikon N-Log	Rec.2020	12	2	50	3
RED Log3G10	REDWideGamutRGB	53	1	50	3
Sony S-Log3	S-Gamut3.Cine	30	2	50	3
非 Log	–	77	–	50	2
总计		**407**	**30**	**600**	**50**

附录 C （图 4）。(1) 合成自然场景集 选取100个公开HDR场景：来自 HDR 摄影调查（Fairchild，2007年）的76张照片，以及来自 Poly Haven（Poly Haven，2026）的24幅全景图；针对11条官方曲线中的每一条，我们从该场景池中编码50个不同场景，再加入50个非 Log 对照组，总计600帧。(2) 合成图表集 包含50帧通过程序绘制的图像，其中有灰卡、24 色块色卡、灰阶渐变，以及纯黑和剪裁色块，涵盖了所有14种具有公开公式的编码。(3) 重新编码集 对6帧真实图像进行反向变换，并分别使用11条曲线和2种非 Log 版本重新编码，总计78帧；该集合保留真实传感器噪声，用于测试预测结果是否遵循底层编码。

规格摘要。 为区分缺乏参数知识与缺乏测量能力，我们从同一实现中导出两级英文参考资料（附录 G）。规格 级别列出公式、锚定代码值和色域色度坐标；规格 +指南 级别增加了六步测量指南，涵盖阴影底限、中性参考、高光剪裁平台，以及对高饱和度颜色的色域检查。

不使用大型模型的两种基线。 两种参考基线用于量化像素本身所承载的信号；其实现方式见第 3.4 节。其公式基线 是一种不含学习参数的解析程序：它依次假设每条候选曲线，将帧解码回场景光和色度，剔除违反物理约束的曲线（例如，若阴影低于 154，就可排除 Apple Log，因为其黑位为 154），并从其余曲线中选出推断曝光最接近标准中灰的那条。小型分类器 是一种轻量级双分支卷积网络，仅使用公式合成图像进行训练，用于检验纯合成监督能否泛化到真实视频素材。

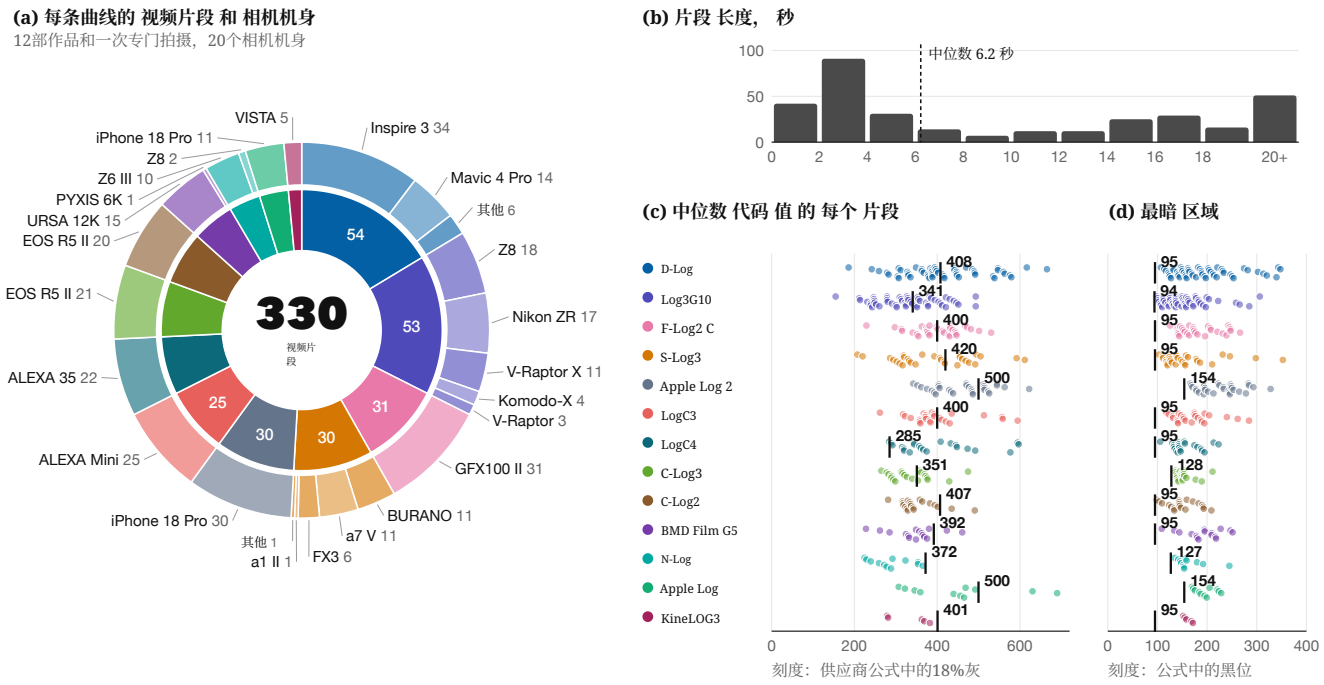


图3： 330个真实Log片段的统计数据。 **(a)** 按编码（内环）和拍摄机身（外环）划分的片段数量。**(b)** 片段时长（中位数为6.2秒）。 **(c)** 每个片段的帧中位代码值（圆点）及其曲线的18%灰度级（垂直线）。 **(d)** 每个片段的最暗区域（圆点）及其曲线的黑电平（垂直线）。代码值为10位。

3.4. 公式基线和小型分类器

公式基线。 公式基线是一种不含学习参数的解析程序，分三个步骤对每幅自然图像进行判定。第一步是统计测量：将帧划分为由 24×24 个像素组成的局部块，其均值可平滑单像素噪声；程序提取帧中最暗和最亮局部块的代码值，并检测高光中是否存在剪切平台，即大量像素堆积在同一个代码值上。第二步是依据物理约束进行筛选：程序依次假设每条候选曲线，利用其逆函数将各局部块解码回场景线性光，再通过配对色域矩阵将其转换为色度，并检查三项约束。帧中最暗部分不得明显低于该曲线的理论黑电平；解码后不得出现负光；转换得到的色度不得落在可见光谱轨迹之外。违反其中任何一项的候选曲线都会被剔除。第三步是根据曝光合理性为剩余候选曲线排序：解码帧的对数平均亮度越接近 18%灰，得分越高。此外，LogC3 和 D-Log 的白皮书给出了随感光度变化的高光剪切代码值；若观测到的高光平台与其中一个值吻合，则可获得加分。综合得分最高的候选曲线是第一预测，排名第二的则是第二候选曲线。

对于经校准的图表项目，公式基线使用已知色块位置的版本。程序直接读取灰卡、灰阶渐变的每一级以及 8 个标准色块的均值代码值。对于每条候选曲线，程序首先在灰阶序列上拟合一个全局曝光偏移，使曲线预测的代码值与测量值之间的均方根残差最小，并记录该残差；然后

使用该曲线及其配对色域将色块解码为色度，并计算它们与标准色度之间的偏差。这两种误差都与拍摄曝光无关。若某个候选项的任一误差超过其容差的约三倍，则将其舍弃；如果所有候选项都被舍弃，程序就输出“非 Log”。

对于真实色卡，色块位置由 ColorChecker 检测器自动定位，并逐帧进行检查。为反映真实相机的物理成像过程，真实拍摄测量模型设有三项设置：色块的参考值采用 X-Rite 发布的测量光谱数据；灰阶渐变拟合会联合估计镜头内部杂散光造成的阴影抬升；色度容差设为 0.01，即真实相机的色彩再现误差水平，而合成图表则设为 0.005。该基线正确回答了 30 个真实色卡帧中的 28 个，仅在 D-Log 和 Log3G10 上出错，这两种编码在数据集中各有一帧。

小型分类器。 该小型分类器用于测量完全由公式合成的监督信号如何泛化到真实视频素材。该模型包含两个分支。第一分支是从头训练的 ResNet-18，其输入为将 384×216 帧进行区域平均并缩小至原尺寸一半后得到的原始代码值张量；不进行 ImageNet 均值和方差归一化，以保留绝对代码值。第二分支计算四个通道（R、G、B 和亮度）中所有像素的 256 桶代码值直方图，并通过一个两层全连接网络提取 128 维特征。两个特征拼接后，输入线性分类头，输出 13 个类别：合成数据集涵盖的 12 种编码以及“非 Log”。

所有训练样本均在线合成。每个样本都是一帧具有普通视场的图像，由线性 HDR 全景图经透视投影获得，朝向、俯仰角和视场均随机选取；以 35% 的概率，将另一个场景的局部区域粘贴进来，以打破图像与场景语义之间的关联。随后对对比度、饱和度、曝光和白平衡施加随机扰动，模拟高光裁切，并加入拍摄噪声、读取噪声和镜头眩光；最后，随机抽取一条官方曲线及其公式对结果进行编码，再量化为 8 位。“非 Log”类别则会进行随机色彩复原和创意调色。训练集由 Poly Haven 的 729 张全景图（Poly Haven, 2026）组成；验证集包含按拍摄地点留出的 80 张全景图，且训练时不会使用与合成测试集来自同一地点的全景图。神经网络使用 AdamW 优化器训练，峰值学习率为 10^{-3}，采用单周期调度，批量大小为 64，共训练 16,000 步；对于真实视频片段，则对三帧的输出概率取平均值。随着测试分布逐步偏离训练域，分类器的曲线识别准确率逐步下降：在留出的验证全景图上为 0.46，在合成自然场景集上为 0.26（表5），在 330 个真实视频片段上为 0.14（表 3）。增加训练全景图数量或将训练步数增加至四倍后，真实视频素材上的准确率仍为 0.14–0.15。

3.5. 评估协议

每帧采用面积平均法缩小至 1920×1080，并以 8 位无损图像格式存储，不进行色调或色彩空间转换；附录 B 将其与 10 位原图逐像素进行比较。提示词分为四个层级，各层级共用同一问题和选项（附录 G，表 11）。仅文本 会隐藏图像，以检查措辞是否会引入偏差；仅帧 用于衡量模型凭借其内部知识能够得出什么；帧 + 规格和 帧 + 规格 + 指南 则会加入第 3.3 节中的两份规格摘要。我们在除 帧 + 规格以外的所有层级上评估真实项目，并在三个视觉层级上评估合成项目；每个项目均按两种选项顺序各提问一次，没有给出有效选项的输出计为错误。

我们报告四项指标：准确率 准确率 为所有调用的总体结果；前二命中率是指前两个候选项中包含正确答案的调用所占比例 es

表 3： 真实视频素材上的主要结果。前六列报告了在两个提示层级下，对 330 个 Log 视频片段进行的曲线识别结果：准确率（Acc.）、前二命中率（Top-2）以及 13 种真实编码的按曲线平均值（Avg.）。品牌和非 Log 仅使用帧：品牌在 330 个 Log 视频片段上的准确率，以及在 77 个分级负样本上的检出率。粗体标出每列中表现最佳的多模态模型。

	仅帧			帧 + 规格 + 指南				
方法／模型	准确率 ↑	Top-2 ↑	平均值 ↑	准确率 ↑	前-2 ↑	平均值 ↑	品牌 ↑	非 Log ↑
随机猜测	0.06	0.12	0.06	0.06	0.12	0.06	0.09	0.06
始终选择 D-Log	0.16	–	0.08	0.16	–	0.08	0.16	–
公式基线	0.09	0.25	0.09	0.09	0.25	0.09	–	–
小型分类器	0.14	0.30	0.15	–	–	–	–	0.96
Claude Opus 5.5	0.13	**0.30**	0.09	0.17	0.30	0.12	**0.22**	0.53
Claude Sonnet 5.5	0.15	0.23	0.12	0.16	0.25	0.11	0.21	0.66
GPT-6.1 Sol	0.14	0.28	0.09	0.25	**0.41**	**0.18**	0.15	0.94
GPT-6 Astra	0.14	0.27	0.08	**0.27**	**0.41**	0.17	0.14	**0.97**
Gemini 3.1 Pro	0.19	0.27	**0.14**	0.14	0.24	0.12	**0.22**	0.22
Gemini 3.8 Flash	**0.20**	0.29	**0.14**	0.20	0.31	0.13	0.18	0.49
Qwen3.8-Max	0.06	0.19	0.05	0.05	0.19	0.07	0.20	0.87
Qwen3.8-Flash	0.14	0.24	0.10	0.16	0.28	0.12	0.19	0.38
Kimi K3.1 Pro	**0.20**	0.28	0.13	0.20	0.26	0.14	0.19	0.61
Kimi K3	0.18	0.27	0.12	0.16	0.25	0.12	0.21	0.42
DeepSeek-V4.1-Flash	0.15	0.25	0.10	0.18	0.31	0.14	0.20	0.60

包含真实标签；按曲线平均值，即每种编码准确率的未加权均值，可消除类别大小的影响；以及弃权率，即回答“无法确定”的探测项所占比例。关于位深度、文件格式和帧数的消融实验见附录F。

4. 实验

4.1. 设置

我们评估了 11 个主流多模态模型：Claude Opus 5.5（Anthropic，2026a）、Claude Sonnet 5.5（Anthropic，2026b）、GPT-6.1 Sol（OpenAI，2026a）、GPT-6 Astra（OpenAI，2026b）、Gemini 3.1 Pro（Google DeepMind，2026a）、Gemini 3.8 Flash（Google DeepMind，2026b）、Qwen3.8-Max（阿里云，2026b）、Qwen3.8-Flash（阿里云，2026a）、Kimi K3.1 Pro（月之暗面，2026）、Kimi K3（Kimi 团队，2026）和 DeepSeek-V4.1-Flash（DeepSeek，2026）。所有模型接收相同的图像和提示词，推理力度统一设置为 `high`，输出上限设置为 32k 词元。

我们将结果与四种基线进行比较。随机猜测在各个选项中均匀随机选择。始终选择 *D-Log*对每个项目均回答 D-Log，即真实样本最多的曲线，其得分仅反映类别不平衡所能带来的效果。公式基线和第小型分类器（第3.3 节）从不输出“无法确定”，也不参与探测。

4.2. 真实视频的主要结果

表3 显示，当前模型在真实视频素材上的表现距离实用水平仍相差甚远。在纯文本对照中，所有模型在 99–100% 的调用中都选择“无法确定”，因此问题和选项并不能提供捷径。仅使用帧时，11 个模型的曲线准确率介于 0.06 至 0.20 之间。始终回答

D-Log（在 330个片段中占 54 个）的得分为 0.16；表现最好的模型 Gemini 3.8 Flash 和 Kimi K3.1 Pro 仅高出 4 个百分点，而 11 个模型中有 7 个低于这一水平。按曲线平均值计算，始终回答 D-Log 的得分为 1/13 ≈ 0.08，模型得分则为 0.05 至 0.14。品牌识别同样接近频率基线：模型得分为 0.14 至 0.22，而始终回答 DJI 的得分为 0.16。使用规格摘要和测量指南后，只有 GPT-6 Astra（0.27）和 GPT-6.1 Sol（0.25）的表现显著提升，分别提高 13 和 11 个百分点；其余 9 个模型的变化介于 −0.05 至 +0.04 之间。

答案偏好与弃答。 输出分布（附录 E，表7）揭示了明显的频率先验。仅提供帧时，11 个模型中有 10 个最常预测 S-Log3，占其输出的 27–60%，尽管该曲线仅占真实集合的 9%；只有 Claude Sonnet 5.5 最常预测 ARRI LogC3，占其输出的 39%。GPT-6 Astra 和 GPT-6.1 Sol 则分别在 46% 和 32% 的调用中选择“无法确定”；仅统计未弃答的调用时，其条件准确率达到 0.26 和 0.21（附录E，表 7），高于最佳总体准确率 0.20。提供规格和指南后，它们的弃答率降至 6–7%，并在 Log3G10 和 C-Log3 等较不常见的曲线上取得明显领先。

各编码的覆盖情况及与基线的比较。正确预测集中在少数常见曲线上，例如 D-Log、S-Log3 和 Apple Log；对于大多数其他编码，准确率接近于零（附录 E，表 8）；LogC3 和 LogC4 上的最高分分别为 0.32 和 0.14。对于其余 8 种编码（共 188 个视频片段），没有任何模型在单一编码上的得分超过 0.11；在 Apple Log 2 和 F-Log2 C 上，得分均未超过 0.03；对于 30 个 Apple Log 2 视频片段，11 个模型总共只给出了 1 个正确答案。提供规格和指南后，两个 GPT 模型在 Log3G10 上的得分跃升至 0.45 和 0.34，而其他模型在该编码上的得分均未超过 0.10；它们在 C-Log3 上的得分也分别升至 0.31 和 0.40。相比之下，强大的模型（GPT-6 Astra 为 0.97，GPT-6.1 Sol 为 0.94）和小型分类器（0.96）都能可靠地检测出 77 个分级负样本。两个不依赖大型模型的基线同样受到真实场景中曝光和成像变化的严重影响。公式基线将 18%灰处的帧平均亮度作为默认曝光；在这一设置下生成的合成自然场景集上，其得分达到 0.71（第 4.4 节），但在真实集合上仅为 0.09，因为曝光会随场景自由变化。小型分类器在真实视频素材上的得分仅为 0.14：合成监督能教会神经网络判断视频素材是否经过调色，却无法让它在真实噪声和未知曝光条件下区分不同曲线。

4.3. 模型为何失败

真实样本上的正确答案可能来自对像素代码值进行光度测量，也可能是根据航拍视角等场景特征猜测品牌，或是退而采用常见的编码名称；四项诊断测试可以将这些情况区分开来。

探测测试。 移除选项中的正确曲线后，如果模型是通过验证像素来作答，就应该选择“无法确定”。然而，仅提供帧时，除 GPT-6 Astra（0.46）和 GPT-6.1 Sol（0.33）外，所有模型的弃答率都不超过 0.03（表 4）。提供规格后，GPT 模型的弃答率降至 0.07–0.12，而最高的 Qwen3.8-Flash 也仅达到 0.17。即使是常规曲线问题中答对的样本，GPT 模型在 80–83% 的探测测试中会改选错误曲线，其他 9 个模型则在 95–100% 的探测测试中会改选错误曲线；对全部 11 个模型而言，最常改选的是 D-Log M，它与 D-Log 属于同一品牌。因此，先前的大多数正确答案都是根据场景和品牌联想得出的先验猜测。

同一品牌的相似编码。按品牌划分的统计数据（附录E，表 9）揭示了两种机制。第一种是识别品牌后混淆编码：在仅提供帧的 54 个真实 D-Log 视频片段中，11 个模型合计以 0.71 的比例识别出 DJI 品牌，但有 15% 的答案选择了对比度更高的 D-Log M；其中 Qwen3.8-Max、Claude Opus 5.5 和 Qwen3.8-Flash 的这一比例分别达到 59%、35%

表 4： 探测与重新编码诊断。弃答：对 330 个真实视频片段的探测项的弃答率。切换：在常规曲线问题中答对、但在探测中切换为错误曲线的回答所占比例。重新编码：对 66 张重新编码的 Log 图像的准确率和相同答案占比。切换和重新编码均仅使用帧； + 指南表示附有测量指南的规格。粗体表示表现最佳的多模态模型。

	弃答 ↑			重新编码	
方法／模型	帧	+指南	切换 ↓	准确率 ↑	相同 ↓
公式基线	–	–	–	0.17	0.52
小型分类器	–	–	–	0.45	0.36
Claude Opus 5.5	0.00	0.00	100%	0.05	0.74
Claude Sonnet 5.5	0.00	0.00	100%	**0.08**	0.68
GPT-6.1 Sol	0.33	0.07	83%	0.03	0.80
GPT-6 Astra	**0.46**	0.12	**80%**	0.02	0.89
Gemini 3.1 Pro	0.00	0.00	100%	0.06	0.67
Gemini 3.8 Flash	0.00	0.00	100%	**0.08**	0.73
Qwen3.8-Max	0.00	0.00	95%	0.03	0.61
Qwen3.8-Flash	0.03	**0.17**	97%	0.07	**0.55**
Kimi K3.1 Pro	0.00	0.00	100%	**0.08**	0.68
Kimi K3	0.00	0.00	99%	0.07	0.64
DeepSeek-V4.1-Flash	0.01	0.01	100%	**0.08**	0.58

分别为 30%。第二种情况是，品牌无法识别时，模型会偏向更熟悉、较旧的编码：除 DJI（0.71）和 Sony（0.58）外，其余 8 个品牌的汇总识别率均不超过 0.07。对于 Apple Log 2，品牌识别率仅为 0.02，模型几乎从未答对，但有 17% 的答案选择了较旧的 Apple Log，其中 Claude Sonnet 5.5 的比例高达 43%；对于 LogC4，也有 8% 的答案选择了较旧的 LogC3，Claude Sonnet 5.5 和 Kimi K3 的比例分别为 32% 和 23%。

品牌与曲线解耦子集。在使用尼康相机机身拍摄的 35 个真实 RED Log3G10 片段中（附录 E，表 9），11 个模型的汇总曲线准确率仅为 0.03，品牌准确率为 0.01；单个模型中最高的曲线准确率为 Claude Opus 5.5 的 0.13。模型的预测退回到常见的 S-Log3 和 LogC3，比例分别为 35% 和 32%。一旦缺少航拍画面等强烈的场景线索，模型便会退回到对编码的频率先验。

重新编码。 将一帧真实画面反转，并使用 11 条官方曲线重新编码，可得到 11 个内容相同、真实标签不同的版本。 相同答案占比 指这 11 个版本中，出现频率最高的预测所占的比例：当所有版本都回答正确时，该值为 $1/11 \approx 0.09$；当模型在所有编码中重复相同答案时，该值为 1。因此，该值越低，说明预测越能反映像素代码值。公式基线和小型分类器的准确率分别为 0.17 和 0.45，相同答案占比分别为 0.52 和 0.36（表 4）。11 个多模态模型的准确率仅为 0.02–0.08，相同答案占比为 0.55–0.89；GPT 模型的相同答案占比较高，主要是因为它们在约 70–80% 的调用中回答“无法确定”。

4.4. 灰卡和规格摘要的作用

表5 揭示了两个瓶颈：画面中缺少光度参考物，以及未能应用书面规格。

画面中有经校准的图表时，规格摘要能带来显著帮助。在合成图表

表5： 光度参考物和规格摘要的影响。在三个提示层级下，对50张合成图表、30张真实图表以及550张合成自然场景Log图像（每条曲线50张）的曲线识别准确率；“+ 指南”表示规格加测量指南。粗体标示每列中表现最佳的多模态模型。

	合成图表 ↑			真实图表，每条曲线的平均值 ↑			合成自然场景 ↑		
方法／模型	帧	+ 规格	+ 指南	帧	+规格	+ 指南	帧	+规格	+ 指南
公式基线，已知色块		1.00			0.83			–	
公式基线，盲测		–			–			0.71	
小型分类器		0.08			–			0.26	
Claude Opus 5.5	0.42	0.47	0.42	0.04	0.13	0.13	0.11	0.13	0.12
Claude Sonnet 5.5	0.30	0.44	0.28	0.04	0.23	0.13	0.09	0.09	0.09
GPT-6.1 Sol	**0.46**	0.68	0.67	0.22	0.43	0.37	0.04	**0.16**	**0.19**
GPT-6 Astra	0.39	**0.75**	**0.82**	**0.30**	**0.46**	**0.56**	0.03	0.15	0.17
Gemini 3.1 Pro	0.14	0.19	0.21	0.21	0.11	0.08	**0.12**	0.10	0.11
Gemini 3.8 Flash	0.19	0.19	0.19	0.22	0.08	0.16	**0.12**	0.13	0.13
Qwen3.8-Max	0.05	0.05	0.05	0.02	0.00	0.00	0.06	0.05	0.06
Qwen3.8-Flash	0.09	0.15	0.18	0.03	0.13	0.14	0.08	0.11	0.08
Kimi K3.1 Pro	0.22	0.22	0.27	0.17	0.11	0.06	0.11	0.13	0.13
Kimi K3	0.15	0.14	0.19	0.08	0.07	0.02	0.11	0.10	0.09
DeepSeek-V4.1-Flash	0.08	0.12	0.10	0.05	0.19	0.16	0.10	0.11	0.11

数据集上，公式基线可读取已知位置的色块，达到 1.00。仅帧条件下，表现最佳的模型是 GPT-6.1 Sol（0.46）、Claude Opus 5.5（0.42）和 GPT-6 Astra（0.39）。加入规格摘要后，GPT-6 Astra 升至 0.75–0.82，GPT-6.1 Sol 升至 0.67–0.68；仅凭规格摘要，两个 Claude 模型也达到 0.44–0.47。因此，前沿模型能够估计色块的代码值，并结合参数表进行推理。

在真实图表上，只有 GPT 模型从规格信息中显著受益。30 个真实图表帧中，LogC4 占 15 个，每种灵敏度设置各 1 个；其他编码各占 1–2 个，因此真实图表列报告的是按曲线平均值。尽管真实相机存在色彩再现误差、片场灯光和传感器噪声，读取已知位置色块的公式基线仍能正确回答 30 帧中的 28 帧，按曲线平均值为 0.83：只要帧中包含反射率和色度已知的参考物，像素本身就足以区分大多数曲线。仅帧输入时，表现最好的模型 GPT-6 Astra 得分为 0.30，GPT-6.1 Sol、Gemini 3.8 Flash 和 Gemini 3.1 Pro 得分为 0.21–0.22。加入规格摘要后，GPT-6 Astra 的得分提升至 0.46–0.56，GPT-6.1 Sol 则提升至 0.37–0.43。其他 9 个模型的得分均未超过 0.23；引入规格信息后，Gemini 3.1 Pro 和 Kimi K3.1 Pro 的得分则分别从 0.21 和 0.17 降至 0.06–0.11。

没有参考物的自然图像显示出明显的规格与测量差距。在合成自然场景集中，每条曲线对应的帧均取自相同的100个场景，因此图像语义无法提供捷径；该数据集的生成方式是将帧平均亮度对齐至18%灰，而将此作为默认曝光的盲公式基线达到了0.71。然而，所有多模态模型在仅帧条件下的得分均为0.03–0.12，接近随机猜测（0.06）。加入规格摘要后，GPT-6.1 Sol升至0.16–0.19，GPT-6 Astra升至0.15–0.17，其他模型的提升均不超过0.03，尽管测量指南明确列出了无需灰卡即可进行的检查，例如阴影底限、高光剪裁平台和饱和色域检查。因此，书面规格无法取代帧中的光度锚点：理解规格尚不能转化为可靠的像素测量。

5. 结论

作为MediaStormBench首个公开发布的子集，CameraLogBench研究多模态模型能否在没有元数据的情况下，仅凭帧识别相机的Log编码；该基准结合了330个专业拍摄片段，以及根据已发布的厂商公式制定的对照项和基线。在11个主流模型中，真实视频素材上的识别远未解决：仅帧条件下的最高准确率只有0.20，加入规格摘要和测量指南后升至0.27，而始终回答最常见曲线D-Log已能达到0.16。正确答案主要源于场景与品牌先验：移除正确选项后，9个模型的弃权率最高仅为0.03，且此前正确答案中的80–100%会切换为错误曲线。厂商规格将模型在合成图表上的最高准确率从0.46提升至0.82，但在没有参考物的合成自然场景上，没有任何模型超过0.19。因此，当前模型缺乏的并不是关于厂商曲线参数的知识，而是定量测量像素代码值的能力；CameraLogBench将这一能力与高层次语义理解分离开来，为基础模型的低层次定量视觉提供了清晰的衡量标准。

局限性与未来方向。 首先，在数据和输入格式方面，该基准测试涵盖未调色Log素材，并使用多模态模型API的标准输入——8位无损帧；对于6条官方曲线，真实视频素材均来自一次专门拍摄，且每条曲线仅使用一台摄影机机身。后续工作可以将真实视频素材扩展到更多制作项目和机身，支持原生10位输入，并还原调色素材的原始编码。更重要的是，评估方式应从单独测试模型转向在真实工作流中借助工具评估模型：本文中的所有模型均未使用任何外部工具进行评估，以确立模型自身的感知基线；而在实际后期制作中，调色师会依靠专业工具进行判断。未来的工作可以将工具调用和智能体框架引入评估，例如让模型通过吸管工具读取像素值、生成波形图和矢量示波器图、编写脚本依据物理约束筛选候选项，甚至在软件中应用候选色彩空间变换并检查转换后的图像，从而测试模型能否在接近真实制作流程的交互式工作流中可靠地完成任务。沿着这一方向，MediaStormBench 后续的子集将继续关注整个影视制作工作流中的其他真实问题。

参考资料

Adobe Inc. 在 Premiere Pro 中自动检测 Log 相机格式和原始媒体。 https://web.archive.org/web/20260210164247/https://helpx.adobe.com/uk/premiere-pro/using/auto-detection-of-log-camera-formats-and-raw-media.html，2025年。Adobe Premiere Pro用户指南，存档于2026年2月10日。

Adobe Inc. 色彩管理选项。https://helpx.adobe.com/premiere/desktop/correct-color/set-up-color-management/color-management-options.html，2026。Adobe Premiere用户指南。

阿里云。Qwen3.8-Flash 模型信息。https://www.alibabacloud.com/help/en/model-studio/qwen3-8-flash，2026a。阿里云模型工作室文档。

阿里云。Qwen3.8-Max 模型信息。https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max，2026b。阿里云模型工作室文档。

Anthropic。《Claude Opus 5.5》介绍。 https://www.anthropic.com/news/claude-opus-5-5，2026a。

Anthropic。《Claude Sonnet 5.5》介绍。 https://www.anthropic.com/news/claude-sonnet-5-5，2026b。苹果公司。《Apple Log 配置文件白皮书》。 https://download.developer.apple.com/Developer_Tools/Apple_Log_profile/Apple_Log_Profile_White_Paper.pdf，2023年。苹果公司。《Apple Log 2 白皮书》。 https://developer.apple.com/download/all/?q=Apple%20Log%202，2025年。ARRI。《ALEXA Log C 曲线——在视觉特效中的应用》。 https://www.arri.com/resource/blob/31918/66f56e6abb6e5b6553929edf9aa7483e/2017-03-alexa-logc-curve-in-vfx-data.pdf，2017年。ARRI。《ARRI LogC4 对数色彩空间规范》。 https://www.arri.com/resource/blob/278790/bea879ac0d041a925bed27a096ab3ec2/2022-05-arri-logc4-specification-data.pdf，2025年。Blackmagic Design。《Blackmagic 第五代色彩科学技术参考》。随 Blackmagic RAW SDK 一同发布；列于《Blackmagic RAW Player 2.1 发行说明》中。 https://www.blackmagicdesign.com/support/content/readme/ea11ce9660c642879612f363ca387c7f，2021年。佳能美国公司。《Canon Log 伽马曲线：Canon Log、Canon Log 2 和 Canon Log 3 伽马曲线说明》。 https://www.usa.canon.com/content/dam/canon-assets/white-papers/pro/white-paper-canon-log-gamma-curves.pdf，2018年。Davide Cozzolino、Giovanni Poggi 和 Luisa Verdoliva。《提取用于视频取证的基于相机的指纹》。载于《IEEE/CVF 计算机视觉与模式识别会议研讨会论文集》，第 130–137 页，2019年。URL https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Verdoliva_Extracting_camera-based_fingerprints_for_video_forensics_CVPRW_2019_paper.html。DeepSeek。《DeepSeek-V4.1-Flash：更智能、更快速、更高效》。 https://api-docs.deepseek.com/news/news260910/，2026。DJI。《DJI 电影色彩系统 D-Log 和 D-Gamut 白皮书：DJI Zenmuse X9 6K 和 8K》。 https://dl.djicdn.com/downloads/DJI_Ronin_4D/X9_D_Log_D_Gamut_Whitepaper.pdf，2022年。Eddie AI。《Log 素材、查找表和 Rec.709 转换》。https://www.heyeddie.ai/help/log-footage-luts-and-rec709，2026。Mark D. Fairchild。《HDR 摄影调查》。载于《色彩与成像会议》，第 15 卷，第 233–238 页。成像科学与技术学会，2007年。doi: 10.2352/CIC.2007.15.1.art00044。I-Sheng Fang 和 Jun-Cheng Chen。《CameraBench：通过摄影对多模态大语言模型中的视觉推理进行基准测试》。arXiv 预印本 arXiv:2504.10090，2025年。doi: 10.48550/arXiv.2504.10090。富士胶片株式会社。《F-Log2 数据表 1.0版》。 https://dl.fujifilm-x.com/support/lut/F-Log2_DataSheet_E_Ver.1.0.pdf，2022年。富士胶片株式会社。《F-Log2 C 数据表 1.0版》。 https://dl.fujifilm-x.com/support/lut/F-Log2C_DataSheet_E_Ver.1.0.pdf，2024年。

Thomas Gloe 和 Rainer Böhme。《德累斯顿图像数据库》：用于数字图像取证基准测试。载于*2010*年*ACM*应用计算研讨会论文集，第 1584–1590 页，2010。doi：10.1145/ 1774088.1774427。

Google DeepMind。Gemini 3.1 Pro 模型卡。 https://deepmind.google/models/model-cards/gemini-3-1-pro/，2026a。

Google DeepMind。Gemini 3.8 Flash 模型卡。 https://deepmind.google/models/model-cards/gemini-3-8-flash/，2026b。

Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让视觉问答中的“视觉”发挥作用：提升图像理解在视觉问答中的作用。载于*IEEE*计算机视觉与模式识别会议论文集，第 6325–6334 页，2017年。doi：10.1109/CVPR.2017.670。

Michael D. Grossberg 和 Shree K. Nayar。相机响应函数的空间是什么？载于 *2003*年 *IEEE*计算机学会计算机视觉与模式识别会议论文集，第 2 卷，第 602–609 页。IEEE，2003年。doi: 10.1109/CVPR.2003.1211522。

Yuchen Guo、Junli Gong、Hongmin Cai、Yiu-ming Cheung 和 Weifeng Su。LumiVideo：用于视频调色的智能代理系统。*arXiv* 预印本 *arXiv:2604.02409*，2026。doi: 10.48550/arXiv.2604.02409。

Brian C. Hosler、Owen Mayer、Belhassen Bayar、Xinwei Zhao、Chen Chen、James A. Shackleford 和 Matthew C. Stamm。使用深度学习与融合的视频相机型号识别系统。载于 *ICASSP 2019 - 2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 8271–8275 页。IEEE，2019年。doi: 10.1109/ICASSP.2019.8682608。

Shima Imani、Seungwhan Moon、Adel Ahmadyan、Lu Zhang、Kirmani Ahmed 和Babak Damavandi。PRiSM：一种通过基于Python的评估开展科学推理的智能体多模态基准。*arXiv* 预印本 *arXiv:2512.05930*，2025年。doi：10.48550/arXiv.2512.05930。

Justin Johnson、Bharath Hariharan、Laurens van der Maaten、Li Fei-Fei、C. Lawrence Zitnick 和 Ross Girshick。CLEVR：用于组合语言和基础视觉推理的诊断数据集。载于*IEEE*计算机视觉与模式识别会议论文集，第1988–1997页，2017年。doi：10.1109/CVPR.2017.215。

Kimi 团队。Kimi K3：开放前沿智能。技术报告。 https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf，2026。API 文档： https://platform.kimi.ai/docs/guide/kimi-k3-quickstart。

Kinefinity。KineLog3技术规格。https://kinefinity.com/support/guides/kinelog3-technical-specifications，2025年。

Iuliia Kotseruba 和 John K. Tsotsos。SNAP：用于测试采集条件对基础视觉任务影响的基准测试。*arXiv* 预印本 *arXiv:2505.15628*，2025年。doi：10.48550/arXiv.2505.15628。

Han Li 和 Pieter Peers。CRF-net：使用卷积神经网络进行单图像辐射度校准。载于第*14*届欧洲视觉媒体制作会议（*CVMP 2017*）论文集，第1–9页，2017年。doi：10.1145/3150165。3150170。

Puyin Li、Tiange Xiang、Ella Mao、Shirley Wei、Xinye Chen、Adnan Masood、Li Fei-Fei 和 Ehsan Adeli。《QuantiPhy：评估视觉语言模型物理推理能力的定量基准测试》

模型。在 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 33174–33184 页，2026 年。网址 https://openaccess.thecvf.com/content/CVPR2026/html/Puyin_QUANTIPHY_A_Quantitative_Benchmark_Evaluating_Physical_Reasoning_Abilities_of_Vision-Language_CVPR_2026_paper.html。

Yijun Liang、Ming Li、Chenrui Fan、Ziyue Li、Dang Nguyen、Kwesi Cobbina、Shweta Bhardwaj、Jiuhai Chen、Fuxiao Liu 和 Tianyi Zhou。ColorBench：视觉语言模型能否看见并理解丰富多彩的世界？一个针对颜色感知、推理和鲁棒性的综合基准测试。载于神经信息处理系统进展，第 38 卷，第 159145–159191 页，2025年。doi: 10.52202/085713-4803。

Yu-Lun Liu、Wei-Sheng Lai、Yu-Sheng Chen、Yi-Lung Kao、Ming-Hsuan Yang、Yung-Yu Chuang 和 Jia-Bin Huang。通过学习逆转相机成像流程进行单幅图像 HDR 重建。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 1648–1657 页，2020年。doi：10.1109/CVPR42600.2020.00172。

Yuan Liu、Haodong Duan、Yuanhan Zhang、Bo Li、Songyang Zhang、Wangbo Zhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu、Kai Chen 和 Dahua Lin。MMBench：你的多模态模型是全能选手吗？载于欧洲计算机视觉会议，第 216–233 页。Springer，2024年。doi: 10.1007/ 978-3-031-72658-3 13._

Ahmed Masry、Do Xuan Long、Jia Qing Tan、Shafiq Joty 和 Enamul Hoque。ChartQA：一项通过视觉和逻辑推理回答有关色卡问题的基准测试。载于计算语言学协会论文集：*ACL 2022*年成果，第 2263–2279 页。计算语言学协会，2022年。doi: 10.18653/v1/2022.findings-acl.177。

月之暗面。Kimi K3.1 Pro。https://platform.kimi.ai/docs/guide/kimi-k3-1-pro-quickstart, 2026。Kimi API 平台文档。

尼康公司。N-Log 规格文档，版本 1.0.0。 https://download.nikonimglib.com/archive3/hDCmK00m9JDI03RPruD74xpoU905/N-Log_Specification_(En)01.pdf，2018年。

OpenAI。GPT-6.1 Sol。 https://developers.openai.com/api/docs/models/gpt-6.1-sol，2026a。OpenAI API 模型文档。

OpenAI。GPT-6 Astra。 https://developers.openai.com/api/docs/models/gpt-6-astra, 2026b。OpenAI API 模型文档。

Larry Pearlstein、Alexander Benasutti、Skyler Maxwell、Matthew Kilcher、Jake Bezold 和 Warren Seto。利用卷积神经网络检索色彩空间转换矩阵。《国际机器学习与计算期刊》，9(4):393–400，2019年。doi: 10.18178/ijmlc.2019.9.4.816。

Poly Haven。Poly Haven：公共 3D 资产库。 https://polyhaven.com，2026。

Daiqing Qi、Handong Zhao、Jing Shi、Simon Jenni、Yifei Fan、Franck Dernoncourt、Scott Cohen 和 Sheng Li。摄影师之眼：教多模态大语言模型像摄影师一样观察和批评。在 *IEEE/CVF* 计算机视觉与模式识别会议论文集中，第 24807–24816 页，2025年。doi：10.1109/CVPR52734.2025.02310。

RED Digital Cinema。关于 REDWideGamutRGB 和 Log3G10 的白皮书。https://docs.red.com/955-0187/PDF/915-0187%20Rev-C%20%20%20RED%20OPS%2C%20White%20Paper%20on%20REDWideGamutRGB%20and%20Log3G10.pdf，2017年。

索尼公司。S-Gamut3.Cine/S-Log3 和 S-Gamut3/S-Log3 技术摘要。 https://download.pro.sony/FNGP/protein/1237494271390/1237494271406.pdf，2014年。

岳翔、倪元盛、张凯、郑天宇、刘若琪、张格、Samuel Stevens、蒋东福、任伟明、孙宇轩、魏聪、于博涛、袁瑞彬、孙仁良、殷明、郑博远、杨珍珠、刘一博、黄文浩、孙欢、苏瑜和陈文虎。MMMU：面向通用人工智能专家的大规模跨学科多模态理解与推理基准测试。发表于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 9556–9567 页，2024年。doi：10.1109/CVPR52733.2024.00913。

邹成科、郭兴刚、杨睿、张俊宇、胡斌和张欢。Dy-naMath：用于评估视觉语言模型数学推理鲁棒性的动态视觉基准测试。发表于国际学习表征会议，第 48337–48383 页，2025年。网址 https://proceedings.iclr.cc/paper_files/paper/2025/hash/78b248ea6f627431bba5029d92be8a3d-Abstract-Conference.html。

A. 曲线和厂商文档的关键值

代码值采用 10 位标度（0–1023），均根据厂商白皮书中公布的数学公式计算得出；每档斜率指在 18% 灰卡附近，场景光每加倍一次时代码值的增量。

曲线	黑位	18%灰	斜率（每挡 CV）	配对色域	厂商文档
ARRI LogC3	95	400	72	ARRI Wide Gamut 3	ARRI，2025年
ARRI LogC4	95	285	57	ARRI Wide Gamut 4	ARRI，2025年
Sony S-Log3	95	420	75	S-Gamut3.Cine	索尼公司，2014年
DJI D-Log	95	408	74	D-Gamut	DJI，2022年
RED Log3G10	94	341	63	REDWideGamutRGB	RED Digital Cinema，2017年
Canon C-Log2	95	407	70	Cinema Gamut	佳能美国公司，2018年
Canon C-Log3	128	351	85	Cinema Gamut	佳能美国公司,2018年
Blackmagic Film Gen5	95	392	60	Blackmagic Wide Gamut	Blackmagic Design，2021年
Apple Log	154	500	83	Rec.2020	苹果公司，2023年
Fujifilm F-Log2	95	400	71	F-Gamut	富士胶片株式会社，2022年
Nikon N-Log	127	372	83	Rec.2020	尼康公司，2018年
Kinefinity KineLOG3	95	401	76	Kinefinity 宽色域	Kinefinity，2025年

在曲线问题的 15 种编码选项中，Apple Log 2（苹果公司，2025年）和 Fujifilm F- Log2 C（富士胶片株式会社，2024年）分别沿用了 Apple Log 和 F-Log2 的传递曲线，并搭配更宽的色域（Apple 广色域和 F-Gamut C），因此共有 14 种编码公布了公式。剩下的选项 DJI D-Log M 没有公布公式，因此未列入表格，基准测试中也没有以其为真实标签的题目（第 3.1节）。

B. 输入帧与 10 位原始帧的比较

对于全部 330 个真实片段，我们将 3840×2160 10 位原始视频帧和提供给模型的 1920×1080 8 位无损帧转换到 10 位代码值尺度，并逐帧比较，同时汇总 R、G 和 B 通道：

统计量	330 个片段的均值	330 个片段的最大偏差
帧均值的变化	−0.5	1.3
帧中位数的变化	−0.5	3.7
最暗1%的变化	−0.3	6.0
最亮1%的变化	0.0	6.8
相对于未舍入的缩小帧的逐像素误差	1.4（均方根）	4.0

在全部330个片段中，无损存储图像与未压缩的8位图像逐像素完全一致。逐像素均方根误差不超过一个8位量化步长，即10位尺度上的4个代码值；而最暗和最亮的1%百分位数处的偏移，则来自 2×2 区域平均下采样，它会平滑孤立的极端噪声像素。

C. 数据示例

在真实图表集中，每台摄影机机身都会以常用的基础感光度拍摄一帧。Sony BURANO、Canon EOS R5 Mark II 和 Nikon ZR 各增加一种高感光度设置，而 ARRI ALEXA 35 则涵盖了从曝光指数 200 到曝光指数 6400 的全部 15 种设置（表 6）。改变感光度会影响图像的噪声水平和传感器的高光剪切点，但除 LogC3 外，传递曲线本身的数学定义不会随感光度改变；因此，可以利用同一曲线在不同感光度下拍摄的色卡来检验识别的稳定性。读取已知位置色块的公式基线在这四款机身的每一种感光度设置下都能正确识别——包括 ALEXA 35 的全部 15 种曝光指数设置，以及 BURANO、EOS R5 Mark II（两条曲线）和 ZR 的两种 ISO 设置——因此其判定不会随感光度而变化。

表 6：真实图表集的相机机身与灵敏度设置：10 台机身、30 帧，每种灵敏度设置对应一帧。

摄影机机身	曲线	感光度	帧
ARRI ALEXA 35	LogC4	曝光指数 200–6400，15 种设置	15
ARRI ALEXA Mini	LogC3	EI 800	1
Sony BURANO	S-Log3	ISO 800, 3200	2
Canon EOS R5 Mark II	C-Log2, C-Log3	ISO 800、4000	4
RED V-RAPTOR 8K VV	Log3G10	ISO 800	1
Nikon ZR	N-Log	ISO 800、6400	2
Blackmagic URSA Cine 17K Film Gen5		ISO 800	1
DJI Ronin 4D	D-Log	ISO 800	1
Apple iPhone 18 Pro	Apple Log，Apple Log 2	ISO 200	2
Fujifilm GFX100 II	F-Log2 C	ISO 800	1

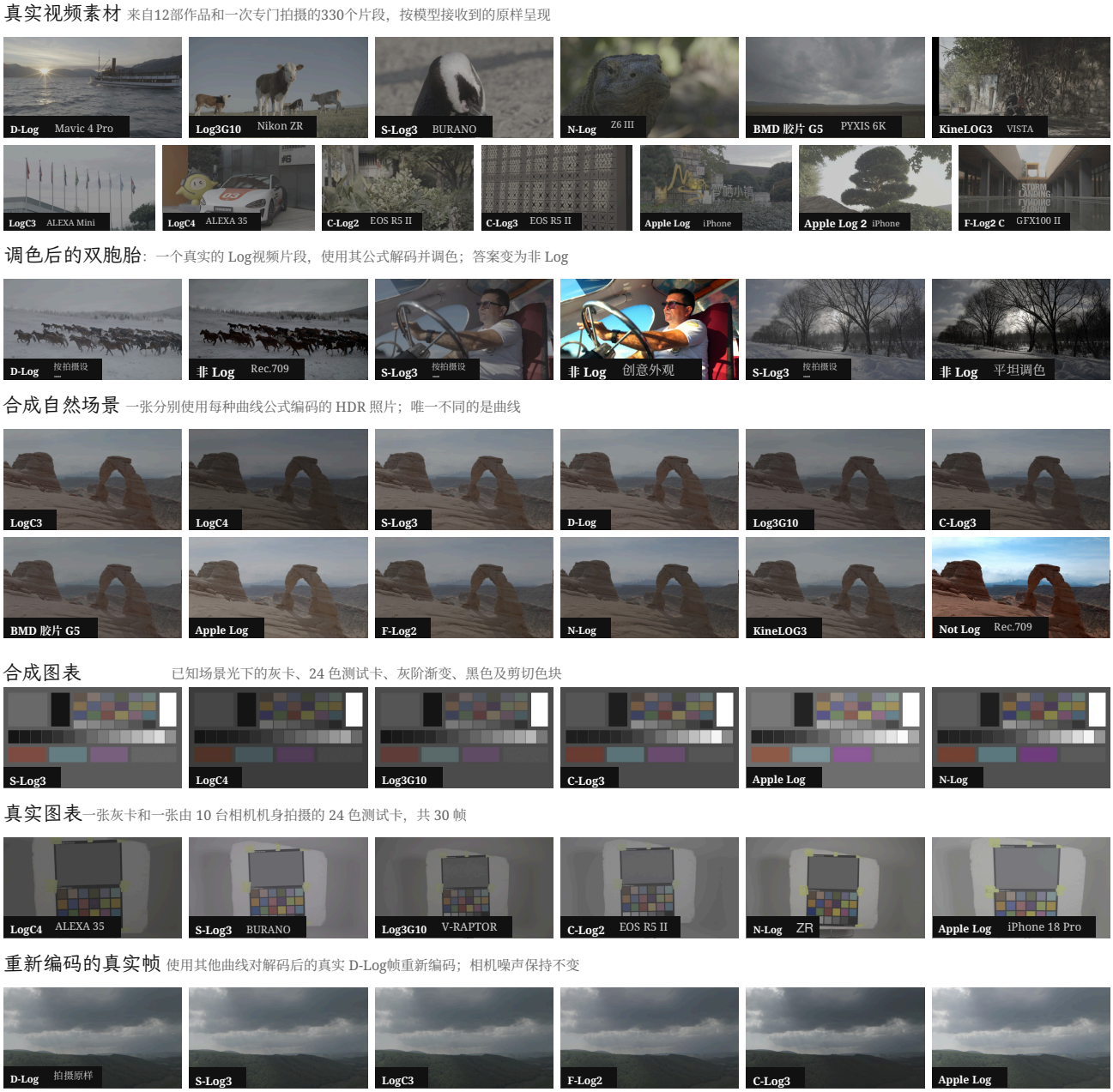


图 4：基准测试中各类数据的输入帧示例，均直接以提供给模型的 8位无损原始码值显示。**第 1–2 行（真实片段）：**13 种真实编码各选取一个代表性片段；第 1 行来自真实制作，第 2 行来自专门拍摄。**第 3 行（分级负样本）：**三个原始真实片段及其标准 Rec.709 转换、创意外观和低对比度仿对数版本，其真实标签始终为“非 Log”。**第 4–5 行（合成自然场景）：**一张 HDR 照片采用 11 种官方公式编码，最后一个单元格为 Rec.709 非对数对照。**第 6 行（合成图表）：**程序生成的校准图表，包含灰卡、24 色块色卡、灰阶渐变、纯黑色块和剪裁色块。**第 7 行（真实图表）：**在片场灯光下，使用不同相机机身拍摄的实体灰卡和 24 色块色卡。**第 8 行（重新编码）：**将一个真实 D-Log帧反转并用其他候选曲线重新编码，同时保留真实传感器噪声。

D. 一个完整的演示示例

本节展示了一个真实曲线识别题目的输入帧，以及模型完整的推理输出。该片段由 DJI Inspire 3 拍摄，其真实标签为 DJI D-Log。测试模型为 Claude Opus 5.5。仅凭帧画面时，模型根据航拍视角识别出 DJI 设备，但错误地选择了同一品牌的 D-Log M。在下方所示的 + 规格 +指南层级中，模型结合阴影底限的代码值和中间调的对比度斜率，排除 D-Log M，并在两种选项顺序下均选择了正确选项 D。



输入帧。从片段时长的 25%、50% 和 75% 处截取的三帧 8位无损帧。

模型输出原文。提示词模板见附录G；在此选项顺序下，选项 D 为 DJI D-Log， E 为 D-Log M，模型最终选择了 D ，并将 E 作为第二选择。


```
{"open_answer": "DJI drone footage in a flat log profile,  most likely DJI D-Log / D-Gamut,
 with D-Log M asthe close alternative.", "evidence": [
"The content points to a DJI drone: an aerial view over a valley with a railway viaduct,  with a slow
camera move across the three frames.", "The image is very flat and desaturated,
 with no display contrast curve. The pink sunset and blue skypatch show only weak saturation,
 which suggests a log encoding plus a wide gamut rather than Rec.709.",
"The darkest foreground ground and trees average about 8-bit 35-45,  roughly 140-180 on the 10-bit scale.
This sits above the 95 black family with no true black anywhere in frame. That is consistent withD-Log,
 while D-Log M would let dusk shadows fall lower.",
"The large mid-tone cloud mass sits around 8-bit 120-130 (about 480-520 10-bit). That is near or
slightly above the D-Log grey of 408 for a bright overcast sky.",
"The brightest areas (horizon glow and the blue sky patch) reach about 8-bit 190-200 (about 760-800 10-bit),
 with no hard clip plateau. This fits below the D-Log clip of about 948.",
"The tonal range of roughly 160 to 800 is compressed with low contrast per stop,  which favours true
D-Log over the higher-contrast D-Log M."],"top2": ["D", "E"],"choice": "D"}
```


E. 答案分布及先前捷径的详细信息

以下三个表格给出了第4.2节和第4.3节中各模型的详细结果：每个模型在真实素材曲线问题上的输出分布；逐编码召回率；以及按品牌划分的识别率、同一品牌相似编码之间的混淆率，以及品牌与曲线解耦子集上的结果。

表 7：仅使用帧时，在 330 个 Log 片段的真实素材曲线问题上的输出分布；每行之和为 1。“其他真实编码”汇总了真实集合中的其他 11 种编码；“不在真实集合中”是 F-Log2，它没有真实视频片段；“无法判断”是指选项“无法判断”。最后一列是排除“无法判断”后，给出具体预测的调用的条件准确率。粗体表示最佳值。

	各类答案所占比例							准确率（当
模型	D-Log	S-Log3	其他真实编码	不在真实集合中	D-LogM	非 Log	无法判断	回答
Claude Opus 5.5	0.06	0.48	0.35	0.01	0.09	0.01	0.00	0.13
Claude Sonnet 5.5	0.14	0.08	0.69	0.06	0.02	0.01	0.00	0.15
GPT-6.1 Sol	0.11	0.36	0.18	0.00	0.03	0.00	0.32	0.21
GPT-6 Astra	0.11	0.27	0.13	0.00	0.03	0.00	0.46	**0.26**
Gemini 3.1 Pro	0.25	0.35	0.38	0.02	0.00	0.00	0.00	0.19
Gemini 3.8 Flash	0.17	0.45	0.30	0.04	0.03	0.01	0.00	0.20
Qwen3.8-Max	0.01	0.39	0.20	0.03	0.22	0.16	0.00	0.06
Qwen3.8-Flash	0.15	0.60	0.12	0.01	0.10	0.01	0.02	0.14
Kimi K3.1 Pro	0.15	0.54	0.25	0.01	0.04	0.00	0.00	0.20
Kimi K3	0.16	0.42	0.39	0.02	0.01	0.00	0.00	0.18
DeepSeek-V4.1-Flash	0.10	0.51	0.32	0.01	0.03	0.01	0.01	0.15

表8： 真实素材曲线问题中逐编码召回率，各列按真实片段数量降序排列：D-Log（54）、Log3G10（53）、F-Log2 C（31）、S-Log3（30）、Apple Log 2（AL 2；30）、LogC3（25）、LogC4（22）、C-Log3（21）、C-Log2（20）、Film Gen5（Gen5；16）、N-Log（12）、Apple Log（AL；11）和 KineLOG3（Kine；5）。粗体表示每个提示层级中各列的最佳值。

模型	D-Log	Log3G10	F-Log2 C	S-Log3	AL 2	LogC3	LogC4	C-Log3	C-Log2	Gen5	N-Log	AL	Kine
仅帧													
Claude Opus 5.5	0.34	**0.11**	0.00	0.47	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.18	0.00
Claude Sonnet 5.5	0.60	0.01	0.00	0.10	0.00	**0.32**	0.07	0.00	0.00	0.03	0.00	**0.41**	0.00
GPT-6.1 Sol	0.58	0.05	0.00	0.33	0.00	0.02	0.02	**0.07**	0.00	0.00	0.00	0.05	0.00
GPT-6 Astra	0.58	0.03	0.00	0.37	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.05	0.00
Gemini 3.1 Pro	0.74	0.00	0.00	0.42	0.00	0.04	0.11	0.00	**0.10**	0.00	**0.08**	0.27	0.00
Gemini 3.8 Flash	**0.78**	0.05	0.00	0.48	0.00	0.00	**0.14**	0.00	0.00	0.00	0.00	0.32	**0.10**
Qwen3.8-Max	0.01	0.00	0.00	0.48	0.00	0.14	0.00	0.00	0.00	0.03	0.00	0.00	0.00
Qwen3.8-Flash	0.39	0.00	0.00	**0.75**	0.00	0.06	0.02	0.00	0.00	0.03	0.00	0.00	0.00
Kimi K3.1 Pro	0.73	0.03	0.00	0.60	0.00	0.08	0.02	0.00	0.00	**0.09**	0.00	0.14	0.00
Kimi K3	0.68	0.02	0.00	0.57	0.00	0.16	0.00	0.00	0.00	0.03	0.00	0.09	0.00
DeepSeek-V4.1-Flash	0.45	0.00	**0.03**	0.52	**0.02**	0.18	0.02	0.00	0.00	**0.09**	0.00	0.00	0.00
帧 + 规格 +指南													
Claude Opus 5.5	0.56	0.04	0.00	**0.62**	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.27	0.00
Claude Sonnet 5.5	0.62	0.00	0.00	0.48	0.00	**0.16**	0.02	0.00	0.00	0.00	0.00	0.14	0.00
GPT-6.1 Sol	0.69	0.34	0.00	0.25	0.00	**0.16**	**0.20**	**0.40**	0.00	0.00	0.00	0.32	0.00
GPT-6 Astra	0.75	**0.45**	0.00	0.23	0.00	0.14	**0.20**	0.31	0.00	0.03	0.00	0.14	0.00
Gemini 3.1 Pro	0.50	0.07	0.00	0.12	0.00	0.06	0.00	0.26	0.00	0.06	0.00	0.45	0.00
Gemini 3.8 Flash	0.71	0.10	0.00	0.57	0.00	0.00	0.14	0.05	0.00	0.00	0.00	0.14	0.00
Qwen3.8-Max	0.11	0.00	0.00	0.08	0.00	0.06	0.00	0.00	0.00	0.00	0.04	**0.64**	0.00
Qwen3.8-Flash	0.63	0.01	0.00	0.38	**0.03**	0.02	0.00	0.00	0.00	0.03	0.00	0.41	0.00
Kimi K3.1 Pro	**0.81**	0.01	0.00	0.45	0.00	0.06	0.05	0.02	0.00	0.12	0.00	0.32	0.00
Kimi K3	0.62	0.00	0.00	0.42	0.00	0.02	0.00	0.00	0.00	0.00	**0.12**	0.36	0.00
DeepSeek-V4.1-Flash	0.67	0.00	0.00	0.35	0.02	0.14	0.00	0.00	0.00	**0.22**	0.00	0.41	0.00

表 9： 按品牌识别、同一品牌相似编码之间的混淆，以及品牌与曲线解耦子集（仅帧）。前四列报告了 330 个真实视频片段在品牌问题上的识别准确率，并按真实品牌分组：DJI（54 个片段）、Sony（30 个）、Apple（41 个），以及其余 7 个品牌合计（205 个）。中间三列报告了在常规曲线问题中，将真实曲线误判为同一品牌另一种编码的答案比例：将 D-Log M（54 个片段）误判为 D-Log、将 Apple Log（30 个）误判为 Apple Log 2，以及将 LogC3（22 个）误判为 LogC4。最后两列报告了由 Nikon 相机机身拍摄的 35 个 RED Log3G10 片段组成的解耦子集中的曲线准确率和品牌准确率。最后一行汇总了 11 个模型的结果。粗体表示每个准确率列中的最高值，以及每个混淆列中的最高比例。

	品牌识别准确率 ↑				与同品牌编码 Log3G10 混淆于 Nikon				
模型	DJI	Sony	Apple	其他 7	D-Log → D-Log M	Apple Log 2 → Apple Log	LogC4 → LogC3	曲线 ↑	品牌 ↑
Claude Opus 5.5	0.79	**0.92**	0.01	0.02	0.35	0.15	0.02	**0.13**	0.00
Claude Sonnet 5.5	0.77	0.67	0.00	0.04	0.07	**0.43**	**0.32**	0.00	0.00
GPT-6.1 Sol	0.65	0.52	0.00	0.00	0.06	0.10	0.00	0.06	0.00
GPT-6 Astra	0.62	0.38	0.00	0.00	0.06	0.12	0.00	0.04	0.00
Gemini 3.1 Pro	**0.81**	0.43	0.04	**0.07**	0.00	0.28	0.05	0.00	**0.07**
Gemini 3.8 Flash	0.69	0.50	0.00	0.04	0.02	0.23	0.00	0.04	0.00
Qwen3.8-Max	**0.81**	0.47	0.02	0.03	**0.59**	0.02	0.07	0.00	0.00
Qwen3.8-Flash	0.61	0.58	0.00	0.05	0.30	0.00	0.02	0.00	0.00
Kimi K3.1 Pro	0.69	0.58	**0.05**	0.03	0.05	0.18	0.09	0.04	0.00
Kimi K3	0.68	0.70	**0.05**	0.05	0.01	0.28	0.23	0.01	0.00
DeepSeek-V4.1-Flash	0.75	0.63	0.02	0.02	0.13	0.08	0.14	0.00	0.00
11 个模型汇总	0.71	0.58	0.02	0.03	0.15	0.17	0.08	0.03	0.01

F. 输入格式和帧数的消融实验

位深度和文件格式。在合成自然场景数据集的 550 张 Log 图像上，以 16 位原始图像、8 位无损图像和 8 位 JPEG 图像作为输入时，公式基线的准确率分别为 0.71、0.71 和 0.68；三种格式之间的差异不超过 3 个百分点。

三帧联合输入与独立单帧输入的比较。当每个真实视频片段的三帧被分开并独立提问时（表10），在1/4、1/2和3/4时间点的单帧准确率差异最多为0.02。与三帧联合输入相比，GPT-6.1 Sol（0。14 → 0.08–0.09）和 Kimi K3.1 Pro（0。20 → 0.14–0.16）的准确率下降，而 Claude Opus 5.5（0。13 → 0.14–0.15）和 Qwen3.8-Max（0。06 → 0.04–0.06）基本保持不变。同一视频片段的三个单帧问题仅在0.17–0.62的情况下得到相同答案，因此单帧预测在很大程度上取决于采样位置，而三帧联合输入能够提供更稳定的片段级表示。

表10： 在330个真实Log片段上进行的仅帧帧采样消融。比较三帧联合输入与在时长的1/4、1/2和3/4处分别取独立单帧时的曲线识别准确率，以及三个单帧问题得到相同答案的视频片段占比。每列中的最佳值以粗体标出。

	联合输入	单帧			Same
模型	3帧	1/4	1/2	3/4	答案
Claude Opus 5.5	0.13	**0.15**	0.14	0.14	0.48
GPT-6.1 Sol	0.14	0.09	0.08	0.08	**0.62**
Qwen3.8-Max	0.06	0.04	0.04	0.06	0.17
Kimi K3.1 Pro	**0.20**	0.14	**0.15**	**0.16**	0.31

G. 提示模板与探测选项

所有项目共用一个英文提示词模板；只有两个占位符——参考材料和选项——会根据提示层级和问题类型进行替换，并在下方模板中用方括号标出。表11列出了四个提示层级的输入及其诊断目的。

表 11： 四个提示层级的输入与诊断目的。我们在除帧 +规格之外的所有层级上评估真实项目，并在三个视觉层级上评估合成项目。

Tier	模型接收的内容	其评估的内容
仅文本	问题和选项，不含图像	问题的措辞以及 选项的措辞是否会造成偏差
仅帧	无损帧、问题和选- 项	仅凭模型的内部 知识进行识别
帧 + 规格	添加规格摘要，其中包含 仅包含公式和关键代码值	模型能否通过 自身测量像素，一旦它知道精确参数
帧 + 规格 + 指南	进一步添加六步测量— 量流程	识别（在测量方法确定后） 也一并提供

系统提示。

```
You are a senior DIT and colorist with deep, hands-on knowledge of camera log encodings, their transfer
functions and native colour gamuts. When images are attached, each one is an 8-bit screenshot of a
video frame exactly as the file stores it: we applied no LUT, colour-space transform, grade or displaytransform of our own,
 so pixel values are the file's code values scaled down to 8 bits (multiply by
about 4 to get 10-bit code values on a 0-1023 scale). The file itself may be a camera-original log recording,
 or it may already be display-referred (e.g. Rec.709 / a baked-in look); deciding which is
part of the task. Answer strictly with a single JSON object and nothing else.
```

用户提示。 下面的模板适用于曲线问题和探测；对于合成样本，第一句话会替换为对单张图像的描述。品牌问题会将问题替换为“最有可能录制这段视频的摄像机品牌是什么？还请列出最多三个具体的摄影机型号，按可能性从高到低排列，填入“camera_candidates”。”，将选项替换为 11 个品牌，并添加输出字段`camera candidates`。_

```
3 frames from the same clip are attached、 in time order、 taken at 25%、 50%、 75% of its duration. They
share one encoding; use all of them.
```

```
Reference material: below is a digest of the camera manufacturers' white papers (encoding formulas and key
    code values). Use its numeric cues to check your reading of the frame.
<reference>
[REFERENCE MATERIAL]
</reference>

Question: Which camera log encoding (transfer function) and native colour gamut was this footage most
    likely recorded in?

Work in this order:
1. Before reading the options, write your own free-text best guess of the encoding and gamut in
"open_answer".
2. List the cues you relied on in "evidence". Be quantitative wherever you can: estimated black level /
darkest code value, where a neutral mid-grey or skin tone sits, highlight clip or plateau code value,
hue bias of strongly saturated colours. Give code values on the 10-bit scale (0-1023) or normalised 0-1.
3. Pick your two most likely options ("top2", most likely first) and your single final answer ("choice",
normally the first entry of "top2"). Use option letters only.

Options:
[OPTIONS]

Respond with a single JSON object and nothing else:
{"open_answer": "<free-text best guess, made before reading the options>",
"evidence": ["<observable cue 1>", "<cue 2>", "..."],
"top2": ["<letter>", "<letter>"],"choice": "<single letter>"}
```

参考材料。仅帧层级和仅文本层级不包含 `Reference material` 块。帧 + 规格层级补充了规格摘要的第 1–5 节，涵盖关键代码值、公式、色域原色、不同感光度下的裁切值以及来源。帧 +规格 +指南层级则进一步附上第 6–8 节，内容原文照录如下。

```
## 6. What actually separates the encodings

- **Black level families**: ~95 = ARRI LogC3 95, ARRI LogC4 95, Sony S-Log3 95, DJI D-Log 95, RED Log3G10 94, BMD Film Gen 5 95,
 Fujifilm F-Log2 95, Kinefinity KineLOG3 95, Canon Log 2 95, Fujifilm F-Log2 C 95; ~128 = Canon Log 3 128,
 Nikon N-Log 127; ~154 = Apple Log 154, Apple Log 2 154. Only true black
(lens cap, deep unlit shadow) shows it; a scene floor above the black level proves nothing, a floor
clearly *below* it rules the curve out.
- **18% grey**: ARRI LogC4 285 < RED Log3G10 341 < Canon Log 3 351 < Nikon N-Log 372 < BMD Film Gen 5 392
< ARRI LogC3 400 < Fujifilm F-Log2 400 < Fujifilm F-Log2 C 400 < Kinefinity KineLOG3 401 < Canon Log 2
407 < DJI D-Log 408 < Sony S-Log3 420 < Apple Log 500 < Apple Log 2 500. Needs a grey card or a
confidently known mid-tone; an exposure error of +-1 stop moves grey by 57-85 CV, more than most gaps.
- **Contrast (CV per stop at grey)**: ARRI LogC4 57, BMD Film Gen 5 60, RED Log3G10 63, Canon Log 2 70, Fujifilm F-Log2 71,
 Fujifilm F-Log2 C 71, ARRI LogC3 72, DJI D-Log 74, Sony S-Log3 74, Kinefinity KineLOG3 76, Nikon N-Log 83, Apple Log 83, Apple Log 2 83,
 Canon Log 3 85. D-Log M (community fit) is ~112, i.e. not a true log.
- **Clip behaviour**: LogC3 and D-Log clip levels move with EI (section 4). Log3G10 reaches 1.0 only at 184 (10 stops over grey),
 BMD Gen 5 at 223, LogC4 at 470, so their highlights usually clip well below 1023 at a camera-specific level.
- **Gamut (blue primary y)**: Kinefinity WG -0.224, AWG3 -0.102, Cinema Gamut -0.100, S-Gamut3.Cine -0.087, RWG -0.085, BMD WG -0.082,
 D-Gamut -0.080, S-Gamut3 -0.050, AWG4 -0.031, Rec.2020 +0.046. RWG (green y 1.494),
 Kinefinity WG (1.148) and Cinema Gamut (1.14) contain large imaginary regions. The
same saturated object therefore has different RGB code ratios in each encoding.
- **EI dependence**: only LogC3 changes the curve itself (section 4); D-Log changes only the clip level; all others are single curves.


## 7. How a DIT uses these numbers

1. Read code values with a 10-bit full-range scope (0-1023). If blacks sit near 64 and nothing exceeds 940,
 the file may have been squeezed to video levels: full-range code y then appears as 64 + 0.856 * y
    (black 95 -> 145, grey 400 -> 406); BMD camera ProRes is recorded this way.
2. Find the darkest flat area (lens cap, deep shadow under an object). A floor at 93-96 points to the 95-family,
 126-130 to Canon Log 3 / N-Log,  ~154 to Apple Log. Noise makes single pixels dip below
```

```
    black; use averages over flat patches.
3. Find a neutral reference (grey card, white card, overcast sky is NOT neutral-reliable). Compare its
code with the grey / white columns, remembering the exposure is unknown: use differences between two
known patches (e.g. grey vs. white = 2.32 stops) to estimate CV/stop, which is exposure-independent.
4. Look for a hard highlight plateau (many pixels at one identical value, often in all channels). Compare
its code with the documented clip levels (section 4). A plateau above a curve's maximum possible code rules that curve out.
5. Decode saturated, well-exposed patches with each candidate curve + gamut to XYZ. Physically real
surface colours stay inside the spectral locus; many negative or out-of-locus results mean the assumed
    gamut is too wide or the curve too steep. Too-small saturation everywhere hints the assumed gamut is too narrow.
6. Combine: black family + grey/CV-per-stop + clip level + gamut plausibility. Each cue alone is weak;
when they disagree,  report the top-2 candidates instead of a single answer.

Worked example (studio chart): averaged floor 128 on a lens-cap black patch, a grey card at ~372, and no
plateau anywhere -> the 128-family (Canon Log 3 / N-Log); grey 372 matches N-Log (Canon Log 3 would
put grey at 351; check with the white card: N-Log 603, Canon Log 3 577). Same chart with floor 95 and
grey 400 but no clip plateau: the 95-family stays open (LogC3 / F-Log2 first by grey, then S-Log3 420, D-Log 408, BMD 392, Log3G10 341,
 LogC4 285 only if the image is clearly under-exposed).

## 8. Caveats

- D-Log M is visibly higher-contrast than D-Log.
- Container colour tags (e.g. Rec.709 nclc) and re-exports say nothing about the original encoding.
```

选项。 该探测使用与常规曲线问题相同的帧和提示词，仅从选项中移除正确的编码。对于正确答案为 RED Log3G10 的问题，下表给出了第一组选项顺序下的两个选项列表。

曲线问题	探测	选项
A	A	ARRI LogC3 / ALEXA Wide Gamut（AWG3）
B	B	ARRI LogC4 / ALEXA Wide Gamut 4（AWG4）
C	C	Sony S-Log3 / S-Gamut3.Cine
D	D	DJI D-Log / D-Gamut
E	E	DJI D-Log M
F （答案）	已移除；稍后字母上移逐一	**RED Log3G10 / REDWideGamutRGB**
G	F	Canon C-Log2 / Cinema Gamut
H	G	Canon C-Log3 / Cinema Gamut
I	H	Blackmagic Film Gen 5 / Blackmagic Wide Gamut Gen 5
J	I	Apple Log / Rec.2020
K	J	Apple Log 2 / Apple 广色域
L	K	Fujifilm F-Log2 / F-Gamut（Rec.2020 原色）
M	L	Fujifilm F-Log2 C / F-Gamut C
N	M	Nikon N-Log / Rec.2020
O	N	Kinefinity KineLOG3 / Kinefinity 宽色域
P	O	不是 Log编码（已经是显示相关，例如 Rec.709 / sRGB）
Q	**P** (答案)	无法从这张图像中确定