arXiv:2603.03714v1 [cs.CL] 4 Mar 2026

顺序并非布局：图像生成中的顺序到空间偏差 n

Yongkan g Zhang *[1,2] Zonglin Zhao *[1,3] Yuechen Zhang *[1,4] Fei Ding [5] Pei Li [2] Wenxuan Wan g [1]

摘要

我们研究了现代图像生成模型中的一种系统性偏差：文本中实体的提及顺序会无端决定空间布局和实体—角色绑定。我们将这一现象称为顺序到空间偏差（*OTS*），并表明它同时存在于文本到图像和图像到图像生成中，常常覆盖基于实体的线索，导致布局错误或分配错位。为量化 OTS，我们提出了 OTS-BENCH，通过成对提示词隔离顺序效应：每对提示词仅实体顺序不同，并从两个维度评估模型：同质化和正确性。实验表明，顺序到空间偏差（OTS）广泛存在于现代图像生成模型中，并提供证据表明，该偏差主要由数据驱动，在布局形成的早期阶段显现。基于这一认识，我们表明，定向微调和早期阶段干预策略都能在保持生成质量的同时，显著减少 OTS。

1. 引言

文本到图像（T2I）和图像到图像（I2I）扩散模型的迅猛发展，从根本上改变了现代图像合成。DALL-E 3（Betker 等，2023年）、Midjourney v7（Midjourney，2025）、Stable Diffusion 3（Esser 等，2024）和 NanoBanana（Google，2025）等模型如今为数百万个日常创意工作流提供支持。这些基于扩散的生成系统通过迭代去噪合成图像（Ho 等，2020），能够生成高度多样化、照片级逼真的图像。然而，在这项令人瞩目的能力之下，隐藏着一个令人担忧的系统性缺陷：模型遵循有依据的布局／身份线索，还是遵循文本顺

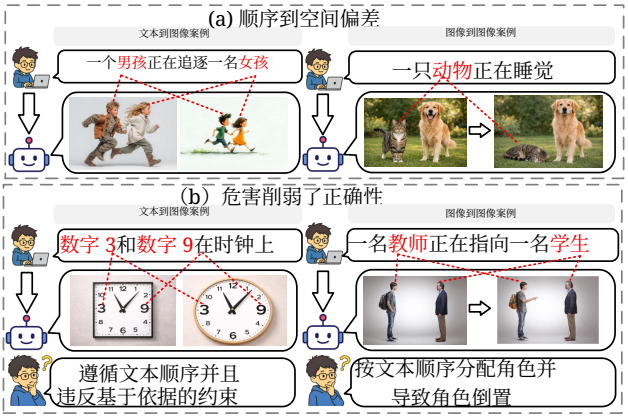


图*1.* **顺序到空间偏差（OTS）**会使模型将提及顺序视为空间布局和角色绑定的线索。(a) 在中性提示词下，首先提及的实体往往出现在左侧，而未充分指定的编辑通常会应用于默认的一侧。(b) 当顺序与基于实体的线索冲突时，模型反而遵循顺序，导致布局错误或角色颠倒。

序——将最先提及的实体放在左侧，并据此绑定角色／动作？

如图1(a)所示，现代生成模型表现出一种简单却持续存在的模式：它们常常将提示词中的提及顺序视为从左到右的布局规则。在文本到图像（T2I）任务中，即使提示词没有包含位置描述（例如“一个男孩正在追逐一个女孩”），生成的图像仍会呈现一致的空间顺序：首先提及的实体出现在左侧，第二个出现在右侧。在图像到图像（I2I）任务中，当提示词未充分指定（例如“一个动物正在睡觉”），且输入图像中包含两只动物时，也会出现类似现象：模型往往默认将所要求的属性或动作赋予左侧实体。我们将这种现象称为**顺序到空间偏差（OTS）**——一种系统性失效，生成模型将文本顺序视为空间或语义分配规则——把“A 和 B”映射为固定的左右布局，或依据提及顺序而非基于实体的视觉线索或现实世界约束来绑定属性和动作。

尽管近期研究在识别和缓解文本引导的图像生成模型中的人口统计偏见、性别刻板印象和文化失实呈现方面取得了显著进展（Bianchi 等人，2023年；Luccioni 等人，2023年；Struppek 等人，2023年；Friedrich 等人，2023年），但

*共同贡献 [1]中国人民大学，中国 [2]华中农业大学，中国 [3]华中科技大学，中国 [4]江南大学，中国 [5]南昌大学，中国。通讯作者：Wenxuan Wang < wangwenxuan@ruc.edu.cn>。

预印本。*2026*年*3*月*5*日。

OTS 引发的空间推理失误仍鲜少受到探索。现有的组合式生成基准测试（Huang 等人，2023年；2025年）主要检验实体是否出现，却往往忽略它们被放在何处，以及在缺乏空间线索的情况下为何会形成特定布局。近期的空间推理基准测试（例如 SPHERE（Zhang 等人，2025年）和 OmniSpatial（Jia 等人，2025年））系统地评估视觉语言模型在不同任务和难度级别下的空间感知与多步空间推理能力，但并未将顺序作为一种生成捷径单独加以考察——也就是说，在空间中性提示词下，提及顺序是否会被错误地映射为左右布局或角色/动作绑定。

这一差距至关重要，因为顺序到空间偏差（OTS）会导致生成结果系统性地出错。如图 1(b) 所示，当文本顺序凌驾于指称落地或语义约束之上时，模型就会生成错误的空间布局——例如，尽管时钟上的数字应按顺时针顺序排列，模型却将“3”放在“9”的左侧。除了布局错误之外，同样的偏差还会导致图生图中更严重的语义失败。模型常常表现出**动作错误归因**：给定一张输入图像和诸如“一位教师正在指着一名学生”这样的提示词时，模型会根据文本中的角色顺序而非视觉锚定来分配动作，导致角色颠倒，违背预期的社会关系和语义关系。

为了量化顺序到空间偏差（OTS），我们推出 OTS-BENCH，这是一个包含 4,300 个测试用例的基准测试，涵盖文本到图像和图像到图像生成。该基准测试由 138 个实体与 172 种动作/状态组合构建而成，可对主体和交互中由顺序引发的影响进行可控评估。我们从两个维度评估 OTS：（1）同质化，衡量模型在左右布局（T2I）或动作/属性分配（I2I）中遵循提示顺序的程度；（2）正确性，评估模型生成是否遵循现实世界约束（T2I），或是否保留输入图像中的指称落地（I2I），而非依赖基于顺序的捷径。

我们对九个最先进模型的评估表明，OTS 普遍存在。在文本到图像和图像到图像任务中，模型生成往往默认采用遵循顺序的布局或角色/动作绑定：面对中性或欠指定提示词时，文本到图像同质化率通常超过 70%；当提及顺序与接地约束相矛盾时，文本到图像正确性会从对齐条件下的 ∼90% 降至反向条件下的 ∼20%。我们还在网络规模的图文数据中观察到强烈的先提及者在左的先验，这表明 OTS 在很大程度上源于数据。此外，我们在扩散采样过程中进行时间干预，将 OTS 定位于早期的布局形成阶段，揭示了顺序线索何时主导生成。最后，我们表明，生成时调度和有针对性的微调都能有效缓解这种偏差，同时保持图像质量。

贡献。 (1)我们在现代文本到图像和图像到图像生成器中发现了顺序到空间偏差（*OTS*）：实体的提及顺序会意外地决定空间布局以及实体与角色/动作的绑定。(2) 我们提出 OTS-BENCH，这是一个受控基准测试，通过成对提示词隔离顺序效应，并从同质化 和正确性两个方面评估模型。(3) 我们对最先进模型开展大规模评估，结果表明 OTS 普遍存在，并会导致严重的指称落地失败。(4) 我们追溯发现，OTS 源于数据偏差和早期布局形成，并通过生成时调度和定向微调加以缓解，同时不降低图像质量。

2. 相关工作

文本到图像与图像到图像模型。 近期的文本到图像模型，如 Stable Diffusion 3 (Esser et al., 2024)、FLUX-dev (Black Forest Labs, 2024) 和 PixArt (Chen et al., 2023)，采用扩散变换器 (Peebles & Xie, 2023)，并基于 CLIP (Radford et al., 2021) 或 T5 编码器 (Raffel et al., 2020) 实现视觉-语言对齐。指令引导的图像到图像模型，如 Instruct-Pix2Pix (Brooks et al., 2023)、ControlNet (Zhang et al., 2023) 和 IP-Adapter (Ye et al., 2023)，则进一步支持对现有图像进行文本控制的编辑。文本到图像和图像到图像系统都会按从左到右的顺序处理提示词，但这种序列编码如何在生成过程中引入系统性的空间行为，长期以来一直未受到足够关注。

组合生成基准。 以往大多数基准测试在显式空间规范下评估组合正确性。T2I-CompBench（Huang 等人，2023年；2025年）使用基于检测器的评分来测试三维关系，而 GenEval（Ghosh 等人，2023年）以及 TIFA（Hu 等人，2023年）和 VPEval（Cho 等人，2023年）等视觉问答式指标则提供更广泛的对象级验证。SPHERE（Zhang 等人，2025年）和 OmniSpatial（Jia 等人，2025年）等空间基准测试进一步在结构化设置中诊断空间感知与推理。相比之下，我们研究一种不同的失效模式：在空间中性提示词下，模型可能依赖一种基于顺序的捷径，将实体提及顺序映射到左右布局或角色/动作绑定。

生成模型与视觉-语言模型中的偏见。以往关于生成模型的研究主要关注人口统计学和文化偏见以及公平性评估（Friedrich 等人，2024；Jha 等人，2024；Wan 等人，2024；Luo 等人，2024；Chinchure 等人，2024），而针对判别式视觉语言模型的研究则考察了位置效应，例如近因偏差和嵌入引发的空间偏差（Tian 等人，2025；Alam 等人，2026）。数据分析（如 Concept-

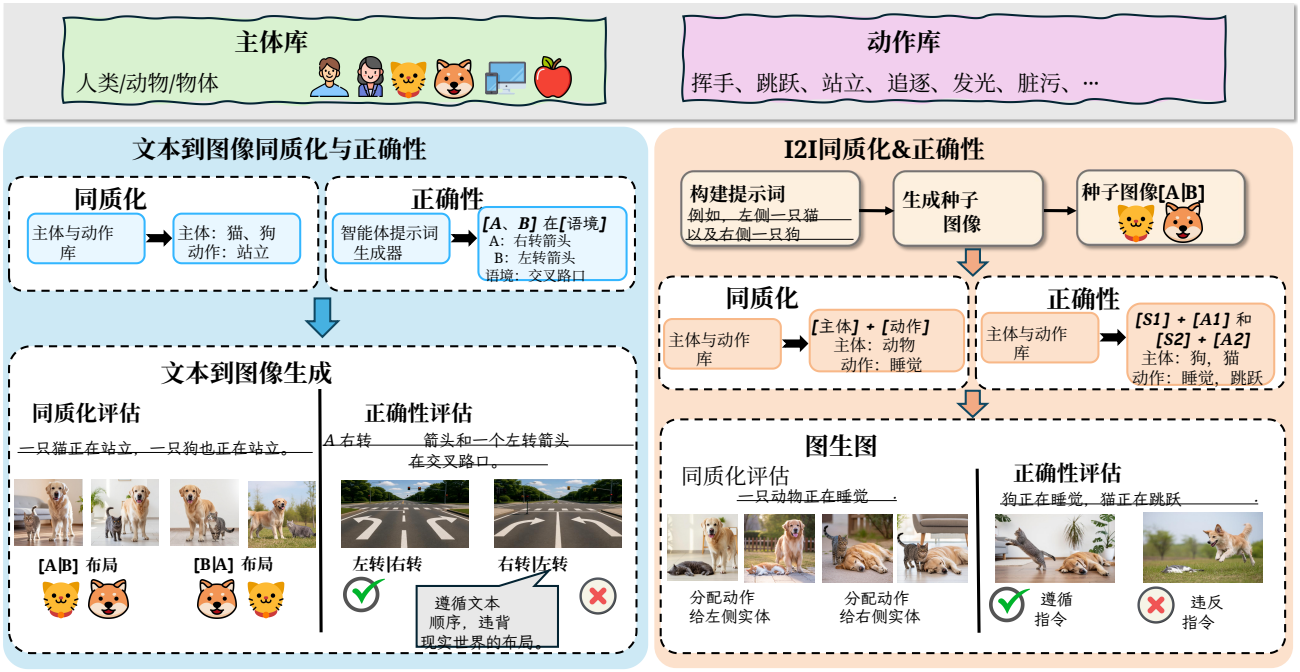


图 *2.* OTS-BENCH 概览。我们在四种设置下评估 OTS：文本到图像和图像到图像，每种设置均考察同质化和正确性。

Mix（Wu 等人，2024））进一步揭示了网络图像描述的局限性，包括 LAION-5B 中空间关系描述不充分的问题（Schuhmann 等人，2022）。然而，这些研究方向关注的是生成什么内容，或位置信号如何影响识别，而非实体被放置在何处，或哪个实体在生成过程中获得某种角色或动作。与此不同，我们研究文本到图像和图像到图像生成中的**顺序到空间偏差（OTS）**——一种在生成时将提及顺序映射为空间布局或角色分配的捷径——并将其与图像描述数据中的顺序不对称联系起来，表明其可能源于数据。

3. OTS-Bench

图 2 概述了 OTS-BENCH 的设计：主体/动作库以及用于衡量文本到图像和图像到图像任务中同质化与正确性的四种设置。

3.1. 基准组成

如表 1 所示，OTS-BENCH 包含 138 个主体和 172 种动作／状态类型。主体按照三级分类体系组织：人类（64 个，分属 8 种职业，且性别均衡）、动物（25 个，按体型划分为 5 个生物类别）和物体（49 个，分属 7 个功能类别）。动作／状态清单包括 35 种人类动作、55 种动物动作和 82 种物体状态。我们参考现有数据集和词典编制这些清单，并整合各类别，以确保一致性和覆盖率。实体、动作和指称落地参照的完整列表见附录 A。

表 *1.* OTS-Bench 分类体系概览。主体和动作/状态清单参考 COCO（Lin 等人，2014）、ImageNet（Deng 等人，2009）、OpenImages（Kuznetsova 等人，2020）、AVA（Gu 等人，2018）、VerbNet（Schuler，2005）、动物王国（Ng 等人，2022）和 Something-Something（Goyal 等人，2017）整理而成。

组成部分	数量	来源
主体		
人类	64	SOC 2018
动物	25	COCO、ImageNet
物体	49	COCO，OpenImages
主体总数	**138**	
动作与状态		
人类动作	35	AVA, VerbNet
动物动作	55	动物王国
物体状态	82	Something-Something
动作与状态总计	**172**	

3.2. 基准构建

OTS-BENCH 包含 **4,300** 个测试用例，涵盖两种模态（文本到图像/图像到图像）和两个评估维度（同质化/正确性），汇总于表2。按实体对类型的完整细分留待附录 B。

表2. OTS-BENCH 概述。

模态	同质化	正确性	总计
T2I	1,700	400	2,100
I2I	1,100	1,100	2,200
总计	2,800	1,500	**4,300**



图3. 用于生成受约束测试用例的多智能体流水线。Task Explorer 提出候选方案，Insight Synthesizer 验证现实世界约束，Diversity Architect 生成结构化数据。

3.2.1. 文本到图像基准

同质化评估。 我们测试模型是否会默认将先提及的实体置于左侧、后提及的实体置于右侧，即使提示词中不含任何空间线索。例如，对于“一只猫站立着，一只狗站立着”，低同质化模型应以相近的频率同时生成 [猫|狗] 和 [狗|猫] 两种布局，而顺序偏置模型则会持续生成 [猫|狗]。

正确性评估。 我们测试模型是否会遵循基于现实的左右约束，尤其是在这些约束与顺序所隐含的映射相冲突时。对于每个受约束对，我们评估两个匹配的提示变体：**对齐（Ali）**，其中提及顺序与现实世界惯例一致；以及**反向（Rev）**，其中提及顺序颠倒，但预期惯例保持不变。这一设计将对文本顺序的依赖作为两个变体之间唯一的差异。例如（图2），在交叉路口，正确的布局是*左转|右转*：**Ali**使用“先左转，再右转”，而**Rev**则颠倒顺序（“先右转，再左转”）。随后，我们检查生成的布局是否仍符合基于现实的惯例。

我们使用多智能体流水线（图3）构建了400个此类受约束对，该流水线提出候选并验证可验证的左右惯例，涵盖功能约束、时间／位置惯例以及常规安排。

3.2.2. 图像到图像基准

同质化评估。 我们检验当生成指令未充分指定时，模型是否会默认采用一致的动作分配。我们首先生成一张具有明确左右定位依据的种子图像（例如，“*左侧一只猫、右侧一只狗*”），然后应用一条未指定动作执行实体的指令（例如，“一只动物在睡觉”）。低同质化模型应以相近的频率将动作分配给任一实体，而有偏模型则会始终将其分配给某一侧（通常是左侧／第一个实体）。

正确性评估。 我们测试模型是否遵循提示词中的实体-动作绑定，而不会调换执行各个动作的实体。我们从一张主体明显分开的种子图像开始（例如，左侧有一只猫，右侧有一只狗）。随后，我们使用一个为两个主体分配不同动作的双实体提示词（例如：“狗在睡觉，猫在跳跃”）。只有当狗 表现出“睡觉”的状态，且猫表现出“跳跃”的状态时，输出才算正确。

为将 OTS 与一般的图像到图像生成难度区分开来，我们针对每个测试用例评估两种匹配的变体。在**对齐**变体中，提示词列出的实体顺序与种子图像中的左右身份顺序相同（先列左侧实体，再列右侧实体）；在**反向**变体中，我们调换提示词中的实体顺序，同时保持预期的实体-动作绑定不变。我们纳入（1）实体彼此独立行动的非交互动作，以及（2）指定谁对谁采取行动的定向交互。

4. 评估流程

我们在两种设置下评估顺序到空间偏差：文本到图像生成和图像到图像生成，并从两个评估维度进行考察：同质化和正确性。对于每种设置，我们根据提示词生成输出，核验实体是否存在以及左右清晰度，并依据相应标准标注结果。同质化用于衡量在中性或欠指定提示词下，模型是否倾向于采用一致的顺序隐含布局或分配方式；正确性则用于衡量输出是否符合现实世界约束（文本到图像），或是否遵循提示词预期的实体-动作绑定（图像到图像）。无效案例（例如缺失实体、布局含糊）不纳入评分。

4.1. 同质化评估

4.1.1. 文本到图像同质化

对于文本到图像同质化，我们测试模型是否会为语义中性的双实体提示词生成平衡的左右布局。每条提示词都包含两个主体（E_A，E_B），从我们的主体库（第 3.2 节）中采样。

对于每张生成图像 $\mathcal{I}$，我们会确认：(1) 两个实体E_A 和 E_B 都清晰可见；(2) 它们的左右排列明确无歧义（排除基于深度或垂直方向的布局）。然后，我们进行标注：

$$y = \begin{cases} \text{L} & E_A \text{ left of } E_B \\ \text{R} & E_A \text{ right of } E_B \\ \text{invalid} & \text{otherwise} \end{cases} \tag{1}$$

同质化得分用于衡量左右分布偏离均衡状态的程度。令 N_{L} 和 N_{R} 分别表示 E_A 位于左侧或右侧的有效输出数量。我们定义

$$\text{Hom}_{\text{T2I}} = 100 \cdot \left| \frac{N_{\text{L}} - N_{\text{R}}}{N_{\text{L}} + N_{\text{R}}} \right|. \tag{2}$$

其中，数值越低，表示布局越不受顺序锁定；$\text{Hom}_{\text{T2I}} = 0$ 对应完美均衡的50/50划分。

4.1.2. 图生图同质化

对于图像到图像同质化，我们检验模型是否会在双实体种子图像中，始终将欠指定动作/状态 α 分配到某一侧。我们仅保留两个实体均可辨认且恰有一个实体表现出α的输出。然后标记为：

$$y = \begin{cases} \text{L} & \text{left entity exhibits } \alpha \\ \text{R} & \text{right entity exhibits } \alpha \\ \text{invalid} & \text{otherwise} \end{cases} \tag{3}$$

令 N_{L} 和 N_{R} 分别表示标记为 L 和 R 的有效输出数量。我们将同质化得分定义为相对于均衡的50/50分配的偏差：

$$\text{Hom}_{\text{I2I}} = 100 \cdot \left| \frac{N_{\text{L}} - N_{\text{R}}}{N_{\text{L}} + N_{\text{R}}} \right|. \tag{4}$$

其中，较低的值表示编辑行为较少受侧向锁定影响，而 $\text{Hom}_{\text{I2I}} = 0$对应于完全均衡的划分。

4.2. 正确性评估

4.2.1. 文生图正确性

为了评估文本到图像正确性，我们特意构建了这样的提示词：其提及顺序与基于现实世界的左右惯例相冲突，以揭示顺序到空间偏差。对于每个测试用例，我们考虑一个实体对（A，B）以及一种可验证的语境，该语境规定了哪个实体应出现在左侧，哪个实体应出现在右侧。我们将参考布局标签 $\ell^* \in${（A-左侧 B-右侧），（B-左侧 A-右侧）} 存储起来。例如，对于“标准钟面上的数字 *3* 和数字 *9*”，ℓ^* =（9 在左侧，3 在右侧）。

我们评估两种匹配的提示变体：**对齐 (Ali)**，其提示顺序与 ℓ^* 相符；以及 **反向 (Rev)**，其中实体顺序互换，而 ℓ^* 保持不变。给定一张生成图像 $\mathcal{I}$，我们验证：(1) 两个实体均存在且可辨认；(2) 描绘了所要求的语境；(3) 左右布局明确无歧义。随后我们指定：

$$y = \begin{cases} \text{correct} & \mathcal{I} \text{ matches } \ell^* \\ \text{wrong} & \mathcal{I} \text{ contradicts } \ell^* \\ \text{invalid} & \text{otherwise.} \end{cases} \tag{5}$$

我们将正确性报告为有效样本中的准确率：

$$\text{Acc}_{\text{T2I}} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{wrong}}}, \tag{6}$$

其中，N_{correct} 和 N_{wrong} 分别表示标记为正确 和错误的有效输出数量。

4.2.2. 图像到图像正确性

对于图像到图像正确性，我们测试模型是否遵循提示词中指定的双实体种子图像中的动作/角色分配。对于每张输出图像 $\mathcal{I}$，我们验证：(1) 两个实体是否仍可辨认；(2) 每项指定动作/角色的证据是否清晰。然后，我们进行如下标注：

$$y = \begin{cases} \text{correct} & \text{actions/roles match the prompt} \\ \text{wrong} & \text{actions/roles are swapped} \\ \text{invalid} & \text{otherwise.} \end{cases} \tag{7}$$

我们在有效样本中报告正确性（准确率）：例：

$$\text{Acc}_{\text{I2I}} = \frac{N_{\text{correct}}}{N_{\text{correct}} + N_{\text{wrong}}}, \tag{8}$$

其中，N_{correct} 和 N_{wrong} 分别表示被标记为正确 和错误的有效输出数量。

5. 实验与发现

5.1. 实验设置

5.1.1. 受评估模型

我们评估了 9 个视觉语言模型在文本到图像和图像到图像任务中的表现：SDXL (Podell et al., 2023年)、SD3.5 (Esser et al., 2024)、FLUX-dev (Black Forest Labs, 2024)、Qwen- Image (Wu 等人, 2025)、DALL-E 3 (Betker et al., 2023)、Midjourney v7 (Midjourney, 2025)、Kling-v2 (快手科技, 2025)、GPT-Image (OpenAI, 2025) 和 NanoBanana (Google, 2025)。

表 *3.* 视觉语言裁判与人工标注之间的一致性结果。**同质化**：偏差分数定义为 Hom $= |2p-100|$，其中 p 是被分配到标签1的有效样本百分比（有效样本不包括标签0）。**正确性**：在*Ali*和*Rev*条件下，有效样本的准确率。**Cohen's** κ：根据完整的 2,400 个样本的三分类混淆矩阵（标签0/1/2）计算。

模型	同质化		文本到图像正确性（%）		I2I正确性（%）		Cohen's $\kappa\uparrow$
	T2I	I2I	Ali	Rev	Ali	Rev	
llama-3.2-11b-vision	97.1	73.3	97.0	2.4	89.7	87.6	0.62
Qwen3-VL-8B-Instruct	94.2	63.9	85.3	12.5	90.4	83.7	**0.81**
InternVL3-8B	90.2	59.8	84.8	24.3	94.9	92.9	0.72
llava-critic-r1-7b	54.9	60.2	91.4	23.8	96.9	95.2	0.47
人类	**90.3**	**68.7**	**82.0**	**9.2**	**89.8**	**82.5**	–

表*4.* OTS-Bench主要结果。**同质化** 衡量左右结果偏离均衡50/50划分的程度：Hom $= |2p-100|$ p 表示被分配到左侧的有效样本百分比（越低越好）。**正确性** 报告提示顺序与指称落地约束一致（*Ali*）或相悖（*Rev*）时的准确率。我们将性能下降 $\Delta =$ Ali $-$Rev作为顺序敏感性的诊断指标。

模型	同质化 ↓		文本到图像正确性 ↑		Δ ↓	I2I正确性 ↑		Δ ↓
	T2I	I2I	Ali	Rev	T2I	Ali	Rev	I2I
SDXL	52.6	55.8	83.3	23.6	59.7	73.5	62.2	11.3
SD3.5	84.2	38.4	84.8	21.1	63.7	79.5	71.8	7.7
FLUX-dev	88.8	42.6	79.8	24.7	55.1	84.3	76.8	7.5
Qwen-Image	91.6	83.4	81.8	28.2	53.6	88.1	78.8	9.3
DALL-E 3	70.4	–	87.7	18.2	69.5	–	–	–
Midjourney v7	86.8	63.2	90.2	21.7	68.5	87.3	82.1	5.2
Kling-v2	77.2	82.6	93.6	14.1	79.5	81.3	75.6	5.7
GPT-Image	86.4	35.4	79.5	15.3	64.2	94.4	79.6	14.8
Nanobanana 1	81.0	51.8	93.2	17.6	75.6	96.3	91.9	4.4

5.1.2. 检测模型选择

为选择自动化视觉语言评判器，我们从所有 9 个受评估模型生成的结果中，按评估指标**均匀抽取 400 个输出**，构建了一个由人类标注的验证集（共 6 项评估指标、2,400 个样本），并由专家标注员依照第 4 节进行标注。随后，我们让候选评判器对同一批样本进行评判，并使用覆盖全部 2,400 个样本的**3 类**混淆矩阵（标签为 0/1/2）计算 Cohen's κ，以评估其与人类标签的一致性。如表 3 所示，Qwen3-VL-8B-Instruct 与人类的一致性最高（$\kappa = 0.81$），因此我们选用它作为所有大规模评估的唯一评判器。详细的提示词模板和更细致的分析见附录E。

5.2. 主要结果

表4总结了文本到图像与图像到图像两种设置下同质化和正确性方面的表现。

同质化。 模型在两种模态中都表现出很强的顺序锁定。在文本到图像中，同质化程度始终较高，从 52.6（SDXL）到 91.6（Qwen-Image）不等，这表明在空间中性提示下，生成的布局绝大多数都遵循提及顺序。在图像到图像中，支持图像到图像的模型之间，同质化程度的范围更广，从 35.4（GPT-Image）到 83.4（Qwen-Image）不等，这表明——明确的视觉锚定在一定程度上减轻了顺序锁定，但许多模型仍存在显著偏差。

正确性（对齐 vs. 反向）。在正确性方面，我们报告了匹配的**对齐（Ali）**对照条件下的结果，其中提示顺序与指称落地一致（文本到图像：现实世界的左右惯例；图像到图像：种子图像身份顺序），以及对比性的**反向（Rev）**设置，其中提示顺序与之相矛盾；性能下降按 $\Delta =$ Ali $-$Rev计算。在对齐设置中，模型取得了较高准确率（文本到图像：79.5–93.6%；图像到图像：73.5–96.3%）。在反向条件下，文本到图像的正确性急剧下降至14.1–28.2%，性能下降幅度较大（Δ：53.6–79.5），表明当提及顺序与基于依据的约束发生冲突时，模型往往遵循与顺序一致的布局。在图像到图像中，该效应较轻但仍持续存在：反向正确性仍为62.2–91.9%（Δ：4.4–14.8），表明视觉锚定部分缓解了这一偏差，但由顺序驱动的角色／动作误归因仍然常见。

5.3. 网络规模语料库分析

为探究 OTS 潜在的数据来源，我们测量网页图像描述–图像对表现出<式样 id='1'>顺序到空间</式样>一致性的频率：对于描述两个实体的标题，较早提及的实体是否出现在配对图像的<式样 id='3'>左侧</式样>，而较晚提及的实体是否出现在<式样 id='5'>右侧</式样>?

表 5. 网络规模语料库中的顺序到空间对齐。**样本** 是筛选后有效图文对的数量。**OTS对齐** 是标题顺序（第一个实体、第二个实体）与图像布局（左侧实体、右侧实体）相匹配的比例。

数据集	样本	OTS对齐（%）
LAION-2B-en-aesthetic	23,020	89.2
DataComp-Large	10,150	87.2

我们使用 LAION-2B-en-aesthetic 和 DataComp-Large 作为公开代理数据集，筛选标题提及两个不同实体，且图像以清晰的左右布局呈现这两个实体的图文对（附录 F）。在这一筛选后的集合上，我们报告<式样 id='1'>顺序到空间匹配率</式样>，其定义为标题顺序（第一个实体、第二个实体）与图像布局（左侧实体、右侧实体）相匹配的案例比例。

如表 5 所示，在两个语料库中，这一匹配率都很高（89.2% 和 87.2%），这表明，顺序到空间的一致性是大规模网络数据中普遍存在的规律。在训练过程中，模型会反复接触到文本提及顺序与水平空间排列之间稳定的对应关系，这种高度一致性便说明了这一点。

因此，我们在基准测试中观察到的系统性顺序到空间偏差（OTS）不太可能是随机伪象。相反，它可以被解释为一种涌现的归纳偏置：当缺少明确的空间约束时，模型似乎会将这种习得的统计规律作为默认先验，实际上是用文本顺序来推断左右放置。

5.4. 通过微调缓解 OTS 偏差

前述实验表明，现代生成器普遍存在顺序到空间 偏差，表现为强烈倾向于将提示顺序映射为空间布局或角色绑-定。为缓解这一偏差，我们在两个具有代表性的开放权重骨- 干模型 **FLUX-dev**和 **Qwen-Image**上进行基于 LoRA 的监- 督微调，使用来自 OTS 对齐图文数据的水平翻转图像对：对于每个双实体样本（例如“一个男子和一个女子”，布局为男子|女子），我们在相同提示词下添加一个布局翻转的对应样本女子|男子 。这一简单的数据增强削弱了提及顺- 序与左右分配之间的相关性，从而减少 OTS 偏差。

训练数据。 我们使用精心整理的反 OTS 数据集，通过监督微调对模型进行微调。该数据集旨在直接打破虚假的顺序→空格键 捷径：对于每个具有清晰水平（左右）布局的双实体实例，我们通过添加水平翻转的对应样本来构建配对监督，同时保持图注未-改变，从而使相同文本 在两种镜像布局下都被观察到。我们仅保留两个实体均可识别且左右排列无歧义的高置信度对，以确保高信噪比的去偏信号。最终的监督微调集包含从公开数据来源整理的 2k 图文对。这种针对性构建之所以有效，是因为 OTS 在很大程度上是由训练数据相关性引起的单一捷径，而基于翻转的配对提供了直接的反向信号，无需大规模概念学习。更多实现细节见附录 D。

结果。 表 6 显示，翻转增强的 LoRA 监督微调降低了 FLUX-dev 和 Qwen-Image 在文本到图像和图像到图像任务中的同质化，表明模型对提示顺序的依赖减弱，在左右分配和角色绑定方面尤为如此。同时，整体正确性保持稳定，并且在对齐设置下有所提升，这表明当提示顺序与预期的基-础信息一致时，模型的遵循能力更强。重要的是，图像质量基本得以保留：基于评分标准的人类评分（整体质量/对齐/保真度）保持相近，ImageReward（Xu 等，2023年）也未表现出系统性的性能下降。整体而言，这些结果表明，简单的基于翻转的训练信号能够缓解顺序跟随行为，同时不牺牲视觉质量。

质量控制。 为验证去偏不会损害视觉质量，我们在留出集上评估自动和人工质量指标。具体而言，我们使用相同的提示词，并收集各模型变体生成的结果，为每个模型抽取 400 组匹配输出。我们报告这些样本的 ImageReward 分数，并请专家标注员依据 1–7 分制，对每张图像的整体质量、提示词对齐和保真度 进行评分，评分标准沿用 ImageReward（Xu 等，2023）。完整的人类评分结果见附录 G。

5.5. 顺序到空间偏差的时间定位

时间视角。 尽管数据层面的分析表明，OTS 源于统计因素，但并未说明顺序线索在生成过程中如何影响空间布局。为更好地理解这一现象，我们考察扩散模型的生成动态。我们的分析基于扩散生成由粗到细的固有特性。已有充分证据表明，早期去噪阶段主要决定全局结构和空间布局，而后期阶段则主要细化局部外观和纹理（Yan et al., 2025; Yue et al., 2024）。由于 OTS 体现在全局布局层面，其影响预计会随生成过程而变化，因此时间轴是一个自然且有参考价值的分析维度。

表 *6.* 对 FLUX-dev 和 Qwen-Image（含 LoRA 变体）的同质化、正确性和图像质量进行综合比较。同质化指标越低越好（↓）；正确性和图像质量指标越高越好（↑）。图像质量以 ImageRe- ward（均值 ± 标准差）报告。

模型	同质化 ↓		文本到图像正确性 ↑		图像到图像正确性 ↑		图像质量 ↑
	T2I	I2I	Ali	Rev	Ali	Rev	ImageReward
FLUX-dev	88.8	42.6	79.8	24.7	84.3	76.8	$0.186_{\pm 0.672}$
FLUX-dev（SFT）	47.4	25.0	83.2	49.3	87.0	79.5	$0.217_{\pm 0.687}$
Qwen-Image	91.6	83.4	81.8	28.2	88.1	78.8	$0.204_{\pm 0.732}$
Qwen-Image（SFT）	50.8	32.8	84.7	57.7	90.5	80.1	$0.223_{\pm 0.717}$

	切换步数					
设置	base	1	2	3	4	5
(a) 切换方向：成功率 (%)						
A2B	100.0	25.0	14.8	11.1	3.2	2.2
B2A	100.0	23.0	17.8	5.5	2.4	1.2
(b) 提示词转换：同质化 (%)						
N2A	89.9	28.8	21.8	27.2	34.1	39.1
N2B	85.0	31.1	15.7	20.0	27.1	41.6

表*7.* 顺序线索的时间效应。(a) A2B/B2A：切换方向。(b) N2A/N2B：从中性提示转换为具有不同实体顺序的丰富提示词。数值以 % 表示。

顺序线索的时间敏感性。 为考察顺序线索在整个生成过程中如何影响空间布局，我们通过在扩散采样期间修改文本条件，开展时序干预实验。对于涉及两个不同实体的提示词，我们在扩散采样期间进行时间提示切换干预。每张图像都沿着一条固定的采样轨迹生成，并在选定的去噪步骤 t 处，通过颠倒提示词中两个实体的提及顺序来修改文本条件，同时保持其他所有采样因素不变。随后，我们考察最终的相对布局遵循的是原始顺序还是调换后的顺序，并将成功率定义为最终布局与调换后提示词相符的样本比例。

我们选取了基线设置下表现出顺序到空间偏差的500个样本，涵盖多种物体类别和场景背景。FLUX-dev上的结果（表7(a)）揭示了明显的时间依赖性。在较早的时间步切换顺序时，一定比例的样本能够成功遵循切换后的顺序，这表明此阶段的空间布局仍可在一定程度上重新配置。相比之下，在较晚阶段，提及顺序的影响变得微乎其微，切换顺序不再改变既定结构。这些结果表明，提及顺序对空间布局的影响主要集中在早期的布局敏感窗口内，而非在整个采样过程中均匀累积。

延迟顺序条件化。上述时间分析表明，提及顺序对空间布局的影响仅限于生成早期的布局敏感阶段。这启发了一种简单的缓解策略：将对顺序敏感的文本条件延迟到全局空间结构确立之后再引入。在基于扩散的生成中，空间布局形成与实体身份绑定在时间上彼此解耦：通过使用顺序中性提示词初始化采样过程，模型可以在不受文本顺序偏置的情况下确定全局空间结构。在早期阶段引入既不偏向任何实体的对称空间布局，中性提示词便能确立一种结构，使后续去噪仅需细化配置，从而令基于顺序的线索无法诱发布局层级重组。

具体而言，我们在扩散采样过程中采用从中性提示切换到丰富提示的方案。对于每个双实体提示词，我们通过匿名化实体身份并删除提及顺序信息，同时保留场景描述和关系结构，从原始提示词中生成中性提示 。原始提示词则作为对应的丰富提示。例如，像“公园里的一个男子和一名女子”这样的丰富提示，会转换为“两个人在公园里”。生成从中性提示开始，并在预先设定的去噪步骤 t 切换到丰富提示，其他所有采样设置均保持不变。

表7(b)报告了延迟顺序条件化在FLUX-dev上的缓解效果。基线表现出严重的布局同质化，通过延迟引入顺序敏感信息，该现象得到大幅缓解；在我们的实验设置下，第2步观察到最强效果。若在该时间点之前或之后切换，缓解效果都会减弱。总体而言，这些结果表明，延迟顺序条件化提供了一种简单而有效的缓解策略。

6. 结论

我们的结果表明，顺序到空格键偏差只是冰山一角，反映了视觉语言系统中更广泛的序列诱导先验：实体顺序会成为一种

视觉结构和语义绑定的默认规则。除了双实体生成之外，一个重要的研究方向是探讨这类捷径如何扩展到多实体构图（3+ 个主体）、需要随时间保持角色和空间关系一致的视频生成，以及图像描述/视频描述；在这些任务中，模型可能会表现出相同的顺序偏好，并强化训练数据中的偏差。

影响声明

本研究探讨了文本引导的图像生成中的一种系统性偏差——顺序到空间偏差（*OTS*），并推出 OTS-BENCH，用于衡量提及顺序如何可能错误地决定空间布局或角色/动作绑定。其主要积极影响在于提升透明度和评估能力：我们的基准测试和分析有助于研究人员和从业者诊断指称落地失败，并制定缓解措施，从而在空间关系和实体角色绑定至关重要时，让生成系统更加可靠。

潜在风险包括助长利用顺序效应生成误导性图像的对抗性提示设计，以及强化对从网络规模数据相关性中习得的捷径的关注。为降低这些风险，我们的贡献侧重于测量和缓解，并提出简单的干预措施，以削弱虚假的顺序与布局相关性，同时不引入额外的控制先验。当生成图像可能影响敏感决策或解读时，我们建议进行审慎审查并实施人工监督。

参考文献

Alam, N., Murali, L. K., Bharadwaj, S., Liu, P., Chung, T., Sharma, D., Kiran, K., Tam, W., Vegesna, B. K. S., 等。《视觉语言模型的空间盲点》。*arXiv* 预印本 *arXiv:2601.09954*，2026年。Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., 等。《通过更优描述改进图像生成》。计算机科学。*https://cdn. openai. com/papers/dall-e-3. pdf*，2(3): 8，2023年。

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., 和 Caliskan, A. 易于获取的文本到图像生成大规模加剧人口统计学刻板印象。载于*2023*年 *ACM* 公平性、问责制与透明度会议论文集，第1493–1504页，2023年。

Black Forest Labs。Flux。https://github.com/black-forest-labs/flux，2024年。Brooks, T., Holynski, A., 和 Efros, A. A. InstructPix2Pix：学习遵循图像编辑指令。载于论文集 *IEEE/CVF*计算机视觉与模式识别会议论文集，第18392–18402页，2023年。

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. Pixart-α：用于照片级真实感文本到图像合成的扩散 Transformer 快速训练。*arXiv* 预印本 *arXiv:2310.00426*，2023年。

Chinchure, A., Shukla, P., Bhatt, G., Salij, K., Hosanagar, K., Sigal, L. 和 Turk, M. Tibet：识别并评估文生图生成模型中的偏见。载于欧洲计算机视觉会议，第 429–446 页。Springer，2024年。

Cho, J., Zala, A. 和 Bansal, M. 用于逐步文本到图像生成与评估的可视化编程。神经信息处理系统进展，36：6048–6069，2023年。

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. 和 Fei-Fei, L. ImageNet：大型分层图像数据库。载于*2009*年 *IEEE* 计算机视觉与模式识别会议，第 248–255 页。IEEE，2009年。

Esser, P., Kulal, S., Blattmann, A., Entezari, R., M¨ uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., 和 Rombach, R. 扩展整流流 Transformer，实现高分辨率图像合成。载于第四十一届国际机器学习会议，2024年。

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., 和 Kersting, K. 公平扩散：指导文本到图像生成模型实现公平性。*arXiv* 预印本 *arXiv:2302.10893*，2023年。Friedrich, F., H¨ ammerl, K., Schramowski, P., Brack, M., Libovicky, J., Kersting, K., 和 Fraser, A. 多语言文本到图像生成会加剧性别刻板印象，提示工程或许也无济于事。*arXiv* 预印本 *arXiv:2401.16092*，2024年。Ghosh, D., Hajishirzi, H., 和 Schmidt, L. GenEval：用于评估文生图对齐的以对象为中心的框架。神经信息处理系统进展，36:52132–52152，2023年。Google。Gemini 2.5 Flash Image。https: //aistudio.google.com/models/gemini-2-5-flash-image，2025年。图像生成与编辑模型（昵称为“Nano Banana”），访问于2025-12-17。

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzyn- ska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M. 等。用于学习和评估的“something something”视频数据库

视觉常识。载于*IEEE*计算机视觉国际会议论文集，第5842–5850页，2017年。

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., 等。AVA：一个包含时空定位原子视觉动作的视频数据集。载于*IEEE*计算机视觉与模式识别会议论文集，第6047–6056页，2018年。

Ho, J., Jain, A., 和 Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020。

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., 和 Smith, N. A. TIFA：通过问答实现准确且可解释的文生图忠实度评估。载于*IEEE/CVF*国际计算机视觉会议论文集，第20406–20417页，2023年。

Huang, K., Sun, K., Xie, E., Li, Z. 和 Liu, X. T2i-compbench：面向开放世界组合式文本到图像生成的综合基准。神经信息处理系统进展，36:78723–78747，2023年。

Huang, K., Duan, C., Sun, K., Xie, E., Li, Z. 和 Liu, X. T2i-compbench++：用于组合式文本到图像生成的增强型综合基准。*IEEE*模式分析与机器智能汇刊，2025年。

Jha, A., Prabhakaran, V., Denton, R., Laszlo, S., Dave, S., Qadri, R., Reddy, C. 和 Dev, S. Visage：对文本到图像生成中视觉刻板印象的全球范围分析。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第12333–12347页，2024年。

Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H. 和 Yi, L. OmniSpatial：面向视觉语言模型的综合空间推理基准测试。*arXiv* 预印本 *arXiv:2506.03135*，2025年。

快手科技。Klingv2.0。https://klingai. com，2025年。文生视频生成模型，访问于2025-12-17。

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A. 等。《Open Images 数据集 v4：大规模统一图像分类、目标检测和视觉关系检测》。国际计算机视觉杂志，128(7):1956–1981，2020。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P. 和 Zitnick, C. L.《Microsoft COCO：语境中的常见物体》。载于欧洲计算机视觉会议，第740–755页。Springer，2014。

Luccioni, S., Akiki, C., Mitchell, M. 和 Jernite, Y. 稳定偏见：评估扩散模型中的社会表征。神经信息处理系统进展，36:56338–56351，2023年。

Luo, H., Deng, Z., Chen, R. 和 Liu, Z. FaintBench：文本到图像模型偏见评估的全面且精准的基准测试。*arXiv*预印本 *arXiv:2405.17814*，2024年。

Midjourney。Midjourney v7。https://www. midjourney.com，2025年。文本到图像模型，访问日期：2025-12-17。

Ng, X. L.、Ong, K. E.、Zheng, Q.、Ni, Y.、Yeo, S. Y. 和 Liu, J. 动物王国：用于动物行为理解的大型多样化数据集。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第19023–19034页，2022年。

OpenAI。GPT图像模型：gpt-image-1与多模态图像生成。https://platform.openai. c om/docs/models/gpt-image-1，2025年。官方模型文档。

Peebles, W. 和 Xie, S. 基于变换器的可扩展扩散模型。载于*IEEE/CVF*计算机视觉国际会议论文集，第4195–4205页，2023年。

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., M¨uller, J., Penna, J., 和 Rombach, R. SDXL：改进用于高分辨率图像合成的潜在扩散模型。*arXiv*预印本 *arXiv:2307.01952*，2023年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., 和 Sutskever, I. 从自然语言监督学习可迁移视觉模型。载于第*38*届国际机器学习会议论文集，第8748–8763页，2021年。URL https://proceedings.mlr. press/v139/radford21a.html。

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. 和 Liu, P. J. 探索使用统一文本到文本Transformer进行迁移学习的极限。机器学习研究期刊，21 (140):1–67，2020。

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. 等。Laion-5b：用于训练新一代图像-文本模型的开放式大规模数据集。神经信息处理系统进展，35:25278–25294，2022。

Schuler, K. K. *VerbNet*：广泛覆盖的综合性动词词库。宾夕法尼亚大学博士论文，2005年。

Struppek, L., Hintersdorf, D., Friedrich, F., Schramowski, P., Kersting, K. 等。通过文本到图像合成中的同形异义字利用文化偏见。人工智能研究期刊，78：1017–1068，2023年。Tian, X., Zou, S., Yang, Z. 和 Zhang, J. 识别并缓解多图视觉语言模型的位置偏差。载于计算机视觉与模式识别会议论文集，第 10599–10609 页，2025年。Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R. 和 Chang, K.-W. 文本到图像生成中的偏差综述：定义、评估与缓解。*arXiv* 预印本 *arXiv:2404.01030*，2024年。Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y. 等。Qwen-Image 技术报告。*arXiv* 预印本 *arXiv:2508.02324*，2025年。Wu, X., Yu, D., Huang, Y., Russakovsky, O. 和 Arora, S. Conceptmix：具有可控难度的组合式图像生成长凳。神经信息处理系统进展，37：86004–86047，2024年。Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J. 和 Dong, Y. ImageReward：为文本到图像生成学习并评估人类偏好。神经信息处理系统进展，36：15903–15935，2023年。Yan, S., Li, M., Xinliang, B., Yang, J., Zhang, Y., Xiong, G., Lan, Y., Zhang, T., Zhai, W. 和 Zha, Z.-J. 超越随机性：理解扩散中的噪声顺序。*arXiv* 预印本 *arXiv:2511.07756*，2025年。Ye, H., Zhang, J., Liu, S., Han, X. 和 Yang, W. IP-Adapter：适用于文本到图像扩散模型的文本兼容图像提示适配器。*arXiv* 预印本 *arXiv:2308.06721*，2023年。Yue, Z., Wang, J., Sun, Q., Ji, L., Chang, E. I., Zhang, H. 等。探索用于无监督表征学习的扩散时间步。*arXiv* 预印本 *arXiv:2401.11430*，2024年。

Zhang, L., Rao, A. 和 Agrawala, M. 为文本到图像扩散模型添加条件控制。载于*IEEE/CVF* 国际计算机视觉会议论文集，第 3836–3847 页，2023年。

Zhang, W., Ng, W. E., Ma, L., Wang, Y., Zhao, J., Koe- necke, A., Li, B. 和 Wanglu, W. Sphere：通过分层评估揭示视觉语言模型中的空间盲点。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第 11591–11609 页，2025 年。

表*8*。 OTS-Bench 实体分类体系，含类别细分和代表性示例。

Type	类别	数量	示例	来源
人类	医疗保健	8	医生、护士、牙医、药剂师	SOC 2018
	教育	8	教师、教授、图书管理员、辅导员	SOC 2018
	工程	8	土木工程师、建筑师、电工、机械工程师	SOC 2018
	餐饮服务	8	chef, 服务员, 咖啡师, 调酒师	SOC 2018
	管理	8	经理、财务经理、施工经理, 销售经理	SOC 2018
	艺术与媒体	8	设计师、摄影师、音乐家、艺术家	SOC 2018
	保护服务	8	警察、消防员、保安、护理人员	SOC 2018
	个人护理	8	美发师, 健身教练、托儿工作者、邮政工作人员	SOC 2018
	小计	**64**	*32*种职业 × *2*性别	
动物	小型宠物	5	猫、兔子、仓鼠、豚鼠、刺猬	COCO、ImageNet
	中型哺乳动物	5	狗、绵羊、猪、山羊、狐狸	COCO，ImageNet
	大型哺乳动物	5	马、牛、大象、长颈鹿、熊	COCO，ImageNet
	鸟类	5	鹦鹉、猫头鹰、鹰、企鹅、鸡	COCO、ImageNet
	爬行动物和两栖动物	5	龟、蜥蜴、蛇、青蛙、鳄鱼	ImageNet
	小计	**25**	*5* 个生物类别	
物体	家具	7	椅子、沙发、床、桌子、书桌、长凳、柜子	COCO, OpenImages
	厨具	7	杯子、瓶子、碗、盘子、叉子、刀、勺子	COCO, OpenImages
	电子产品	7	笔记本电脑、键盘、鼠标、手机、电视、遥控器, 相机	COCO, OpenImages
	体育器材	7	球、网球拍、滑板、冲浪板、滑雪板、飞盘, 棒球棒	COCO, OpenImages
	交通工具	7	自行车、摩托车、汽车、公共汽车、卡车、船、飞机	COCO, OpenImages
	乐器	6	吉他、钢琴、小提琴、鼓、萨克斯管、小号	OpenImages
	日常物品	7	书、雨伞、背包、手提包、行李箱、时钟, vase	COCO, OpenImages
	小计	**49**	*7* 个功能类别	
实体总数		**138**		

A. OTS-Bench分类体系

本附录详细说明 OTS-Bench 分类体系。我们的基准测试旨在通过基于成熟视觉数据集精心构建的实体-动作组合，系统评估视觉语言模型中的顺序到空间偏差。

A.1. 实体分类体系

我们构建了一个涵盖人类、动物和物体的三级实体分类体系，确保语义领域的多样性，同时保留数据集溯源，以实现可复现性。表8展示了完整的实体分类体系，包括类别细分和数据集来源。

A.2. 动作库

我们的动作库涵盖了不同程度的视觉复杂度，从简单的静态姿势到动态动作和状态转换。除了基础动作，我们还纳入了交互组合（例如握手、拍照）以及针对不同身体类型的变化，以便全面评估空间推理。对于人类，我们按身体部位

的参与情况来组织动作；对于动物，涵盖从基本姿势到特定物种行为的内容；对于物体，则涵盖功能、视觉和形态上的状态变化。表 9 展示了完整的动作和状态定义。

A.3. 评估设计

为了系统地探究空间推理，我们使用对比鲜明的动作对（例如，挥手 ↔ 敬礼、坐着 ↔ 站立）以及具有相似视觉特征的实体对（例如，猫 ↔ 兔子、笔记本电脑 ↔ 平板电脑）构建测试用例。每个测试用例都涉及生成具有特定文本顺序的实体和动作的提示词，然后评估输出是否在空间排列中正确反映了这一顺序。我们还生成反向提示词以验证一致性，从而在文本到图像生成和图像到图像编辑模态中产生了超过2,000个测试用例。

代表性示例。人类主体：医生（白大褂、听诊器），警察（制服、徽章），摄影师（相机背带）。**动物：**猫（坐着，耳朵竖起），大象（站立，象鼻放松），鹦鹉（栖息，翅膀收拢）。**物体：**笔记本电脑（通电/断电，屏幕开/关），杯子（空/满，干净/脏污），汽车（移动/停放，完好/撞毁）。

示例测试用例。人类（T2I）：“一名医生正在挥手，一名护士正在敬礼，并排站立”——验证空间位置是否与文本顺序一致。**动物（I2I）：**给定猫和狗坐着的图像，将其编辑为“猫坐着，狗正在奔跑”——检查动作分配。**物体（T2I）：**“一台笔记本电脑已开机，一台平板电脑已关机，并排放置”——验证状态对应关系。

A.4. 数据集溯源

为确保可复现性并与既有计算机视觉基准保持一致，所有实体均溯源至广泛使用的数据集。这有助于开展独立的实体识别验证，并将评估问题追溯至具体的数据集类别。

数据集覆盖范围。COCO 2017：41 个实体（9 种动物，32 种物体）。**ImageNet：**25 个动物同义词集。**OpenImages V6：**49 个物体类别。**SOC2018：**32 个职业代码。**动作来源：**AVA（人类动作）、AnimalKingdom（动物行为）、Something-Something（物体可供性）、VerbNet（动作语义）。

A.5. 文本到图像危害生成

B. 按配对类型划分的基准构成

表 10 报告了 OTS-BENCH 按模态、评估维度和实体对类型划分的完整构成。我们纳入同类型配对（人类–人类、动物–动物、物体–物体），以及适用情况下的跨类型配对，以确保覆盖各类语义类别，同时保证布局在视觉上可验证。

C. 智能体提示工程细节

为构建用于正确性评估的受约束实体对，我们采用了由三个专业角色组成的多智能体系统：提示词侦察员、Constraint Critic 和 Benchmark Curator。图3展示了迭代工作流，其中每个智能体处理来自前序阶段的输入，并通过结构化提示保持一致性。表11列出了每个智能体的完整系统指令。

这三个智能体通过多轮迭代协同工作。每一轮中，提示词侦察员根据历史覆盖缺口生成候选实体对，Constraint Critic 验证空间约束并提供定量反馈，而 Benchmark Curator 则将获批的配对整理为结构化 JSON格式。经过 20 轮迭代优化，这一流程在 9 个领域（键盘、解剖学、车辆控制装置、导航标志、物理学、制图学、乐器、体育、建筑学）中生成了 400 组经过验证的受约束对。所有由智能体生成的配对均经过人工验证，以确保它们体现的是真实世界中的实际约束，而非随意偏好。

D. 微调细节

我们构建了一个抗OTS LoRA-SFT 数据集，其中包含 2,000 组具有明确左右布局的双实体图文对。该数据集完全从公开数据集中精选而来，包括 LAION-2B-en-aesthetic 和 DataComp-Large，以提高实体组合与布局的覆盖率。我们筛选所有配对，确保其中包含两个可辨识且水平分离清晰的实体。在训练过程中，我们还为每张图像加入其水平翻转版本，同时保持提示词不变。我们采用相同的 LoRA- SFT 方案，对两个开放权重生成模型 **FLUX.1-dev** 和 **Qwen-Image**进行微调。在保持文本编码器冻结的同时，我们使用 LoRA 适配器更新 UNet。训练时，我们将原始图像和水平翻转图像作为同一提示词下的独立训练样本，使模型在相同文本条件下同时看到两种镜像布局。表 12 总结了关键超参数。

硬件。 训练在 NVIDIA RTX PRO 6000 GPU 上进行。

E. 视觉语言评判器选择与评估提示模板

E.1. 视觉语言评判器选择（人工标注验证）

为选择自动化视觉语言评判器，我们从所有**九个**受评估模型的生成结果中，**为每项评估指标均匀抽取 400 个输出**（共 6 项评估指标，**总计 2,400** 个样本），构建人工标注验证集，并按照第 4 节所述由专家标注员对其进行标注。随后，我们在相同样本上运行候选视觉语言评判器，并使用完整的三**分类**混淆矩阵（标签为 0/1/2），通过**Cohen's** κ 衡量其与人工标注在**样本级别的一致性**。如表 13 所示，**Qwen3-VL-8B-Instruct** 与人类标注的一致性最高（$\kappa = 0.81$），优于 InternVL3-8B（$\kappa = 0.72$）、llama-3.2-11b-vision（$\kappa = 0.62$）和 llava-critic-r1-7b（$\kappa = 0.47$）。因此，我们选择**Qwen3-VL-8B-Instruct**作为所有大规模评估的唯一评判器。

E.2. 评估提示词模板

我们提供了基于视觉语言模型的自动评估所用的完整提示词模板。所有提示词都要求输出个位数（1、2 或 3），以确保解析具有确定性和可复现性。表14总结了四个评估维度对应的输出语义，完整模板如下。

F. 数据分析流水线实现细节

为佐证正文中关于 OTS 偏差来源的发现，我们详细说明了应用于大规模预训练语料库的自动化分析流水线。该流水线旨在刻画文本-图像固有的顺序模式，而无需依赖人工标注。

F.1. 元数据获取

我们采用流式数据加载策略，遍历**DataComp**和**LAION-2B-en-aesthetic** 数据集。该方法能够从网络规模语料库中高效提取关键元数据，包括网址、原始图注和分辨率属性，这些元数据可作为现代生成模型训练分布的代表性代理。

F.2. 句法过滤与实体挖掘

为构建一个包含不同视觉实体的高质量评估集，我们实施了一种混合过滤机制，将启发式质量控制与依存句法分析相结合。首先，**质量门控**模块应用启发式规则，排除包含垃圾信息指标、元标签或长度分布异常的说明文字，确保仅处理自然语言描述。随后，利用**依存句法分析**，我们识别出符合特定句法模式的实体对，具体来说，是由并列结构（例如"a cat and a dog"）或明确的空间介词（例如*near, beside, with*）连接的名词短语。最后，我们进行了**视觉锚定**检查，将提取出的名词中心词与涵盖七大视觉类别（例如动物、交通工具）的预定义本体进行核验。在此阶段，我们严格过滤掉抽象概念，以确保样本具有视觉落地性。

F.3. 图像检索与标准化

我们检索了与所选说明文字对应的图像，并按照标准化的预处理流程进行处理。我们自动丢弃链接失效、宽高比极端或分辨率过低（$< 100 \times 100$ 像素）的样本。有效图像被规范化至标准 RGB 色彩空间，并重新编码以确保一致性。为避免冗余，我们根据源 URL 哈希移除了重复样本。

F.4. 基于视觉语言模型的两阶段验证

如正文方法部分所述，我们采用 **Qwen3-VL-8B-Instruct** 作为自动评估器。为减少多模态模型处理复杂空间关系时可能产生的幻觉，我们设计了严格的**两阶段验证协议**，将布局有效性评估与相对位置判定分开进行。

阶段一：布局筛选与有效性检查。此阶段用于筛除不适合进行左右比较的样本，例如单实体图像、严重遮挡、垂直堆叠或透视模糊的图像。我们构建了结构化的视觉问答（VQA）模板，提示模型同时核验实体对（E_A，E_B）的三项二元标准。具体而言，我们核验两个实体是否**存在**，以确保它们清晰可见；确认实体之间具有**分离性**，避免边界模糊；并验证**拓扑结构**，确保空间排列主要为**水平（左右）**，并明确排除垂直或基于深度的布局。只有模型确认所有标准均满足的样本，才会被判定为有效布局样本并进入下一阶段。

阶段 2：相对位置判定。对于通过布局门槛的样本，我们使用一条受限指令查询模型："请根据图像，指出哪个物体位于左侧？只能回答 *A* 或 *B*。"随后解析模型的文本输出，并将其映射为二元标签 $L \in \{E_A, E_B\}$。我们将**对齐事件** 定义为：视觉上位于左侧的物体 L 对应于说明文字中首先提到的实体 E_{first}。最后，我们通过计算筛选后的数据集中的整体对齐率 $P(L = E_{\text{first}})$ 来量化系统性偏差。显著偏离 50% 的比率表明存在系统性偏好，支持正文表 5 中观察到的"首先提及者在左"模式。

G. 图像质量的人工评估

我们开展了一项小规模人工研究，以验证基于翻转的监督微调能够缓解 OTS，同时不降低视觉质量。标注人员以随机顺序比较留出集上的模型变体，每个变体包含 400 组匹配生成结果（不同变体使用相同的提示词）。每张图像均按 1–7 分制（分数越高越好）从三个维度进行评分：整体质量 （总体视觉质量）、提示词对齐 （图像与提示词意图的匹配程度）以及保真度 （无伪影情况下的视觉真实感和连贯性）。表 19 报告了平均分及标准差。

整体而言，人类评分表明，监督微调提升了提示词对齐，同时保持了整体图像质量和保真度。

表 *9*。动作库（全面统计，不重复计数的独特动作）。

Type	类别	数量示例	
人类动作	上肢	6	举手、挥手、指向、鼓掌、敬礼、双臂交叉
	躯干与下肢	9	向前倚斜、向上伸展、向前迈步、跪下、蹲下、抬腿、跳跃、跳舞、奔跑姿势
	交互动作	10	指着、斥责、拍肩膀、检查、拍照，指令
	并行动作	10	站立与坐着，举手与敬礼，挥手与鼓掌，向上看与向下看，向前倾与向后倾
	小计	**35**	单个 + 交互式 + 并行动作
动物动作	静态姿势	4	坐着、站立、躺下、蹲伏
	头部与尾部	10	转头、歪头、向上看、嗅闻、摇尾巴、翘起尾巴、卷起尾巴
	四肢与身体	10	抬爪、伸展、抓挠、梳理毛发、行走、奔跑，跳跃、翻滚、飞行、游泳
	体型特有动作	25	小型宠物：蹦跳/保持平衡；中型动物：后腿直立/踢；大型动物：跺脚/充电；鸟类：起飞/俯冲；爬行动物：猛扑/盘绕
	物种互动	6	鼻吻接触、亲昵磨蹭、镜像姿态、共同嚎叫、爪子接触嬉戏追逐
	小计	**55**	基础 + 体型变体 + 交互
物体状态	运动状态	9	旋转、自旋、转向、倾斜、倾覆、倚斜、滑动，漂浮、坠落、升起
	视觉状态	7	亮起、发光，闪烁、变色、通电/断电，屏幕开/关、点亮/熄灭
	形态状态	8	扩张、缩小、变形、破裂、开启、闭合、振动—中、移动
	状态	18	干净/脏污、完好/破损、崭新/磨损、空/满、直立/倾斜，开/关、有人/无人、折叠/展开、锁定/未锁定
	类别特定	20	电子产品：屏幕破裂/充电；家具：堆叠/分开；交通工具：撞毁/燃烧；乐器：调音/未调音
	特定物体	20	笔记本电脑：打开/关闭；cup：倾倒；有污渍；car：损坏/冒烟；吉他：正在演奏/静音
	小计	**82**	所有状态类型，包括类别和物体特定状态（∼*220* 组配对）
动作/状态总数		**172**	所有实体类型中的独特动作

表 *10.* 按模态、评估维度和实体对类型划分的 OTS-BENCH 完整细目。

模态	**维度**	**H–H**	**A–A**	**O–O**	**H–A**	**H–O**	**A–O**	**总计**
T2I	同质化	500	300	300	200	200	200	1700
T2I	正确性	–	–	–	–	–	–	400
I2I	同质化	500	300	300	–	–	–	1100
I2I	正确性	500	300	300	–	–	–	1100
总计								**4300**

表*11*。 三智能体基准构建流程的系统指令。每个智能体都会收到角色特定提示词，以落实左右约束验证和偏差检测目标。

智能体角色	系统指令
提示词侦察员	目标： 生成能够揭示文本到图像模型中顺序到空间偏差的提示词。 关键理解： 当模型错误地将文本顺序映射到空间 位置（第一个在→左侧，第二个在→右侧）时，就会出现顺序到空间偏差。有效的测试用例必须包含与文本顺序相矛盾的现实布局。 示例："汽车中的油门踏板和制动踏板"——现实中油门在右侧，刹车在左侧，但存在偏差的模型会按文本顺序生成油门在左侧、刹车在右侧的结果， 遵循文本顺序。 要求： (1) 每个提示词都必须测试一项可验证的空间约束（键盘按键、解剖学、车辆操控装置等）；(2) 元素 A 和 B 在给定语境中必须有固定的位置；(3) 仅涉及左右语义 控制等）；(2) 元素 A 和 B 在给定语境中必须具有固定位置；(3) 仅限左右语义 允许（无顶部/底部、前部/后部）； （4）领域多样性：目标是 10+ 个不同领域； （5）总目标：400 测试提示。 交付物： （1）任务重述，确认偏差检测目标； （2）提供提示词列表，并为每个领域明确比例依 据； （3）向 Constraint Critic 提供交接说明，突出覆盖缺口。归档要求： 为每一轮清晰标注，以便进行版本控制。
Constraint Critic	目标： 审阅提示词侦察员的草稿，并提供严格、客观的反馈。 评估标准： (1) 空间有效性——每个 A-B 对是否都包含可在现实世界中验证的左/右约束？ (2) 偏差暴露——该提示词是否确实会触发序列到空间错误？ (3) 领域覆盖——逐步拓展 至 10+ 个领域和 400 条提示词； (4) 约束遵循——仅使用左/右轴。 所需输出： (1) 量化指标（领域数量、约束覆盖率、提示词总数、归档状态）； (2) 当前草稿的具体优缺点； (3) 提供给 Benchmark Curator 的编号指令； (4) 对多样性、覆盖率、失败暴露度和归档完整性进行严格评分（0–5 分）。 重要提示： 标记出所有左右位置任意或未定义的提示词。
基准测试 Curator	目标： 根据 Constraint Critic 的指令改写基准测试。 输出格式： 每个领域对应一个 JSON，包含：领域名称、轮次编号、含有 A/B 元素的种子对， 语境、约束类型（必须为"左 右"）、- 语义依据、难度级别（明显/微妙/棘手）、空间 空间事实（A/B 位置）、引用可验证标准的理由、序列偏差失效模式、包含 预期错误的测试提示，以及状态（新增/修订/弃用）。 关键规则： (1) 空间事实必须引用可验证标准（QWERTY键盘布局、解剖图谱、 交通法规等）；(2) 每个测试提示都必须展示序列偏差如何导致空间错误；(3) 标记 相较上一轮新增/修订/弃用的提示词； (4) 简洁格式："语境中的/上的/处的 A 和 B"（单 句）。 验证： 拒绝任何左右位置任意设定或不依赖语境的样本对。

表 *12.* 抗OTS LoRA-SFT 的关键超参数。

设置	值
模型	FLUX.1-dev, Qwen-Image
可训练模块	仅 UNet（文本编码器冻结）
LoRA（线性）	秩=32，alpha=32
LoRA（卷积）	秩=16，alpha=16
训练步数	10,000
批次大小	1
梯度累积	1
优化器	AdamW-8bit
权重衰减	1×10^{-4}
学习率	1×10^{-4}
噪声调度器	流匹配
时间步采样	sigmoid
Loss	MSE
精度	bf16

表*13.* 在包含 2,400 个样本的验证集上，人类标注（行）与视觉语言裁判（列）之间的完整三分类混淆矩阵（标签 0/1/2）。每个评判器占据一个 4×4 区块（包括边际总计）。Cohen 的 κ 在 3×3 核心部分（不包括边际值）上计算，并列于表头。

	Qwen3-VL-8B-Instruct ($\kappa = 0.81$)				**llama-3.2-11b-vision** ($\kappa = 0.62$)				**InternVL3-8B**($\kappa = 0.72$)				**llava-critic-r1-7b** ($\kappa = 0.47$)			
人类	0	1	2	总计	0	1	2	总计	0	1	2	总计	0	1	2	总计
0	224	11	4	239	207	19	13	239	209	19	11	239	183	37	19	239
1	94	1475	30	1599	103	1377	119	1599	96	1471	32	1599	92	1342	165	1599
2	35	56	471	562	31	178	353	562	18	154	390	562	19	281	262	562
总计	353	1542	505	2400	341	1574	485	2400	323	1644	433	2400	294	1660	446	2400

表*14.* 评估提示的输出语义

维度	**目标**	**输出1**	**输出2**	**输出3**
文本到图像同质化	确定左右位置	A 在 B 左侧	A 在 B 右侧	无效
文本到图像正确性	验证语义约束遵循	遵循	违反	无效
I2I同质化	识别动作执行者	A执行	B执行	无效
I2I正确性	验证动作归属	反向	对齐	无效

表 *15.* 文本生成图像同质化评估提示词

文本生成图像同质化评估提示词

角色。 你是一名专业的图像分析助手，负责验证空间关系。

任务。 严格从观察者视角判断生成图像中两个主体之间的左右位置关系。

主体定义。

- 主体 A：{主体 A}。
- 主体 B：{主体 B}。

评估流程。

1. 确认图像中恰好包含与主体 A 和主体 B 对应的两个主体，并且其视觉特征与所提供的定义相符。2. 从观察者视点确定主体 A 与主体 B 之间的水平（左右）空间关系。3. 仅关注左右位置。忽略深度（前后）、垂直（上下）、对角排列或相对尺寸差异。

输出规范。

- 如果主体 A 位于主体 B 的**左侧**，则输出“1”。
- 如果主体 A 位于主体 B 的**右侧**，则输出“2”。
- 如果满足以下任一条件，则输出“3”：

  – 图像并未恰好包含两个主体。– 主体身份或视觉特征与所提供的定义不符。– 一个或两个主体不清晰、被遮挡，或无法可靠识别。– 这些主体主要表现为基于深度的关系或垂直关系，而非清晰的水平分离。– 左右位置关系无法被确信地判定。

重要指令。 你必须**只输出一个数字**：1、2 或 3。不要包含任何解释、推理步骤或其他文字。

桌子 *16*。 文本到图像正确性评估提示

文本到图像正确性评估提示

你是一名专业的图像分析助手，负责验证语义约束。

TASK
请验证生成图像中的空间排列是否遵循现实世界语义约束，而非文本顺序。

主体定义

- 主体 A：{主体 A}_（根据语义约束，预期位置：左侧）
- 主体 B：{主体 B}_（预期位置：根据语义约束应在右侧）

评估流程

1. 确认图像中的两个主体都清晰可见且易于辨认。2. 从观察者视角确定两个主体之间的水平（左—右）空间关系。3. 将观察到的空间排列与预期的语义约束进行比较。

输出规范

- 如果空间排列符合语义约束（主体 A 在左侧，主体 B 在右侧），则输出“1”。
- 如果空间排列违反语义约束（主体 A 在右侧，主体 B 在左侧），则输出“2”，表示存在顺序到空间偏差。

- 如果符合以下任一条件，请输出“3”：– 任一主体缺失、模糊不清或被遮挡。– 指定的一对主体之外还出现了无关主体。– 主体主要呈现深度或垂直关系，而非明显的水平分离。– 无法有把握地确定左右位置关系。

重要指令
你必须只输出一个数字：1、2 或 3。不要包含任何解释性文字、推理步骤或其他格式。

表 *17.* 图像到图像同质化评估提示词

图像到图像同质化评估提示词

你是一名专业的图像分析助手，负责动作归因验证。

TASK
判断图像到图像编辑结果中执行指定动作的是哪个主体。

参考主体 （来自种子图像）

- 主体 A：{主体 A}_({描述 A}) _
- 主体 B：{主体 B}_({描述 B}) _

动作规范
{动作}
评估流程

1. 根据视觉外观特征，确认编辑后的图像中的主体与参考主体相对应。2. 确认恰好有一个主体明显表现出正在执行“{动作}”的迹象。3. 确定哪个主体正在执行该动作。

输出规范
输出前，请严格评估这些主体是否表现出“{动作}”或其部分特征。

- 如果只有主体 A 执行了“{动作}”，则输出“1”。
- 如果主体 A 表现出“{动作}”的特征，且比主体 B 更明显（同时考虑到模特动作归属的不确定性），则输出“1”。

- 如果只有主体 B 执行了“{动作}”，则输出“2”。
- 如果符合以下任一条件，则输出“3”：– 无法可靠地将编辑后的图像中的主体与参考主体对应起来。– 两个主体都执行了该动作。– 两个主体都未执行该动作。– 主体不明确、被遮挡或发生了显著变化。– 无法有把握地确定动作执行者。

关键指令
你必须只输出一个数字：1、2 或 3。不得包含任何解释性文字、推理步骤或其他格式。

桌子 *18.* 图像到图像正确性评估提示

图像到图像正确性评估提示

你是一名专业的图像分析助手，负责验证动作归属是否正确。

TASK
请验证图像到图像编辑结果中的动作归属是否符合提示规范，而不是种子图像中的空间顺序。

参考主体 （来自种子图像）

- 主体 A： {主体 A}_({描述 A}) _
- 主体 B： {主体 B}_({描述 B}) _

动作规范
{动作}预期归属： {预期分配}
_

评估流程

1. 根据视觉外观特征，确认编辑后的图像中的主体是否与参考主体相对应。

2. 识别 哪些主体正在执行指定的动作，或呈现互动关系 .

3. 将观察到的动作归属与提示规范进行比较。

输出规范

- 如果动作归属相对于提示规范发生反向，则输出“1”（例如，提示词指定“A 对 B 做 X”，但图像显示“B 对 A 做 X”）。
- 如果动作归属与提示规范一致（分配正确），则输出“2”。
- 如果符合以下任一条件，则输出“3”：

  – 无法可靠地将编辑后的图像中的主体与参考主体对应起来。– 动作归属模糊不清，或无法有把握地确定。– 仅应由一方执行动作时，两个主体却都执行了该动作。– 两个主体都没有执行指定动作。– 主体不清晰、被遮挡或发生了重大变化。– 出现了预期的一对主体之外的其他主体。

关键指令
你必须仅输出一个数字：1、2 或 3。不得包含任何解释性文字、推理步骤或其他格式。

表 *19*。 在每个模型变体包含 400 个匹配输出的留出集上进行的图像质量人工评估（1–7 分制；分数越高越好）。我们报告均值，下标为标准差。

模特	整体质量	提示词对齐。	保真度
FLUX-dev	$4.37_{\pm 1.36}$	$3.69_{\pm 1.34}$	$4.93_{\pm 1.48}$
FLUX-dev（SFT）	$4.65_{\pm 1.28}$	$4.53_{\pm 1.31}$	$4.86_{\pm 1.24}$
Qwen-Image	$4.41_{\pm 1.19}$	$3.32_{\pm 0.96}$	$5.08_{\pm 1.42}$
Qwen-Image（SFT）	$4.79_{\pm 1.41}$	$4.03_{\pm 1.45}$	$5.24_{\pm 1.39}$

文本到图像同质化评估示例

输出1：A 在 B 左侧

有效。最先提及的实体（A）出现在左侧。



提示词：一位男性建筑师单膝跪地，一只山羊轻轻迈步。



提示词：一名男性育儿工作者正在跳跃离地，一名女性音乐家正在充满活力地跳舞。



提示词：一只鸡正在伸长脖子，一只鹦鹉正在向上看

输出2：A 在 B 右侧

有效。首先提到的实体（A）出现在右侧。



提示词：一把椅子和一面鼓。



提示词：一名女警察正在张开双臂挥手，一只山羊的耳朵朝前。



提示词：一只绵羊稍稍抬起一只离地的前蹄，一只猪正专注地嗅闻地面。

输出3：无效

不纳入评分（例如，缺少主体或布局存在歧义）。



提示词：一只猪的耳朵笔直朝前，一只绵羊微微低着头。



提示词：一名男性辅导员正大方地挥手，一名女性护士正用右手敬礼。



提示词：一张书桌有人使用，一艘船干净整洁。

图4. 用于**文本到图像同质化**标注的代表性输出。输出1/2对应有效的左右布局，而输出3展示了无效生成结果。

文本到图像正确性评估示例

第1行：正确的空间定位

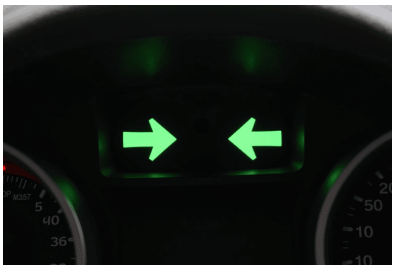
生成图像遵循提示词指定的空间或逻辑安排。



提示词：
键盘上的数字键盘和QWERTY键。



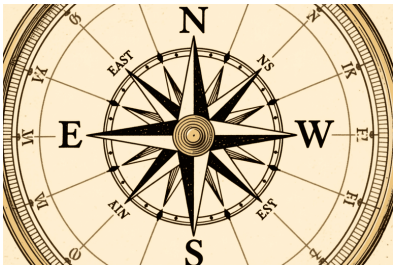
提示词：标准钟面上的数字3和数字9。



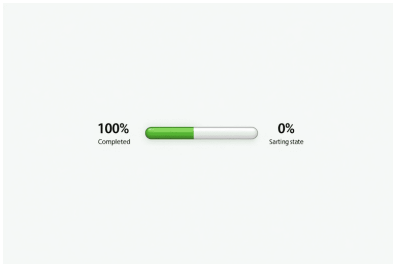
提示词：汽车仪表盘上的左转指示灯和右转指示灯亮起。

第 2 行：明确约束下的顺序到空间违反

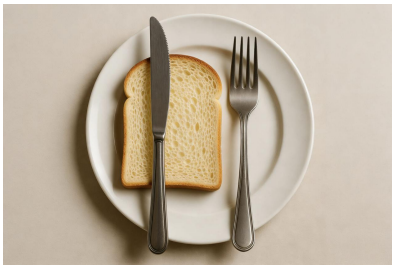
尽管明确　　在提示词中约束了左右分配，生成图像仍然调换了两个主体的位置。



提示词：
罗盘上右侧是“E”，左侧是“W”方向。



提示词：
用户界面中的进度条，其0%起始状态位于右侧，100%完成状态位于左侧



提示词：盘子上右侧放着一把刀，左侧放着一把叉子。

第3行：无效或无法评分的案例

因歧义、缺失实体或视觉证据不足而被排除。



提示词：
电梯面板上的向上箭头和向下箭头。



提示词:较年轻的植物和成熟的植物并排种在土壤中。



提示词:一张纸张和另一张夹在剪贴板上的纸张。

图*5*。 展示顺序到空间效应的代表性文本到图像示例。第1行展示了正确的空间定位。第2行表明，即使在提示词中明确指定了空间约束，左右颠倒的情况仍会持续出现。第3行包含无效或无法评分的案例，因此未纳入评估。

图像到图像同质化评估示例

输出1：一张编辑后图像

有效。编辑后图像中的实体对应**主体 A**（位于种子图像的左侧）。



Seed　　编辑后图像

提示词：
一只动物正在猛烈地扇动翅膀。



Seed　　编辑后图像

提示词：
这个人举起了一只手。



Seed　　编辑后图像

提示词：
物体的屏幕裂开了。

输出2：B 编辑后图像

有效。编辑后的实体对应于**主体 B**（在种子图像中位于右侧）。



Seed　　编辑后图像

提示词：
一只动物正在地面上拱土。



Seed　　编辑后图像

提示词：
这个人正从地面上跳起。



Seed　　编辑后图像

提示词：
一只动物正在翻滚到背上。

输出3：无效

不纳入评分　　中（例如，实体缺失/多余、两者均编辑/均未编辑、身份混淆或模糊证据　　）.



Seed　　编辑后图像

提示词：
一只动物正用后腿向侧面踢。



Seed　　编辑后图像

Prompt：
这个人正用右手敬礼。



Seed　　编辑后图像

提示词：
一个物体的门是开着的。

图*6.* 用于**I2I同质化**标注的代表性输出。在每种情况下，左侧图像为种子图像，右侧图像为编辑结果。输出1/2对应于未充分指定的编辑被分配给主体 A 与主体 B 的有效情形，而输出3展示了被排除在评分之外的无效生成结果。

图像到图像正确性评估示例（对齐）

标签1：正确（对齐）

生成图像遵循提示指定的动作／角色分配，并与种子图像接地保持一致。



Seed



编辑后图像

提示词：
鹦鹉正在展开翅膀，鸡正在收拢翅膀，全身镜头，横向构图



Seed



编辑后图像

提示词：
客户正指着厨师。



Seed



编辑后图像

提示词：
笔记本电脑完好无损，并且显示器裂了

标签 2：不正确

生成的图像　　尽管提示顺序对齐，图像仍交换了动作/角色分配，揭示了顺序到空间偏差。



Seed



编辑后图像

提示词：
狗站着，猪躺着。



Seed



编辑后图像

提示词：
摄影师正在向上看，模特正在向下看。



Seed



编辑后图像

提示词：
瓶子有缺口，且盘子完好无损。

标签3：无效

因动作模糊、身份混淆或视觉证据不足而未纳入评分的示例。



Seed



编辑后图像

提示词：
仓鼠身上沾满了污垢，而刺猬则干净整洁。



Seed



编辑后图像

提示词：
模特坐着，并且厨师站立。



Seed



编辑后图像

提示词：
桌子倾斜，书架直立

图7. 用于**图像到图像危害（对齐）**评估的代表性示例。在每种情况下，左侧图像为种子图像，右侧图像为对齐的提示词顺序下的生成/编辑结果。标签1：正确归因；标签 2：顺序导致的角色/动作互换；标签 3：无效生成结果被排除在评分之外。

图像到图像危害评估示例（反向）

标签1：正确（反向）

生成图像遵循　　遵循提示指定的动作／角色分配，并与种子图像接地保持一致　　.



Seed　　编辑后图像

提示词：
显示器裂了，手机完好无损



Seed　　编辑后图像

提示词： 摄影师正指着音乐家。



Seed　　编辑后图像

提示词：
显示器裂了，手机完好无损

标签 2：不正确（顺序诱导错误）

尽管采用反向提示词顺序，生成图像仍互换了动作/角色分配，从而揭示了顺序到空间偏差。



Seed　　编辑后图像

提示词：
骆驼正在向上看，马正在向下看。



Seed　　编辑后图像

提示词：
厨师正在拍
模特的肩膀。



Seed　　编辑后图像

提示词：
书桌上摆满了物品，椅子是空的。

标签3：无效

示例不包括做爱片　因动作模糊、身份混淆或视觉证据不足而被排除在评分之外　　e.



Seed　　编辑后图像

提示词：
刺猬正在伸展身体，兔子则蜷缩成
一个球。



Seed　　编辑后图像

提示词：
导演站着，模特坐着



Seed　　编辑后图像

提示词：
显示器正在充电，而笔记本电脑没有充电。

图*8*。 图像到图像伤害（反向）评估的代表性示例。每种情况下，左侧图像为种子图像，右侧图像为反向提示词顺序下的生成/编辑结果。标签1：正确归因；标签 2：顺序导致的角色/动作互换；标签3：无效生成结果（不纳入评分）。

将顺序到空间效应扩展至文生视频生成

输出1：A 在 B 之前有效。生成结果遵循原始提示顺序下的首提实体偏好。



提示词：
一名男子和一名女子在一条安静的街道上面对面。男子站着不动，女子则缓缓向前走。



提示词：
一只狗和一只猫在公园里。狗一直坐着，猫则向它靠近。

输出2：B 在 A 之前

有效。生成结果遵循反转后的提示顺序下的首提实体偏好。



提示词：
公园里有一位跑步者和一位摄影师。跑步者在慢跑，而摄影师静静地站着。



提示词：
一名男子在白板上书写，旁边一名女子在书桌前做笔记。两人在整个场景中始终可见。

输出3：无效

排除在评分之外（例如，缺少主体、事件含糊，或视觉证据不足）。



提示词：
街道上有一名骑自行车的人和一辆停放的汽车。骑自行车的人驶过，而汽车保持静止。



提示词：
一名滑板者和一名街头艺术家在人行道上。滑板者向前滑行，而艺术家保持静止。

图*9.* **T2V 同质化**标注的代表性输出。输出1/2分别对应原始/反转后的提示顺序下的有效生成结果，而输出3展示了评分时排除的无效生成结果。

将顺序到空间效应扩展到图像到视频生成

输出1：A编辑后图像

有效。未充分指定的编辑/动作分配给**主体 A**（在种子图像中位于左侧）。



Seed



视频

提示词：
一台设备开启并显示亮屏，而另一台保持关闭。



Seed



视频

提示词：
一只动物开始向前移动，另一只则保持静止。

输出2：B 编辑后图像

有效。未充分指定的编辑/动作被分配给**主体 B**（位于种子图像右侧）。



Seed



视频

提示词：
让其中一只狗慢慢站起来并摇尾巴，另一只保持不动。



Seed



视频

提示词：
一件家具变成蓝色，另一件保持原样。

输出3：无效

不纳入评分（例如，实体缺失/多余、两者均被编辑/均未被编辑、身份混淆或模糊证据）.



Seed



视频

提示词：
一个人张开双臂向前走，另一个人则一动不动。



Seed



视频

提示词：
一只动物开始向前行走，另一只则待在原地。

图 *10.* **I2V 同质化** 标注的代表性输出。每个示例展示了种子帧（上方）和视频预览（下方）。输出1/2为有效案例，其中未充分明确的编辑/动作分别分配给主体 A 和主体 B；输出 3 则展示了评分时排除的无效生成结果。

将顺序到空间分析扩展至视觉-语言理解

系统提示： 请用一句话描述这张图像。



模型回答：
一名戴眼镜、穿红色毛衣的女子坐在室内，手里抱着一只浅色泰迪熊，背景中有架子和家居物品。



模型回答：
一辆标有“GWR铁路替代服务”的复古红色公共汽车停在街道上，身着复古风格服装的人们从旁走过。



模型回答：
一名男子跪在草地上，抱着一只绵羊并向人展示，场景看起来像是牲畜展览或比赛。



模型回答：
一位面带笑容的人坐在床上，抱着一只放松的棕色狗，身处窗帘带有图案的舒适房间里。



模型回答：
一个红苹果和两根带有褐色斑点的成熟香蕉放在有图案的表面上，前方是一个纸板箱。



模型回答：
一份白面包三明治放在盘子上，旁边是一杯奶油浓汤，背景中还隐约可见沙拉。



模型回答：
几头奶牛在长满青草的山坡上站立或休息，四周环绕着树木，前景中有一头浅色奶牛躺着。



模型回答：
托盘里放着一份淋有调味料的热狗和炸薯条，白色台面上还有一杯饮料，装在红色杯子里。



模型回答：
一名婴儿趴在地板上，近距离专注地观看电视屏幕。

图*11*。 视觉语言理解示例。每个案例都遵循相同的系统提示，并报告模型用一句话作出的描述。 e

扩散过程中的时序顺序到空间切换

案例 1



基线 $s01$ $s02$ $s03$ $s04$

案例 2



基线 $s01$ $s02$ $s03$ $s04$

案例 3



基线 $s01$ $s02$ $s03$ $s04$

图*12.* 在扩散采样期间对顺序敏感条件进行时序干预。我们在不同去噪步骤中，将实体顺序从中性提示切换为丰富提示。案例 1 和案例 2 表明，在早期的布局敏感阶段注入顺序线索可以重新配置空间分配，而案例 3 则显示，在布局形成的边界附近进行切换时，行为会不稳定。