Type 4 (Logic Trap) 样本构造与原理白皮书

核心目标： 构造 25 条“高仿”谣言，能够骗过 CLIP (Channel 2)，但被 Logic Engine (Channel 3) 拦截。

Part 1. 核心原理：为什么要这么找?

1.1 黄金法则 (The Golden Rule)

> **同主语，异属性 (Same Subject, Different Attribute)**

- **Type 3 (语义不符)** 是“指鹿为马”。图是鹿，文是马。CLIP 一眼就能看出来。
- **Type 4 (逻辑陷阱)** 是“白马非马”。图是白马，文是黑马。
  - **CLIP 的弱点：** CLIP 模型像一个“只看关键词”的浏览者，它对**实体（名词）**极其敏感，但对**属性（形容词、状态、否定词、隐喻）**很不敏感。
  - **我们的策略：** 利用这一点，构造**“实体高度一致，但属性截然相反”**的样本。

1.2 为什么 CLIP 检测不出来?

CLIP 是基于对比学习训练的，它的特征提取器倾向于**“词袋模型” (Bag-of-Words)** 的特性。

- 当你给它一张“正午故宫”图，配文“深夜故宫”。
- CLIP 看到图里有 Forbidden City，文里也有 Forbidden City。
- 它会认为两者**高度相关**，给出高分（Pass）。
- **只有我们的第三通道（逻辑层）**，通过提取 Time=Day vs Time=Night，才能发现这是谣言。

Part 2. 样本搜集 SOP (由浅入深)

请按照以下 **5 个梯度** 搜集样本，每个梯度 **5 张**。

梯度 1：时间光影陷阱 (The Temporal Trap)

难度： ⭐
原理： 欺负 CLIP 对光照特征的不敏感。

- **图片要求：** 必须有强烈的**正午阳光**（短影子、高光）或**漆黑夜景**。
- **文本构造：** 必须包含**反向时间词**。
- **执行清单：**

ID	图片内容 (Visual)	文本关键词 (Text Keywords)	Excel Meta (真值)	逻辑冲突点
T01	**正午**的故宫/天安门	"深夜", "月光", "静谧"	Meta_Time = Day	Day vs Night

T02	**阳光直射**的沙漠	"凌晨", "满天星斗"	Meta_Time = Day	Day vs Night
T03	**灯火通明**的夜上海	"清晨", "第一缕阳光"	Meta_Time = Night	Night vs Day

梯度 2：天气氛围陷阱 (The Environmental Trap)

难度： ⭐⭐ **原理：** 欺负 CLIP 忽略环境背景。

- **图片要求：** 极端天气特征（雪、积水、撑伞）。
- **文本构造：** 描述相反的季节或气象。
- **执行清单：**

ID	图片内容 (Visual)	文本关键词 (Text Keywords)	Excel Meta (真值)	逻辑冲突点
E01	**漫天大雪**的长城	"酷暑", "烈日", "炎热"	Meta_Weather = Snow	Snow vs Summer
E02	**暴雨**积水的街道	"干旱", "滴雨未下"	Meta_Weather = Rain	Rain vs Dry
E03	**晴空万里**的海滩	"台风", "暴风雨", "巨浪"	Meta_Weather = Sunny	Sunny vs Storm

梯度 3：数量与事实陷阱 (The Fact Trap)

难度： ⭐⭐⭐ **原理：** 欺负 CLIP 是“数盲”，分不清空和满。

- **图片要求：** 状态极端的场景（空无一人 vs 人山人海）。
- **文本构造：** 描述相反的状态。
- **执行清单：**

ID	图片内容 (Visual)	文本关键词 (Text Keywords)	Excel Meta (真值)	逻辑冲突点
F01	**空荡荡**的会议室/教室	"座无虚席", "拥挤", "人山人海"	Meta_Fact = Empty	Empty vs Crowded
F02	**枯黄**的树叶/草地	"春意盎然", "生机勃勃", "翠绿"	Meta_Fact = Withered	Withered vs Fresh

F03	**脏乱差**的垃圾堆	"一尘不染", "整洁", "卫生"	Meta_Fact = Dirty	Dirty vs Clean

梯度 4：地标实体陷阱 (The Entity Trap)

难度： ⭐⭐⭐⭐ **原理：** 利用外观相似性进行欺骗。

- **图片要求：** 长得像但名字不同的地标。
- **文本构造：** 指着 A 说 B。
- **执行清单：**

ID	图片内容 (Visual)	文本关键词 (Text Keywords)	Excel Meta (真值)	逻辑冲突点
L01	**广州塔** (小蛮腰)	"上海", "东方明珠"	Meta_Location = Guangzhou	GZ vs SH
L02	**伦敦塔桥** (双塔)	"金门大桥" (红色悬索)	Meta_Location = London	UK vs USA
L03	**白宫**	"国会大厦" (圆顶)	Meta_Location = White House	Object Mismatch

梯度 5：双关语与话题陷阱 (The Polysemy Trap) - 最强杀招

难度： ⭐⭐⭐⭐⭐ **原理：** **利用词语的多义性 (Polysemy)**。这是师兄建议的重点，也是 CLIP 最容易挂的地方。

- **操作核心：** 图和文都包含同一个词（如“牛”、“跳水”），但一个是**实体**，一个是**隐喻（金融/体育）**。
- **执行清单：**

ID	图片内容 (Visual)	文本描述 (Text Strategy)	为什么 CLIP 会挂?	Excel Meta (真值)
P01	一头真实的**公牛** (Bull)	"A股今日迎来**牛市 (Bull Market)**，全线飘红。"	CLIP 匹配到 "Bull"	Meta_Topic = Animal
P02	运动员在**跳水** (Diving)	"科技股今日**大跳水 (Diving)**	CLIP 匹配到 "Dive"	Meta_Topic = Sports

		，跌幅惨重。"		
P03	一只**黑天鹅** (Animal)	"金融市场爆发**黑天鹅 (Black Swan)** 事件。"	CLIP 匹配到 "Swan"	Meta_Topic = Animal
P04	真实的**战场/开火** (Fire)	"两名选秀歌手在网上**激烈交火 (Open Fire)** 互喷。"	CLIP 匹配到 "Fire"	Meta_Topic = War
P05	**寒冬**雪景 (Winter)	"互联网行业的**寒冬 (Winter)** 来了，大厂纷纷裁员。"	CLIP 匹配到 "Winter"	Meta_Topic = Nature

Part 3. 验收标准 (Quality Control)

每做完一张图，请自问三个问题：

1. **Ch1 会报警吗？** -> 图必须是原图，不能 P 过。（答案应为 No）
2. **Ch2 会报警吗？** -> 图和文必须有共同的关键词（如“牛”、“故宫”）。（答案应为 No，**必须骗过 CLIP**）
3. **Excel 填对了吗？** -> Meta_ 必须填图片里看到的真实情况，不要被文本带偏。（答案应为 Yes）

特别提示：
梯度 5 的样本（双关语）是本项目申请专利和竞赛的亮点案例，请务必找高清、典型的图片！