arXiv:2603.01063v1 [cs.CV] 1 Mar 2026

通过从失败中显式学习释放视觉-语言-动作模型在自动驾驶中的潜力

Yuechen Luo[1*]、Qimao Chen[1*]、Fang Li[2*]、Shaoqing Xu[2、‡]、Jiaxin Liu[1]、Ziying Song[3],Zhi-xin Yang[2、✉]、Fuxi Wen[1、✉] [1] 清华大学、 [2] 澳门大学、 [3] 北京交通大学luo-yc24@mails.tsinghua.edu.cn

摘要

用于自动驾驶的视觉-语言-动作（*VLA*）模型在强化学习（*RL*）优化过程中通常会陷入性能平台期。这种停滞源于其探索能力受到先前监督微调（*SFT*）的限制，导致模型在长尾场景中出现“持续失败”。在这些关键情境下，所有探索出的动作都会产生零驾驶得分。这种信息稀疏的奖励信号能够表明失败，却无法确定其根本原因——究竟是规划错误、推理有误，还是轨迹执行不佳。为解决这一局限，我们提出了具有从失败中显式学习（***VLA**with* 显式学习失败（***ELF-VLA***）框架的方法），通过结构化诊断反馈增强强化学习。我们的方法不依赖模糊的标量奖励，而是生成详细且可解释的报告，用于识别具体的失败模式。随后，*VLA*策略利用这一显式反馈生成反馈引导修正。通过将这些经过纠正的高奖励样本重新注入强化学习训练批次，我们的方法提供了有针对性的梯度，使策略能够解决无引导探索无法应对的关键场景。大量实验表明，我们的方法释放了*VLA*模型的潜在能力，在公开*NAVSIM*基准上，于整体*PDMS*、*EPDMS*和高层规划准确率方面取得了最先进水平的性能。

1. 引言

自动驾驶系统的发展正经历一场范式转变，从传统的模块化架构转向端到端框架 [4, 11]。视觉-语言-动作（VLA）模型处于这一转变的前沿 [15]。这些模型通过监督微调（SFT）和强化学习（RL），将原始摄像头传感器输入映射为连贯的车辆运动指令

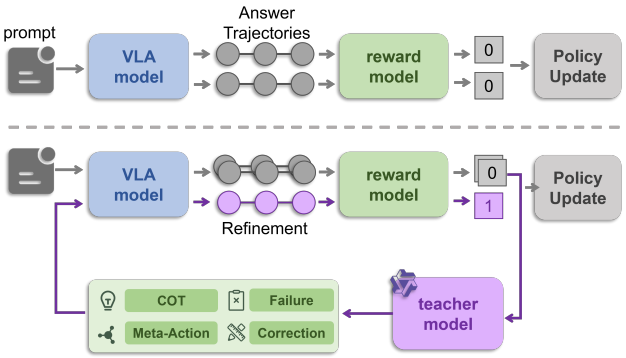


图 1. **通用视觉-语言-动作模型与 ELF-VLA 的强化学习微调对比**。上：使用强化学习算法训练视觉-语言-动作模型时会遭遇性能平台期：在某些场景中，策略模型的推演持续产生低分答案，使智能体陷入困境，无法发现更优策略。下：ELF-VLA 通过教师模型提供结构化反馈，并利用这些反馈重新进行推演以优化策略，从而促使策略突破这一性能平台期。

应用于大型视觉-语言模型（VLM）。这种集成式设计消除了人工设计的接口，并支持大规模、数据驱动的策略学习。值得注意的是，VLA模型可以通过“思考”模块生成中间推理轨迹，模仿人类的问题解决策略，为实现可解释、可信赖的自动驾驶提供了一个很有前景的方向 [17, 25, 41]。

尽管取得了这些进展，强化学习微调仍持续表现出性能平台期：我们观察到，经过监督微调后，模型的策略探索能力受到监督微调数据集局限性的严重制约；该数据集中常见场景占比很高，而严格考验自主系统能力的安全关键场景则十分罕见 [8, 23]。因此，在安全关键且具有挑战性的场景中（例如复杂的无保护左转或紧急避让），所有探索性推演都会持续失败，驾驶得分为零，如图1上排所示。现有的视觉-语言-动作强化学习方法

*同等贡献。‡项目负责人。 ✉ 通讯作者。

将训练期间的性能评估简化为单一标量奖励（例如PDMS [7]）。模型发生失败时，这种信息稀疏的奖励不足以确定错误的根本原因，因此难以判断失败究竟源于“思考”模块中高层规划的累积误差、对关键目标的认知推理存在缺陷，还是低层轨迹的动态能力不足。

为克服这些局限并实现持续学习，本文提出了一种面向自动驾驶的全新视觉-语言-动作训练框架，以弥合失败诊断与策略修正之间的鸿沟。如图1底行所示，该方法的核心思想是提供包含结构化失败分析的反馈，帮助视觉-语言-动作模型发挥其“先思考后行动”架构的作用，而非依赖简单的标量奖励。该方法有两项核心创新：

- **VLA能力对齐反馈：**我们引入了一种由教师模型提供的反馈机制，当视觉-语言-动作模型遭遇持续性失败时，该机制便会触发。教师模型会生成一份与VLA能力相匹配的结构化诊断报告，指出视觉-语言-动作模型在规划、推理或执行层面的具体错误。
- **反馈引导精炼与再注入：**视觉-语言-动作策略模型（学生模型）利用这份诊断报告生成修正轨迹。随后，这个高奖励的修正样本会被重新注入GRPO训练批次。该过程提供了目标导向的梯度信号，而这种信号此前在推演批次中并不存在。

通过在NAVSIM基准上的广泛评估，我们的方法相较现有视觉-语言-动作基线方法展现出显著的性能提升。我们的方法在总体驾驶指标（PDMS）和高层规划准确率上均达到最先进性能。通过将可解释反馈与策略修正相衔接，本工作为VLA模型克服自动驾驶中的性能平台期提供了一条切实可行的路径。

2. 相关工作

用于自动驾驶的视觉-语言-动作模型。 近年来，通过整合视觉与文本数据，实现统一的感知、规划和决策，视觉语言模型（VLMs）在自动驾驶领域受到广泛关注。目前，该领域主要有两种范式。第一种专注于场景理解和高层推理 [13, 14, 26, 28, 33]。Senna [13], 是这种方法的一个例子，它处理传感器输入，为下游规划器生成元动作，但实际驾驶性能尚未得到显著提升。另一种范式则专注于直接根据原始输入预测驾驶轨迹 [10, 12, 19, 24, 30, 34, 35, 40]。为了提高模型的可解释性和准确性，一项值得关注的发展是越来越多地采用中间推理（例如，思维链）来揭示模型内部的认知过程。来自EMMA [12], 、ReasonPlan [24], 和Sce2DriveX [40]的证据证实，领域特定推理能够显著提高轨迹预测的精度

.VLA模型的强化学习微调。 目前，自动驾驶中的VLA模型通常采用两阶段范式进行训练：先在驾驶数据集上进行初始监督微调（SFT），随后进入强化学习（RL）训练阶段。在该框架中，强化学习阶段的效果在很大程度上取决于先前监督微调的性能。当前方法 [17,25, 41] 采用组相对策略优化（GRPO） [31] 算法进行强化学习训练，其中奖励由视觉-语言-动作模型的驾驶得分（例如PDMS）衡量。这导致一个显著的训练缺陷：视觉-语言-动作模型在监督微调阶段之后，难以处理训练数据集中存在的罕见长尾场景。因此，当模型进入强化学习阶段时，无论推演次数多少，其在这些特定场景中的驾驶得分都仍然极低。这使模型的整体学习陷入停滞，从而形成性能平台期。在大语言模型文献中，已有方法成功使用非数值反馈（如文本批评）来提供详细指导 [3, 39]。另一些工作 [27, 37] 则采用混合策略方法，从高质量数据中内化知识，从而增强探索并提升策略质量。受这些方法启发，本工作引入一种反馈机制来解决自动驾驶领域中的这一局限，该机制采用教师模型，通过结构化反馈分析并纠正视觉-语言-动作模型的错误驾驶行为，将其修正为正确动作。

3. 方法

在本节中，我们介绍所提出的方法（图2），该方法包含两个主要组成部分：（1）两阶段监督微调（SFT）流程；（2）结合失败反馈的强化学习（RL）框架。

3.1. 视觉-语言-动作输入的形式化

在我们的方法中，视觉-语言-动作模型同时充当生成器和优化器，旨在接收两种不同类型的输入：不含反馈的原始**基础输入**，以及包含纠正性指导的**反馈输入**。

基础输入。基础输入查询 q^{base} 包括一张前视图像（记作 I_{cam} ）、高级导航指令 q_{com} （例如直行、左转、右转）、自车状态信息q_{ego} （例如速度和加速度），以及以 2Hz 频率记录的最近三帧历史轨迹 $q_{\text{his}} = \{h_{t-3}, h_{t-2}, h_{t-1}\}$ 。

反馈输入。 基于基础输入，视觉-语言-动作模型输出一个原始响应 o，由轨-

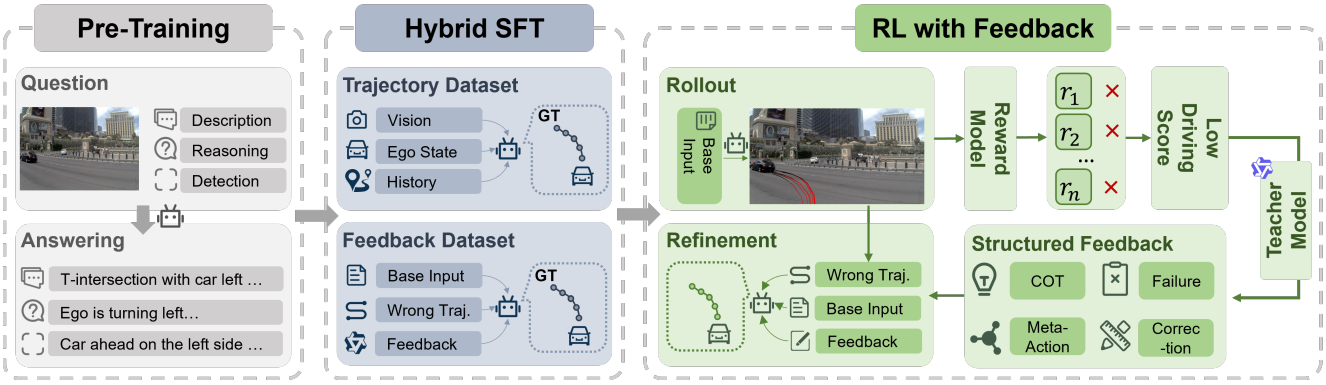


图 2. **ELF-VLA 概览**。首先，在自动驾驶问答数据集上对模型进行预训练，为其提供基础驾驶知识。随后，使用由“基础输入”和“反馈输入”组成的混合数据集对模型进行监督微调，使其能够同时学习轨迹预测和基于反馈的优化。最后，在强化学习阶段，使用教师模型生成反馈，从而降低零奖励推演的比例。

迹及思维链组成（详见附录6.2），随后根据阈值 s进行分类：PDMS超过 s 的响应被视为“正确”（o_c），而得分低于 s 的响应则被视为“错误”（o_w）。

$$q_{fb} = \begin{cases} < q_{base}, o_c, f^{rule} > & \text{if } o_c, \\ < q_{base}, o_w, f^{teacher} > & \text{if } o_w. \end{cases} \tag{1}$$

对于正确响应，相应的反馈输入由三个部分组成：原始基础输入 q_{base}、正确响应 o_c 本身，以及基于规则的正向反馈 f^{rule}。对于错误响应，我们通过视觉语言模型教师进行外部干预。该教师模型以基础输入 q^{base}、错误轨迹 o_w和真实轨迹o_{gt} 为输入，生成结构化反馈 $f^{teacher}$。该反馈包括：(1) 元动作分析；(2) 思考过程分析；(3) 安全失效分析；(4) 效率失效分析；以及 (5) 可执行的修正（包含横向和纵向分量）。随后，将基础输入、原始错误响应 o_i和生成的结构化反馈组合起来，构成视觉-语言-动作模型的最终反馈输入。基础输入和反馈输入的详细示例见附录6.2。

3.2. 认知与优化的两阶段监督微调

我们采用两阶段监督微调流程，以开发一个兼具驾驶知识和轨迹规划能力的模型。第一阶段旨在向模型注入通用驾驶知识。第二阶段专注于赋予模型轨迹预测能力，以及根据接收到的反馈进行优化的能力。

如图所示2，在第一阶段，模型在一个大规模驾驶相关问答对数据集上进行预训练，以增强其对驾驶领域认知的理解能力。

该数据集整合了多种开源驾驶问答数据集，包括 DriveLM [32], LingoQA [28], ImpromptuVLA [5], 以及其他开源驾驶数据集 [9, 29, 34]。此外，我们遵循思维链范式，为 NAVSIM 构建了一个多轮问答推理数据集。此阶段涵盖道路边界估计（可行驶区域）、关键目标识别（目标定位）、自车动作预测以及相关交通语义等任务。关于数据集组成的更多详情，请参阅附录6.1。

随后，第二阶段引入轨迹预测与优化任务。对于每个查询 q^{base} 和q^{fb} （定义见第 3.3节），模型的输出由真实轨迹 o_{gt}进行监督，旨在最大化条件似然：

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(q,o)\sim\mathcal{D}}\big[-\log \pi_\theta(o \mid q)\big]. \tag{2}$$

$\mathcal{D}$ 表示包含 $\{(q_{base}, o_{gt}), (q_{fb}, o_{gt})\}$的数据集，而 π_θ 表示视觉-语言-动作模型。这种混合数据集训练方法赋予模型轨迹预测和基于反馈进行轨迹优化的双重能力，从而使模型能够在强化学习阶段利用失败反馈。

3.3. 利用失败反馈进行强化学习

我们的失败反馈机制受 [39] 启发，并应用于 GRPO 算法的推演阶段。传统的自动驾驶视觉-语言-动作模型在强化学习训练期间通常会遇到性能瓶颈。这是因为它们无法应对复杂的长尾场景；因此，在这些场景中采样的轨迹会获得极低的驾驶得分，导致奖励稀疏问题。我们通过引入一种反馈机制来解决这一问题，该机制能够有效提升模型在这些关键场景中的驾驶得分，从而帮助智能体突破性能平台期。

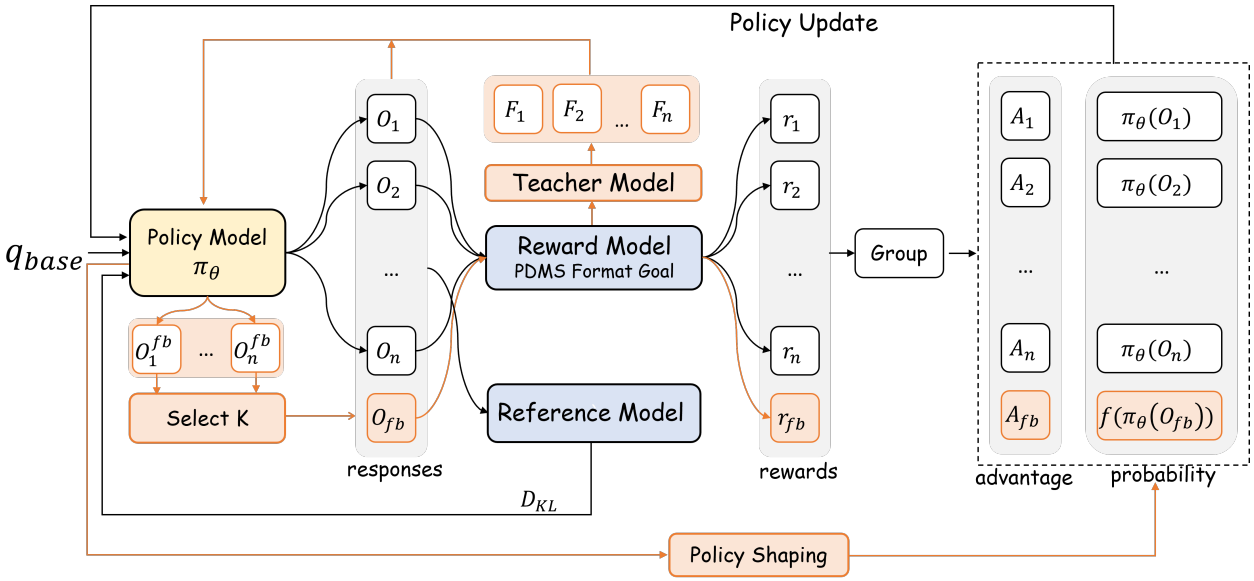


图 3. **带反馈的 GRPO 概览**。策略模型生成初始响应。教师模型（Qwen3-VL-32B）根据奖励提供反馈，引导策略采样更优的精炼响应。选取一条高质量的精炼响应，并将其与初始响应集结合进行联合优化。对最终概率应用策略塑形。

高效困难样本筛选。在介绍带有反馈机制的 GRPO 之前，我们首先以较低成本筛选数据，以最大限度地提高训练效率。朴素的强化学习训练往往会将资源浪费在过于简单（已掌握）的场景上，这些场景提供的学习信号较弱。我们的筛选旨在过滤掉这类样本，让智能体专注于高价值且信息丰富的场景，包括困难样本（模型持续失败的场景）和模糊样本（模型最不确定的场景）。为此，我们利用监督微调模型为每个查询采样 N 条推演，估算其平均奖励和奖励方差。随后，我们剔除平均奖励高且方差低的样本，因为这表明模型能够持续成功应对这些场景。该策略有效地将训练集中在困难场景（平均奖励低、方差低）和模糊场景（方差高）上。通过这种方法，我们将最初的 8.5 万条训练数据筛选为包含 2.4 万个高价值场景的核心数据集。

奖励建模。 为了激励视觉-语言-动作模型学习有效的驾驶行为，并确保其输出格式稳定，我们设计了一个包含三个组成部分的奖励函数：PDMS奖励、格式奖励和目标奖励。 **PDMS奖励** r_{traj} 是一种基于预测驾驶员模型评分的综合轨迹评估指标 [7]。其表示为一个介于0和1之间的连续值。计算该评分所用的具体公式见附录 8.1。 **格式奖励** r_{fmt} 是一种二元（1或0）奖励，旨在严格确保遵循所要求的输出格式。最后，**目标奖励** r_{goal} 通过赋予分级奖励来激励终点精度，该奖励基于与真实终点之间的L1距离。各项奖励的详细计算方法见附录7.1。强化学习过程中的总体奖励通过整合设计的四个奖励组成部分计算得出，具体如下：

$$r = r_{traj} + r_{fmt} + r_{goal}. \tag{3}$$

带反馈的 GRPO。 如图3所示，我们的方法采用反馈机制来优化轨迹并提高奖励，从而使视觉-语言-动作模型突破其性能平台期。具体而言，首先使用基础输入 $\{o_i\}_{i=1}^n$ 对 q^{base}进行采样，生成一批轨迹响应$\{r_i\}_{i=1}^n$ 。随后计算这批响应的奖励 $\{r_{traj,i}\}_{i=1}^n$，$\{r_{fmt,i}\}_{i=1}^n$以及 $\{r_{goal,i}\}_{i=1}^n$。根据预定义的阈值 s，再将这些响应分为两组：正确响应 $\{o_c\}$ 和错误响应 $\{o_w\}$。最后，按照第 3.1节所述，分别处理正确响应和错误响应，然后将其整合，构成最终的反馈输入 $\{q_i^{fb}\}_{i=1}^n$。

随后，视觉-语言-动作模型根据反馈输入（$\{q_i^{fb}\}_{i=1}^n$）生成一批新的响应 $\{o_i^{fb}\}_{i=1}^n$，并计算其奖励 $\{r_i^{fb}\}_{i=1}^n$。我们从轨迹奖励（r_{traj}^{fb}）高于原始批次最高轨迹奖励 $max\ (r_{traj})$ 的响应子集中，随机选取 k 个响应 $\{o_i^{fb}\}_{i=1}^n$ 。如果此类“更优”响应少于 k 个，则用达到最高奖励的原始响应重复填充剩余名额。由此得到最终一批 $n+k$ 个展开样本 $\{(q_i^{final},\ o_i^{final})\}_{i=1}^{n+k}$，并据此计算相对优势

$\{A_i\}_{i=1}^{n+k}$。推演更新算法的详情见算法 1。GRPO 的最终优化目标定义如下：

$$\mathcal{J}(\theta) = \mathbb{E}_{D_{final}\sim\pi_{\theta_{old}}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{J}_i + \frac{1}{k}\sum_{j=1}^{k}\mathcal{J}_j^{fb} - \beta\mathbb{D}_{KL}\right], \tag{4}$$

$$\mathcal{J}_i = \min\left(c_i(\theta)A_i, \text{clip}\left(c_i(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right), \tag{5}$$

$$\mathcal{J}_j^{fb} = f(c_j^{fb}(\theta))A_j^{fb}, \tag{6}$$

其中，$D_{final} = \{(q_i^{final}, o_i^{final})\}_{i=1}^{n+k}$，$\pi_{ref}$ 是初始监督微调模型的参考策略，β 是一个超参数。我们仅对原始批次推演样本应用裁剪，而不应用于经反馈优化后的响应。为计算优势值，我们首先将两组奖励合并为一个统一集合 r_{union}。随后，使用该集合的均值和标准差对奖励进行归一化，并计算相对优势A_i 和 A_j^{fb}，如下所示：

$$r_{union} = \{r_j\}_{j=1}^{n} \cup \{r_{j'}^{fb}\}_{j'=1}^{k}, \tag{7}$$

$$A_i = \frac{r_i - \text{mean}(r_{union})}{\text{std}(r_{union})}, A_i^{fb} = \frac{r_j^{fb} - \text{mean}(r_{union})}{\text{std}(r_{union})}, \tag{8}$$

对于反馈生成输出的token级概率比，一个挑战在于条件不匹配。这些样本是使用反馈查询 q^{fb}生成的，但我们的优化目标以基础查询 q^{base}为条件。这种差异可能导致优化策略下精炼后的响应 $\{o^{fb}\}$具有极低的概率，进而造成高方差、潜在的梯度爆炸以及训练不稳定。因此，受 LUFFY [37], 启发，我们采用策略塑形 $f(x) = \frac{x}{x+\gamma}(0 < \gamma < 1)$。该技术会提高 $\{o^{fb}\}$中低概率token的权重。这种机制鼓励模型从罕见但正确的轨迹中学习有价值的知识，否则这些轨迹可能会被忽视。标准比率 $c_i(\theta)$和整形比率$f(c_j^{fb}(\theta))$定义如下：

$$c_i(\theta) = \frac{\pi_\theta(o_i|q^{base})}{\pi_{old}(o_i|q^{base})}, f(c_j^{fb}(\theta)) = \frac{\pi_\theta(o_j^{fd}|q^{base})}{\pi_\theta(o_j^{fd}|q^{base}) + \gamma}. \tag{9}$$

算法 1 结合反馈的 GRPO 采样更新

输入：

$\{q_i^{base}\}_{i=1}^{n}$：基础输入

s：分数阈值

k: 精炼次数

输出：

$\{q_i^{final}\}_{i=1}^{n+k}$：最终查询

$\{o_i^{final}\}_{i=1}^{n+k}$：最终推演响应

//步骤***1***：初始推演

$\{o_i\}_{i=1}^{n} = \{\text{VLA模型}(q_i^{base})\}_{i=1}^{n}$

$\{r_{traj,i}\}_{i=1}^{n} = \{r_{traj}(o_i)\}_{i=1}^{n}$

$\{f_i^{teacher}\}_{i=1}^{n} = \{\text{教师模型}(q_{base}, o_i, o_{gt})\}_{i=1}^{n}$

$$\{q_i^{fb}\}_{i=1}^{n} = \begin{cases} < q_i^{base}\ o_i, f^{\ rule} >, & \text{如果 } r_{traj}(o_i) \geq s \\ < q_i^{base}, o_i, f_i^{teacher} >, & \text{如果 } r_{traj}(o_i) < s \end{cases}$$

//步骤***2***：带反馈的推演

$\{o_i^{fb}\}_{i=1}^{n} = \{\text{VLA模型}(q_i^{fb})\}_{i=1}^{n}$

$\{r_{traj,\ i}^{fb}\}_{i=1}^{n} = \{r_{traj}(o_i^{fb})\}_{i=1}^{n}$

//步骤***3***：最终查询与输出构建

$r_{\max} \leftarrow \text{最大值}_i(\{r_{traj,i}\}_{i=1}^{n})$

$\{o_j^{fb}\}_{j=1}^{k} = \text{选择}k,\ \{o_m^{fb} \mid r_{traj,m}^{fb} > r_{\max}\})$

$\{q_i^{final}\}_{i=1}^{n+1} = \{q_i^{base}\}_{i=1}^{n} + \{q_j^{base}\}_{j=1}^{k}$

$\{o_i^{final}\}_{i=1}^{n+1} = \{o_i\}_{i=1}^{n} + \{o_j^{fb}\}_{j=1}^{k}$

返回 $\{q_i^{final}\}_{i=1}^{n+k}$，$\{o_i^{final}\}_{i=1}^{n+k}$

4. 实验

4.1. 实现细节

数据集。我们在 NAVSIM 上进行了全面的实验和评估 [7], 。NAVSIM 是基于 OpenScene 构建的面向规划的自动驾驶数据集。除了从 NAVSIM 收集的推理数据外，我们还使用了若干开源数据集，详见第3.2节。

指标。我们从自动驾驶的两个不同方面评估了该方法的性能：**高层规划**和**轨迹预测**。

对于高层规划评估，我们使用高层规划准确率，该指标严格要求完整的元动作（包括纵向速度和横向路径）与真实值完全一致。这里的真实值由GT轨迹生成，详细信息见附录8.1。

对于NAVSIM基准上的轨迹预测评估，我们使用预测驾驶员模型评分（PDMS）作为**NAVSIMv1** [7] 的闭环规划指标，并使用扩展预测驾驶员模型评分（EPDMS）作为**NAVSIMv2** [2] 的闭环规划指标。

训练细节。我们使用 InternVL3-8B [42]作为基座模型，并分三个阶段进行训练。首先，在大规模驾驶知识数据集上进行预训练。其次，在混合数据集上对模型进行微调，该数据集由精选的 Navsim 规划数据集（带有思维链标注）和第 3 节中的反馈数据集组成。第三，使用 32 块 NVIDIA H20 GPU 进行强化学习。我们使用 Qwen3-VL-32B [1]作为教师模型。强化学习的关键参数包括每批次 8 次推演、一个阈值 $s = 0.8$、一个策略塑形参数$\gamma = 0.1$，以及 $k = 1$ 精炼回复。更多细节和超参数见附录 8.1。

表1. 在 NAVSIMv1 上使用 PDMS 与最先进方法进行比较。

方法	图像	激光雷达	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
恒定速度			68.0	57.8	50.0	100	19.4	20.6
自车状态多层感知机			93.0	77.3	83.6	100	62.8	65.6
UniAD [11]	✓		97.8	91.9	92.9	100	78.8	83.4
TransFuser [6]	✓	✓	97.7	92.8	92.8	100	84.0	84.0
DiffusionDrive [22]	✓	✓	98.2	96.2	94.7	100	82.2	88.1
WoTE [18]	✓	✓	98.5	96.8	94.9	99.9	81.9	88.3
Hydra-NeXt [21]	✓		98.1	97.7	94.6	100	81.8	88.6
AutoVLA-3B [41]	✓		98.4	95.6	98.0	100	81.9	89.1
DriveVLA-W0-3B [36]	✓		98.7	**99.1**	95.3	99.3	83.3	90.3
GoalFlow [36]	✓	✓	98.4	98.3	94.6	100	85.0	90.3
InternVL3-8B-SFT	✓		98.5	95.5	95.3	100	81.2	87.4
InternVL3-8B-RL	✓		98.5	96.7	95.4	100	83.2	89.0
ELF-VLA-8B（我们的方法）	✓		**98.9**	98.1	**96.0**	**100**	**85.3**	**91.0**

表2. 使用 EPDMS 在 NAVSIMv2 上与最先进方法进行比较。

方法	NC ↑	DAC ↑	DDC ↑	TLC ↑	EP ↑	TTC ↑	LK ↑	HC ↑	EC ↑	EPDMS ↑
HydraMDP++ [20]	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprem [38]	97.5	96.5	99.4	99.6	88.4	96.6	95.5	**98.3**	77.0	83.1
Recogdrive-8B [19]	98.3	95.2	**99.5**	**99.8**	87.1	97.5	96.6	**98.3**	86.5	83.6
DiffusionDrive [22]	98.2	95.9	99.4	**99.8**	87.5	97.3	96.8	**98.3**	**87.7**	84.5
DriveVLA-W0-3B [16]	98.5	**99.1**	98.0	99.7	86.4	98.1	93.2	97.9	58.9	86.1
ELF-VLA-8B（我们的方法）	**98.9**	98.1	99.4	**99.8**	**88.5**	**98.4**	**96.9**	**98.3**	87.2	**87.1**

4.2. 性能比较

Navsim 基准。 表 1 展示了 ELF-VLA 与当前领先方法在 NAVSIMv1基准测试上的性能比较。在纯视觉设置下，ELF-VLA 的 PDMS 达到 91.0，创下新的最先进水平（SOTA）。这一结果比此前表现最佳的纯视觉方法 DriveVLA 高出 0.7 PDMS。此外，ELF-VLA 的表现优于仅SFT（InternVL3-8B-SFT）和传统强化学习（InternVL3-8B-RL）基线方法，分别高出 3.6 和 2.0 PDMS。在 NAVSIMv2基准测试（表2）上，ELF-VLA 延续了强劲表现，以 87.1 EPDMS 创下新的最先进水平。这一分数比此前表现最佳的 DriveVLA-W0 高出 1.0 EPDMS。这些发现表明，与传统强化学习方法相比，我们的方法 ELF-VLA 显著增强了模型的驾驶能力，尤其是在应对具有挑战性的驾驶场景时。此外，该方法在两个基准测试上的出色表现证实，ELF-VLA 并非仅仅过拟合 PDMS 指标；相反，它在更具区分性且更全面的 EPDMS 指标上同样表现出色，展现出稳健的泛化能力。

定量评估。我们将 ELF-VLA 的性能与若干经过精心设计的消融模3型进行比较（表

表 3. ELF-VLA 与传统 GRPO 及其他反馈策略的性能比较。

方法	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
SFT	98.5	95.5	95.3	100	81.3	87.4
GRPO	98.5	96.7	95.4	100	83.2	89.0
GT-GRPO	98.1	97.1	93.5	100	85.2	89.2
Rule-GRPO	98.3	97.3	94.8	100	84.5	89.6
ELF-VLA	**98.9**	**98.1**	**96.0**	**100**	**85.3**	**91.0**

型的详细定义见附录8.1）：

- **监督微调（基线）：** 仅通过监督微调训练的基座模型。
- **GRPO：** 使用传统的 GRPO 算法对监督微调模型进一步微调。
- **GT-GRPO：** 在通过直接添加真实值（GT）轨迹扩充的响应集上进行微调的监督微调模型。
- **Rule-GRPO：** 在通过新响应扩充的响应集上进行微调的监督微调模型，这些新响应是根据预定义规则的反馈重新生成的。
- **ELF-VLA：** 在通过教师模型提供的结构化反馈重新生成新颖、优化后的响应，并将其加入响应集扩充后，对该响应集进行微调得到的监督微调模型。

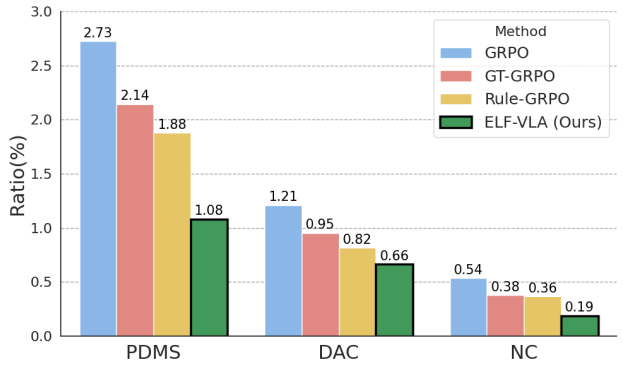


图 4. **强化学习训练阶段测得的完全失败样本比例**，适用于 GRPO、GT-GRPO、Rule-GRPO 和 ELF-VLA。完全失败是指某个样本的所有推演在某项特定指标上均失败（分别为 PDMS 低于 s、NC 为 0 和 DAC 为 0）。

值得注意的是，ELF-VLA取得了最佳的整体表现。我们的方法比传统GRPO方法高出2.0 PDMS。这表明，通过引入结构化反馈并重新生成更优且符合分布的轨迹，我们的方法有助于模型解决长期存在的失败问题。此外，ELF-VLA分别以1.8和1.4 PDMS的优势超越GT-GRPO和Rule-GRPO。这凸显了这两种基线方法各自的明显局限。对于GT-GRPO，GT轨迹与原始视觉-语言-动作模型生成的响应存在显著的分布偏移。这些GT响应的似然较低，导致优化困难。对于Rule-GRPO，预定义规则提供的反馈对模型影响有限。这一过程类似于简单的自我优化，缺乏细粒度指导，导致模型无法从这类简单反馈中学会有效的轨迹修正。相比之下，ELF-VLA利用教师模型丰富的通用知识，对原始响应进行深入的结构化分析。视觉-语言-动作模型接收这些全面反馈后，便能够从失败中学习并优化轨迹。这一过程最终生成更优、也更易于优化的优化后轨迹。

完全失败率分析。 如图所示，我们分析了这些模型在强化学习训练阶段的失败率。4。具体而言，我们测量了所有展开的轨迹在关键指标 PDMS、DAC 和 NC 上同时失败的样本比例。如图所示，尽管 GT-GRPO 和 Rule-GRPO 等中间策略有助于降低失败率，但 ELF-VLA 在所有指标上都表现出最显著的改进。ELF-VLA 将完全失败的 PDMS 比率从 GRPO 的 2.73% 降至仅 1.08%，NC 和 DAC 也取得了同样显著的降幅。这一结果进一步验证了我们的方法能够使模型从错误中学习，解决持续存在的失败问题，最终提升整体驾驶安全性和鲁棒性。

表 4. NAVSIM 上高层规划的比较。

方法	速度准确率↑	路径准确率↑	准确率↑
Qwen2.5-VL-7B	37.8	61.3	19.1
InternVL3-8B	40.9	58.7	20.1
Qwen2.5-VL-32B	46.6	55.3	27.6
Qwen2.5-VL-72B	49.4	62.6	28.7
SFT	84.2	90.7	79.2
GRPO	84.3	90.8	79.3
GT-GRPO	83.5	90.5	78.4
Rule-GRPO	84.5	91.2	79.5
ELF-VLA	**85.8**	**92.5**	**80.3**

表 5. 强化学习中训练数据数量的消融研究。 †: 随机采样。*: 根据第 3.3 节筛选。

编号	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDM$↑
85k	98.5	96.8	95.3	100	83.4	89.1
24k†	98.4	96.8	95.2	100	83.1	88.9
24k*	**98.9**	**98.1**	**96.0**	**100**	**85.3**	**91.0**

高层规划评估。 如表所示。4，我们的结果凸显了 ELF-VLA 在高层规划方面的明显优势。ELF-VLA 在纵向速度准确率和横向路径准确率两项指标上均取得最佳结果，整体规划准确率达到 80.3%，比传统 GRPO 高出 1.0%。此外，与开源模型相比，ELF-VLA 的准确率比规模大得多的 Qwen2.5-VL-72B 高出 51.6%。这一提升得益于教师模型提供了更精细的元动作，并由视觉-语言-动作模型学习并内化这些动作。这表明 ELF-VLA 能够从失败案例中学习，以改进其高层规划。

4.3. 消融研究

GRPO 中训练数据的影响。如表 5 所示，我们研究了训练数据量及其组成对强化学习的影响。使用完整的 85k数据集（PDMS 为 89.1）或随机采样的 24k 子集（PDMS 为 88.9）都只能取得次优结果。相比之下，我们依据第 3.3 节筛选的24k数据集（24k*）取得了最佳表现，PDMS 达到 91.0。这表明，完整的85k数据集由大量简单场景构成，所提供的学习信号有限，从而削弱了整体梯度信号，并导致策略更新效率低下，主要集中于已经掌握的场景。我们的筛选策略有效提炼出最有价值的数据。结合反馈机制，这些数据可用于针对这些复杂场景进行训练，最终提升模型性能和训练效率。

关于带反馈的 GRPO。 表。 6 分析了我们的反馈组件。我们首先改变精炼响应的数量， k。最佳性能（91.0 PDMS）在 $k = 1$时达到。增加 k 会导致性能下降，降至

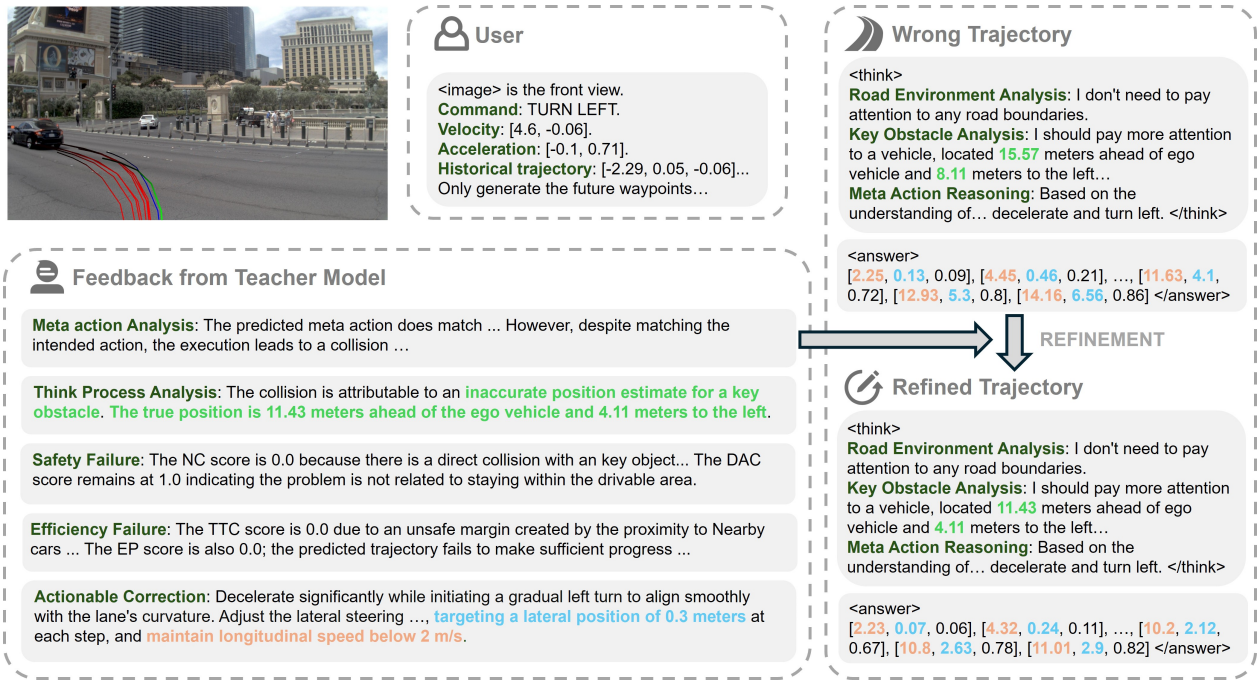


图 5. **ELF-VLA 在 NAVSIM数据集上的轨迹优化过程可视化**。 初始错误轨迹（红色）、真实值（绿色）以及最终优化后的轨迹（蓝色）的可视化。教师模型生成的反馈引导错误轨迹优化为优化后的轨迹。反馈中的彩色文字详细说明了已进行的具体优化。

表 6. 对精炼响应数量 k 及策略塑形（PS）使用情况的消融研究。

k	PS	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
4	✓	98.5	96.7	95.4	100	83.3	89.0
2	✓	98.2	97.5	94.3	100	84.9	89.7
1	✗	98.5	97.0	94.9	100	83.9	89.3
1	✓	**98.9**	**98.1**	**96.0**	**100**	**85.3**	**91.0**

89.0（在 $k = 4$ 时）。这表明，虽然单次针对性的精炼很有效，但多次基于反馈的响应可能会分散策略的注意力。我们还评估了策略塑形（PS）。移除 PS（在 $k = 1$ 时）会使 PDMS 显著下降 1.7%，从 91.0 降至 89.3。这证实了 PS 对于防止训练崩溃和格式错误至关重要，从而确保模型能够正确地从高优势、低概率的精炼轨迹中学习。

4.4. 轨迹优化过程可视化 s

图 5 展示了一个定性示例，其中 ELF-VLA 在复杂的左转场景中纠正了一条错误轨迹。初始错误轨迹（红色曲线）可能导致碰撞，原因是对一个关键障碍物的位置估计严重错误（预测位置：前方 15.57 米、左侧 8.11 米）。教师模型提供了结构化反馈，准确识别出这一“思考过程”错误，并估算出更准确的位置（前方 11.43 米、左侧 4.11 米）。同时，模型还提供了可执行的修正，例如调整目标横向位置和纵向速度。根据这些反馈，模型生成了优化后的轨迹（蓝色曲线）。精炼方案中相应的“关键障碍物分析”反映了修正后的感知结果，使智能体能够规划出更安全的轨迹，成功避开障碍物。更多结果见附录 8.2。

5. 结论

本文提出了 ELF-VLA，这是一种从失败中显式学习的框架。我们的方法为视觉-语言-动作策略引入了一个强大的教师模型，该模型会生成结构化诊断报告，并在每次发生失败时识别其根本失败模式。随后，该策略利用这种明确、类人的反馈，合成经过纠正且具有高奖励的轨迹。通过将这些纠正后的样本重新注入强化学习训练批次，ELF-VLA 提供针对性梯度，使策略能够解决具有挑战性的场景，而无引导探索很少能够克服这些场景。

该方法的主要局限在于依赖外部教师模型，因此学生模型的性能在本质上受限于教师模型的分析能力。此外，所有实验均在 NAVSIM 基准这一非反应式仿真环境中进行。未来的工作将探索不同教师模型的作用，并在更多样化的数据集上进行闭环评估。

参考文献

[1] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-vl 技术报告。arXiv 预印本 arXiv:2502.13923，2025。5[2] Wei Cao、Marcel Hallgarten、Tianyu Li、Daniel Dauner、Xunjiang Gu、Caojun Wang、Yakov Miron、Marco Aiello、Hongyang Li、Igor Gilitschenski 等。面向自动驾驶的伪仿真。arXiv 预印本 arXiv:2506 04218，2025 5，3 . . [3]Angelica Chen、J´er´emy Scheurer、Jon Ander Campos、Tomasz Korbak、Jun Shern Chan、Samuel R Bowman、Kyunghyun Cho 和 Ethan Perez。基于自然语言反馈进行学习。《机器学习研究汇刊》，2024。2[4] Li Chen、Penghao Wu、Kashyap Chitta、Bernhard Jaeger、Andreas Geiger 和 Hongyang Li。端到端自动驾驶：挑战与前沿。《IEEE模式分析与机器智能汇刊》，2024。1[5]Haohan Chi、Huan-ang Gao、Ziming Liu、Jianing Liu、Chenyu Liu、Jinwei Li、Kaisen Yang、Yangcheng Yu、Zeda Wang、Wenyi Li 等。Impromptu VLA：面向驾驶视觉-语言-动作模型的开放权重与开放数据。arXiv 预印本 arXiv:2505.23757，2025。3，1[6] Kashyap Chitta、Aditya Prakash、Bernhard Jaeger、Zehao Yu、Katrin Renz 和 Andreas Geiger。TransFuser：采用基于Transformer的传感器融合进行自动驾驶模仿。《IEEE模式分析与机器智能汇刊》，45(11)：12878–12895，2022。6[7] Daniel Dauner、Marcel Hallgarten、Tianyu Li、Xinshuo Weng、Zhiyu Huang、Zetong Yang、Hongyang Li、Igor Gilitschenski、Boris Ivanovic、Marco Pavone 等。NAVSIM：数据驱动的非反应式自动驾驶车辆仿真与基准测试。《神经信息处理系统进展》，37：28706–28719，2024。2，4，5，3[8] Wenhao Ding、Chejian Xu、Mansur Arief、Haohong Lin、Bo Li 和 Ding Zhao。安全关键驾驶场景生成综述——方法论视角。《IEEE智能交通系统汇刊》，24(7)：6971–6988，2023。1[9] Xinpeng Ding、Jianhua Han、Hang Xu、Xiaodan Liang、Wei Zhang 和 Xiaomeng Li。通过注入鸟瞰视角的多模态大模型实现整体自动驾驶理解。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第13668–13677页，2024。3，1[10] Haoyu Fu、Diankun Zhang、Zongchuang Zhao、Jianfeng Cui、Dingkang Liang、Chong Zhang、Dingyuan Zhang、Hongwei Xie、Bing Wang 和 Xiang Bai。Orion：通过视觉-语言指令生成动作的整体端到端自动驾驶框架。arXiv 预印本 arXiv:2503 19755，2025 2 . . [11] Yihan Hu、Jiazhi Yang、Li Chen、Keyu Li、Chonghao Sima、Xizhou Zhu、Siqi Chai、Senyao Du、Tianwei Lin、Wenhai Wang 等。面向规划的自动驾驶。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第17853–17862页，2023。1，6

[12] Jyh-Jing Hwang、Runsheng Xu、Hubert Lin、Wei-Chih Hung、Jingwei Ji、Kristy Choi、Di Huang、Tong He、Paul Covington、Benjamin Sapp 等。Emma：面向自动驾驶的端到端多模态模型。arXiv 预印本 arXiv:2410 23262，2024 2。[13] Bo Jiang、Shaoyu Chen、Bencheng Liao、Xingyu Zhang、Wei Yin、Qian Zhang、Chang Huang、Wenyu Liu 和 Xing- gang Wang。Senna：连接大型视觉语言模型与端到端自动驾驶。arXiv 预印本 arXiv:2410.22313，2024。2[14] Bo Jiang、Shaoyu Chen、Qian Zhang、Wenyu Liu 和 Xing- gang Wang。Alphadrive：通过强化学习与推理，释放视觉语言模型在自动驾驶中的潜能。arXiv 预印本 arXiv:2503.07608，2025。2[15] Sicong Jiang、Zilin Huang、Kangan Qian、Ziang Luo、Tianze Zhu、Yang Zhong、Yihong Tang、Menglin Kong、Yunlong Wang、Siwen Jiao 等。自动驾驶视觉-语言-动作模型综述。arXiv 预印本 arXiv:2506.24044，2025。1[16] Yingyan Li、Shuyao Shang、Weisong Liu、Bing Zhan、Haochen Wang、Yuqi Wang、Yuntao Chen、Xiaoman Wang、Yasong An、Chufeng Tang 等。DriveVLA-W0：世界模型放大自动驾驶中的数据缩放定律。arXiv 预印本 arXiv:2510.12796，2025。6[17] Yue Li、Meng Tian、Dechang Zhu、Jiangtong Zhu、Zhenyu Lin、Zhiwei Xiong 和 Xinhai Zhao。Drive-r1：通过强化学习，在视觉语言模型中衔接自动驾驶的推理与规划。arXiv 预印本 arXiv:2506 18234，2025 1，2。[18] Yingyan Li、Yuqi Wang、Yang Liu、Jiawei He、Lue Fan 和 Zhaoxiang Zhang。借助鸟瞰图世界模型进行在线轨迹评估的端到端驾驶。arXiv 预印本 arXiv:2504.01941，2025。6[19] Yongkang Li、Kaixin Xiong、Xiangyu Guo、Fang Li、Sixu Yan、Gangwei Xu、Lijun Zhou、Long Chen、Haiyang Sun、Bing Wang 等。ReCogDrive：面向端到端自动驾驶的强化认知框架。arXiv 预印本 arXiv:2506.08052，2025。2，6，1[20] Zhenxin Li、Kailin Li、Shihao Wang、Shiyi Lan、Zhiding Yu、Yishen Ji、Zhiqi Li、Ziyue Zhu、Jan Kautz、Zuxuan Wu 等。Hydra-mdp：采用多目标 Hydra 蒸馏的端到端多模态规划。arXiv 预印本 arXiv:2406 06978，2024 6。[21] Zhenxin Li、Shihao Wang、Shiyi Lan、Zhiding Yu、Zux- uan Wu 和 Jose M Alvarez。Hydra-NeXt：通过开环训练实现稳健的闭环驾驶。arXiv 预印本 arXiv:2503.12030，2025。6[22] Bencheng Liao、Shaoyu Chen、Haoran Yin、Bo Jiang、Cheng Wang、Sixu Yan、Xinbang Zhang、Xiangyu Li、Ying Zhang、Qian Zhang 等。DiffusionDrive：用于端到端自动驾驶的截断扩散模型。载于《计算机视觉与模式识别会议论文集》，第 12037–12047 页，2025。6[23] Henry X Liu 和 Shuo Feng。自动驾驶汽车的稀有性诅咒。《自然-通讯》，15(1):4808，2024。1[24] Xueyi Liu、Zuodong Zhong、Yuxin Guo、Yun-Fu Liu、Zhiguo Su、Qichao Zhang、Junli Wang、Yinfeng Gao、Yupeng Zheng,

Qiao Lin 等。ReasonPlan：面向闭环自动驾驶的统一场景预测与决策推理。arXiv 预印本 arXiv:2505.20024，2025。2[25] Yuechen Luo、Fang Li、Shaoqing Xu、Zhiyi Lai、Lei Yang、Qimao Chen、Ziang Luo、Zixun Xie、Shengyin Jiang、Ji- axin Liu 等。Adathinkdrive：通过强化学习实现自动驾驶的自适应思考。arXiv 预印本 arXiv:2509.13769，2025。1，2[26] Ziang Luo、Kangan Qian、Jiahua Wang、Yuechen Luo、Jinyu Miao、Zheng Fu、Yunlong Wang、Sicong Jiang、Zilin Huang、Yifei Hu 等。Mtrdrive：面向极端场景下稳健自动驾驶的记忆-工具协同推理。arXiv 预印本 arXiv:2509.20843，2025。2[27] Lu Ma、Hao Liang、Meiyi Qiang、Lexiang Tang、Xiaochen Ma、Zhen Hao Wong、Junbo Niu、Chengyu Shen、Runming He、Yanhao Li 等。学习强化学习做不到的事：针对最难问题的交错在线微调。arXiv 预印本 arXiv:2506.07527，2025。2[28] Ana-Maria Marcu、Long Chen、Jan H¨unermann、Alice Karn- sund、Benoit Hanotte、Prajwal Chidananda、Saurabh Nair、Vijay Badrinarayanan、Alex Kendall、Jamie Shotton 等。LingoQA：面向自动驾驶的视觉问答。载于《欧洲计算机视觉会议》，第 252–269 页。Springer，2024。2，3，1[29] Tianwen Qian、Jingjing Chen、Linhai Zhuo、Yang Jiao 和 Yu-Gang Jiang。NuScenes-QA：面向自动驾驶场景的多模态视觉问答基准。载于 AAAI 人工智能会议论文集，第 4542–4550 页，2024。3，1[30] Zhijie Qiao、Haowei Li、Zhong Cao 和 Henry X Liu。LightEMMA：轻量级端到端自动驾驶多模态模型。arXiv 预印本 arXiv:2505 00284，2025 2。[31] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu 等。DeepSeekMath：推动开放语言模型数学推理能力的极限。arXiv 预印本 arXiv:2402.03300，2024。2[32] Chonghao Sima、Katrin Renz、Kashyap Chitta、Li Chen、Hanxue Zhang、Chengen Xie、Jens Beißwenger、Ping Luo、Andreas Geiger 和 Hongyang Li。DriveLM：基于图视觉问答的驾驶。载于《欧洲计算机视觉会议》，第 256–274 页。Springer，2024。3，1[33] Xiaoyu Tian、Junru Gu、Bailin Li、Yicheng Liu、Yang Wang、Zhiyong Zhao、Kun Zhan、Peng Jia、Xianpeng Lang 和 Hang Zhao。DriveVLM：自动驾驶与大型视觉语言模型的融合。arXiv 预印本 arXiv:2402.12289，2024。2[34] Shihao Wang、Zhiding Yu、Xiaohui Jiang、Shiyi Lan、Min Shi、Nadine Chang、Jan Kautz、Ying Li 和 Jose M Al- varez。OmniDrive：一个结合反事实推理的整体式自动驾驶视觉语言数据集。载于《计算机视觉与模式识别会议论文集》，第 22442–22452 页，2025。2，3，1[35] Shuo Xing、Chengyuan Qian、Yuping Wang、Hongyuan Hua、Kexin Tian、Yang Zhou 和 Zhengzhong Tu。OpenEMMA：用于端到端自动驾驶的开源多模态模型驾驶。在《冬季计算机视觉应用会议论文集》中，第 1001–1009 页，2025 年。2[36] Zebin Xing、Xingyu Zhang、Yang Hu、Bo Jiang、Tong He、Qian Zhang、Xiaoxiao Long 和 Wei Yin。Goalflow：面向端到端自动驾驶多模态轨迹生成的目标驱动流匹配。载于《计算机视觉与模式识别会议论文集》，第 1602–1611 页，2025 年。6[37] Jianhao Yan、Yafu Li、Zican Hu、Zhi Wang、Ganqu Cui、Xiaoye Qu、Yu Cheng 和 Yue Zhang。学习在离策略引导下进行推理。arXiv 预印本 arXiv:2504 14945，2025 年。2、5。[38] Wenhao Yao、Zhenxin Li、Shiyi Lan、Zi Wang、Xinglong Sun、Jose M Alvarez 和 Zuxuan Wu。Drivesuprim：迈向端到端规划中的精确轨迹选择。arXiv 预印本 arXiv:2506.06659，2025 年。6[39] Xiaoying Zhang、Hao Sun、Yipeng Zhang、Kaituo Feng、Chaochao Lu、Chao Yang 和 Helen Meng。Critique-GRPO：通过自然语言和数值反馈推进大语言模型推理。arXiv 预印本 arXiv:2506.03106，2025 年。2、3[40] Rui Zhao、Qirui Yuan、Jinyu Li、Haofeng Hu、Yun Li、Chengyuan Zheng 和 Fei Gao。Sce2DriveX：用于从场景到驾驶学习的通用多模态大语言模型框架。arXiv 预印本 arXiv:2502.14917，2025 年。2[41] Zewei Zhou、Tianhui Cai、Seth Z Zhao、Yun Zhang、Zhiyu Huang、Bolei Zhou 和 Jiaqi Ma。AutoVLA：一种结合自适应推理与强化微调、用于端到端自动驾驶的视觉-语言-动作模型。arXiv 预印本 arXiv:2506.13757，2025 年。1、2、6[42] Jinguo Zhu、Weiyun Wang、Zhe Chen、Zhaoyang Liu、Shenglong Ye、Lixin Gu、Hao Tian、Yuchen Duan、Weijie Su、Jie Shao 等。InternVL3：探索开源多模态模型的先进训练与测试时方案。arXiv 预印本 arXiv:2504.10479，2025 年。5、2

通过从失败中显式学习，释放视觉-语言-动作在自动驾驶中的潜力

补充材料

在本补充材料中，我们详细介绍数据构建（第 6 节）和奖励设计（第 7 节）的具体实现细节，并展示其他实验结果，包括训练流程的消融研究以及轨迹优化过程的可视化（第8 节）。

6. 数据构建细节

6.1. 预训练数据细节

我们汇集了多种开源驾驶问答数据集，随后是 ReCogDrive [19], , 包括 Driv- eLM [32], LingoQA [28], I mpromptuVLA [5], NuScenes- QA [29], NuInstruct [9], OminiDrive [34]。

6.2. 监督微调数据集详情

思维链构建。 遵循 [25], 中概述的数据构建范式，我们通过系统地综合未来轨迹数据与场景级语义，生成高质量的思维链监督。对于动态实体，我们基于时空关系筛选并识别与自车主动交互的智能体。这些智能体分为三类：CIPO-1 指当前车道上的前车，主要施加纵向约束；CIPO-2 包括相邻车道上根据车道几何和相对速度推断出可能并道或切入的车辆；运动交互则涵盖预测轨迹与自车在空间上相交、表明存在较高碰撞风险的实体。对于静态元素，我们利用 NAVSIM地图重建车道拓扑并提取关键边界特征，例如道路曲率和车道中心线，以严格界定可行驶区域。此外，我们采用 Qwen3-VL-32B 对环境属性进行细粒度描述，将视觉线索转化为文字。通过整合这些动态交互、静态约束和语义描述，我们整理出各类场景，构建逐步推理标注，以逻辑连贯的方式衔接感知、预测与规划。思维链详情见图 10 和图11。**基础输入与反馈输入构建。** 对于基础输入，我们将车辆当前的速度、加速度和历史轨迹作为提示，帮助模型更好地预测其行驶路径，如图 6所示。对于反馈输入，我们使用 Qwen3VL-32B，通过向其提供错误轨迹（o_w）、真实值（o_{gt}）、详细的 Navsim指标分数，以及任务要求。这些明确的输入使教师模型能够诊断失败原因并生成详细的纠正指导。详情见图7。

监督微调数据构建。 为使模型具备初步的轨迹优化能力，我们在推理阶段从生成的轨迹中随机采样 4k 个正确响应和 4k 个错误响应，构建反馈数据集。这些响应根据其相对于 PDMS阈值 s 的评估结果，与相应的反馈配对。具体而言，得分高于 s 的响应轨迹被指定为“正确”（o_c），并附带基于规则的反馈（f^{rule}），以强化成功行为。相反，得分低于 s 的响应轨迹被标记为“错误”（o_w），并与教师引导的反馈（$f^{teacher}$）关联；该反馈通过 Qwen3-VL-32B 的推理提供纠正性指导。

7. 方法详情

7.1. 强化学习中的详细奖励设计

PDMS奖励。 我们将模型预测的轨迹输入 Navsim 模拟器进行评估。模拟器根据多项关键指标评估每条轨迹的驾驶质量，包括无责任碰撞、可行驶区域合规性、自车进展、碰撞时间和驾驶舒适度。随后，模拟器生成一个综合得分，即预测驾驶员模型评分（*PDMS*），作为该组件的奖励信号。该评估指标*PDMS*用于轨迹奖励 r_{traj}，其取值范围为 0 到 1 的连续值。

格式奖励。 奖励项 r_{fmt} 旨在确保结构有效性，总分为 1.0 分，平均分配给两项要求。若输出正确包含两个独立部分：<think>...</think>和<answer>...</answer>，则可获得前一半（0.5 分）。剩余的 0.5 分用于验证预测轨迹点的语法，确保输出格式正确且可由机器解析。

目标奖励。 为使预测终点与真实值精确对齐，我们采用分段奖励函数 r_{goal}，并通过 L1距离进行计算。具体形式定义如下：

System: Your task is to first output the planning reasoning process within the <think> </think> tags ... Then, based strictly on this predicted behavior in the <think> section, generate the corresponding 8-step trajectory in the <answer> </answer> tags... If you fail to provide a response, it may lead to potential harm or injury.
Generate the planning reasoning in the <think></think>: First, examine whether the ego vehicle is close to any road boundaries... Then, describe only the agents that may affect the ego vehicle's driving... Based on this analysis, reason about the ego vehicle's next driving behavior. You must predict the ego vehicle's future driving behavior.

Prompt: Command: … Velocity: … Acceleration: … Historical trajectory: … Output the reasoning in <think></think> and the predicted trajectory in <answer></answer>… Output exactly 8 waypoints in the format [x, y, heading]...

图 6. 视觉-语言-动作模型（基础输入）的提示设计。

System: You are an expert autonomous driving … The PDMS is … Output Requirements: Provide your critique inside <feedback> tags.
1. Meta action analysis: State if … **2. Think process analysis**: Evaluate… **3. Safety failure**: base on NC/DAC, state the reason…
4. Efficiency failure: base on TTC/EP, quantify the failure… **5. Actionable correction**: Provide a consolidated correction plan…

Prompt: Command: … Velocity: … Acceleration: … Fault Trajectory: … Traj Score: … GT Trajectory: … Feedback Format: …

图 7. 教师模型的提示设计（反馈输入）。

$$r_{goal} = \begin{cases} 1 & \text{if } 0 < dis < 2 \\ 0.8 & \text{if } 2 \leq dis < 4 \\ 0.6 & \text{if } 4 \leq dis < 6 \\ 0.4 & \text{if } 6 \leq dis < 10 \\ 0.2 & \text{if } 10 \leq dis < 15 \\ 0 & \text{if } dis > 15 \end{cases} \tag{10}$$

8. 实验细节

8.1. 实现细节

模型架构。我们使用 InternVL3-8B [42]，，这是一个视觉语言基础模型，它将参数量为 3 亿的 InternViT 视觉编码器与参数量为 70 亿的 Qwen2.5 语言模型相结合。它采用分辨率自适应视觉输入机制，根据内容动态调整特征提取的尺度与粒度来处理图像。该设计对复杂区域进行细粒度处理，对较简单区域进行粗粒度特征提取。这使得模型能够在优化计算效率的同时保持较高的视觉保真度。

训练参数与硬件配置。训练过程包括三个阶段：第一阶段在大规模、高质量的驾驶知识数据集上进行监督微调，该数据集包含多样化的指令遵循示例和场景感知标注，使用 2 个训练轮数、批次大小为 1、学习率为 1×10^{-5}、4 个梯度累积步数、0.05 的权重衰减以及 0.05 的权重比。第二阶段在 NAVSIM 规划数据集（带有思维链标注）和反馈数据集上进一步微调模型，采用 2 个训练轮数、批次大小为 2、学习率为 4×10^{-5}、2 个梯度累积步数，并保持相同的权重衰减和权重比（均为 0.05）。第三阶段使用 GRPO 在经过筛选的 Navsim 规划数据集上进行强化学习，学习率为 2×10^{-6}、批次大小为 3、16 个梯度累积步，权重衰减为 0.05，权重比为 0.05，生成 8 个样本，温度设为 1.2，迭代 2 次。此阶段使用 32 块 NVIDIA H20 GPU。此外，我们将阈值 $s = 0.8$、策略塑形参数 $\gamma = 0.1$设为 $k = 1$，并将精炼回复设为该值。所有实验均在 Ubuntu 系统下使用 PyTorch 2.5.0 和 CUDA 12.3，运行于 NVIDIA H20 GPU 上。前两个阶段分别使用 16 块 GPU 训练约 2 天和 8 小时，第三阶段使用 32 块 GPU 训练 18 小时。推理期间，为生成思维链和轨迹，我们的模型延迟为 0.1 秒（由 vLLM 加速）。详细超参数见表 7。

表 7. ELF-VLA 的训练超参数

阶段	超参数	值
预训练	训练轮数	2
	批次大小	1
	学习率	1×10^{-5}
	梯度累积步数	4
	权重衰减	0.05
	权重比	0.05
2	训练轮数	2
	批次大小	2
	学习率	4×10^{-5}
	梯度累积步数	2
	权重衰减	0.05
	权重比	0.05
3	训练轮数	3
	批次大小	2
	学习率	2×10^{-6}
	梯度累积步数	16
	权重衰减	0.05
	权重比	0.05
	生成次数	8
	迭代次数	2
	温度	1.2
	阈值(s)	0.8
	细化次数	1
	策略塑形权重(γ)	0.1

The driving score is 0.0. The score is below the threshold, and the result is wrong. Please follow the prompts to regenerate the correct trajectory.

(a) Rule-Based Feedback (Rule-GRPO)

1. Meta action Analysis: The predicted meta action in the <think> block matches the Ground Truth. However, the implementation of this action resulted in a problematic trajectory due to flawed execution rather than an incorrect action choice.

2. Think Process Analysis: Overlooking key obstacle contributes directly to the failure... The car should pay more attention to a vehicle located 18.23 meters ahead of ego vehicle and 6.31 meters to the right.

3. Safety Failure: The safety failure is due to the NC=0 score, indicating a collision. The DAC score remains at 1.0 indicating the problem is not related to staying within the drivable area.

4. Efficiency Failure: The TTC score is 0, showing an unsafe following distance, likely because the vehicle veers off its intended path. For EP, the predicted position at t=8 seconds ([16.33, -6.22]) deviates from the target ([16.27, -4.43]), indicating insufficient progress along the intended turning route due to poor lateral control.

5. Actionable Correction: Adjust the steering angle to ensure the vehicle maintains a precise path closer to the ideal trajectory during the right turn. Target a lateral deviation of less than 0.2 meters from the center of the lane throughout the turn. Maintain a consistent longitudinal speed of approximately 5 m/s. Ensure the vehicle avoids any collision by aligning correctly with the curve.

(b) Teacher-Based Feedback (ELF-VLA)

图 8。**反馈机制对比**。不同于 Rule-GRPO（a）中的二元信号，ELF-VLA（b）中基于教师的反馈提供了结构化诊断和具体、可执行的策略，以指导轨迹优化。

指标。 为了评估 NAVSIM基准中的轨迹预测，我们采用预测驾驶员模型得分（PDMS）作为 **NAVSIMv 1** [7] 的主要指标，并采用扩展预测驾驶员模型得分（EPDMS）作为**NAVSIMv2** [2]的主要闭环规划指标。

对于 NAVSIMv1，PDMS 综合五项子指标：无责碰撞（NC）、可行驶区域合规性（DAC）、碰撞时间（TTC）、舒适度（C）和自车进展（EP），以生成全面的闭环规划得分。其计算公式定义如下：

$$PDMS = NC \times DAC \times \left(\frac{5 \times EP + 5 \times TTC + 2 \times C}{12} \right), \tag{11}$$

对于 NAVSIMv2，PDMS指标包含多个组件，这些组件分为惩罚项或加权子得分。其关键指标包括无责碰撞（NC）、可行驶区域合规性（DAC）、行驶方向合规性（DDC）、交通信号灯合规性（TLC）、自车进展（EP）、碰撞时间（TTC）、车道保持（LK）、历史舒适度（HC）和扩展舒适度（EC）。其计算公式定义如下：

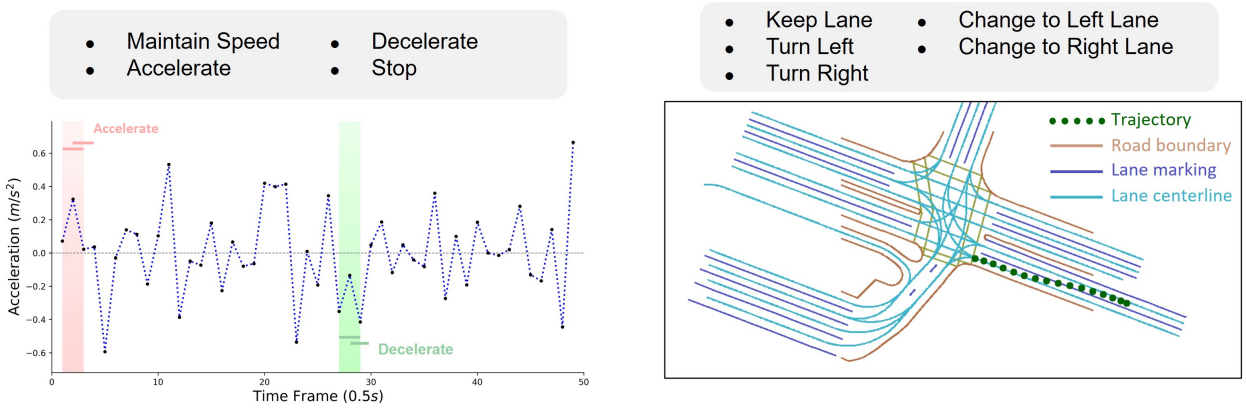
$$EPDMS = NC \times DAC \times DDC \times TLC \times \left(\frac{5EP + 2LK + 2HC + 5TTC + 2EC}{16} \right). \tag{12}$$

反馈机制比较。 如第 4.2节所述，Rule-GRPO 与 ELF-VLA 的根本区别在于反馈机制及其粒度。Rule-GRPO 依赖静态启发式规则，提供仅用于指示正确或失败的二元反馈，随后参考真实值来引导重新生成正确答案。相比之下，ELF-VLA 会生成在线实例级反馈。通过利用 PDMS 中的细粒度指标，它会分析每条错误轨迹的具体成因，提供详细推理和建设性指导，以引导纠正过程。如图 8所示，基于规则的反馈提供的约束有限，而 ELF-VLA 的反馈则能提供全面诊断，从而实现精确的轨迹优化。

高层规划准确率。 为评估模型的决策能力，我们采用高层规划准确率指标，从两个不同维度衡量精度：纵向速度与横向路径。纵向维度将车辆行为划分为四种离散状态：加速、减速、保持速度和停车。横向维度涵盖五种方向机动：保持车道、左转、右转、向左变道和向右变道。如图9所示，真实值标签直接从轨迹中导出。具体而言，纵向状态通过滑动窗口计算加速度来确定。对于横向机动，我们构建道路拓扑图，根据未来路径与车道结构的对齐关系识别具体动作。

8.2. 更多消融研究

关于训练流程。 表。 8展示了 ELF-VLA 三阶段训练流程的消融实验结果。仅使用 NAVSIM 轨迹数据进行监督微调时，模型的 PDMS 达到 85.3。在大规模驾驶问答数据集上进行预训练后，得分提升至 87.4 PDMS，增加了 2.1。进一步采用反馈 GRPO 后，性能提升至 91.0 PDMS，增加了 3.6。这些结果表明，预训练和反馈 GRPO 策略对于提升模型的理解能力与推理能力都发挥着关键作用。



(a) Longitudinal Action (b) Lateral Action

图 9. **高层动作与真实标注生成示意图。** 上方各图列出了纵向规划和横向规划的离散类别。下方各图展示了标注标准：通过滑动窗口加速度分析确定纵向状态，并根据车辆轨迹与地图拓扑之间的关系识别横向行为。

表 8. **ELF-VLA 组件消融研究。** 我们使用 NAVSIM基准评估预训练、监督微调和强化学习对驾驶性能的影响。

模型	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
SFT	98.5	93.4	95.1	100	78.8	85.3
预+监督微调	98.5	95.5	95.3	100	81.2	87.4
预+监督微调+强化学习	**98.9**	**98.1**	**96.0**	**100**	**85.3**	**91.0**

表 9. **未经过预训练的 ELF-VLA 的性能。**“w/o”表示“没有”。

方法	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
InternVL3-8B(未预训练)	97.9	93.9	**94.4**	**100**	82.8	86.9
ELF-VLA-8B(未预训练)	**98**		**94.4**	1		**90.0**

关于反馈 GRPO 的预训练。 我们进行了排除预训练数据集的消融研究（见表。 9）。即使没有预训练，ELF-VLA 的 PDMS 仍达到 90.0，超过基线（86.9 PDMS） +3.1。这证实，性能提升主要源于我们的 GRPO 设计，而非仅仅来自预训练数据。

关于反馈阈值 s。 我们研究了阈值 s 的敏感性，该阈值用作触发教师模型反馈机制的标准（如第 3.1 节所述）。具体而言，仅当模型的响应分数低于 s 时，才会调用教师模型生成优化指导。如表 10 所示，模型在 $s = 0.8$ 时达到最佳性能。较低的阈值（例如 0 和 0.5）会导致次优结果，主要是因为它们将纠正范围限制在完全失败的情况（分数为 0），忽略了那些虽然表现略差但仍需要改进的样本。将阈值设置为 $s = 0.8$ 后，我们有效扩大了优化范围，从而捕获并纠正这些次优情况。相反，将阈值激进地提高到 0.9 会导致

表 10. 反馈阈值消融研究 s。

阈值	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
0	98.4	97.1	94.3	100	84.7	89.4
0.5	98.6	97.4	95.3	100	84.9	90.1
0.9	98.4	97.3	94.9	100	84.8	89.8
0.8	**98.9**	**98.1**	**96.0**	**100**	**85.3**	**91.0**

性能下降。这表明，得分处于 $[0.8, 0.9)$ 范围内的样本已经具有足够高的质量。对这些有效响应强制进行优化不会带来任何积极收益，反而可能引入噪声，导致优化过程偏离最优策略。因此，我们的 ELF-VLA 采用 $s = 0.8$，以平衡纠正覆盖率和训练稳定性。

8.3. 精炼过程的可视化

图 10 和图11 展示了更多定性示例，证明 ELF-VLA 在复杂场景中的有效性。如图所示，我们的结构化反馈机制在精炼循环中发挥着关键作用。首先，它验证高层规划决策的正确性。其次，也是至关重要的一点，它纠正中间的思维链推理过程。这一步有效降低了错误累积的风险，确保思维链发挥其提升轨迹预测的预期作用，而非引入幻觉或级联故障。为此，我们利用教师模型（Qwen3-VL-32B）的空间推理能力，精确定位关键障碍物。此外，借助 PDMS 中的细粒度评分，我们对失效模式进行详细诊断，并将其具体分为安全失效和效率失效。通过这一全面分析，最终生成具体且可执行的纠正策略，以精炼最终输出。

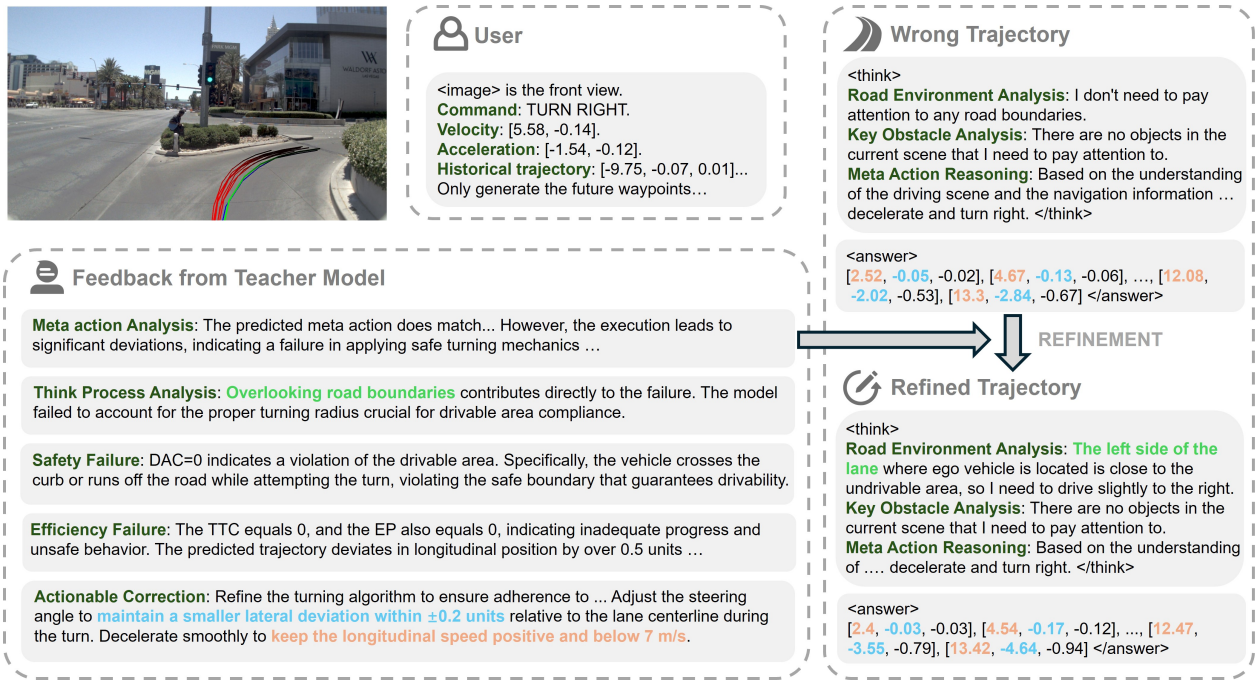


图10. **ELF-VLA在NAVSIM数据集上的轨迹优化过程可视化。** 初始错误轨迹（红色）、真实值（绿色）以及最终优化后的轨迹（蓝色）的可视化。

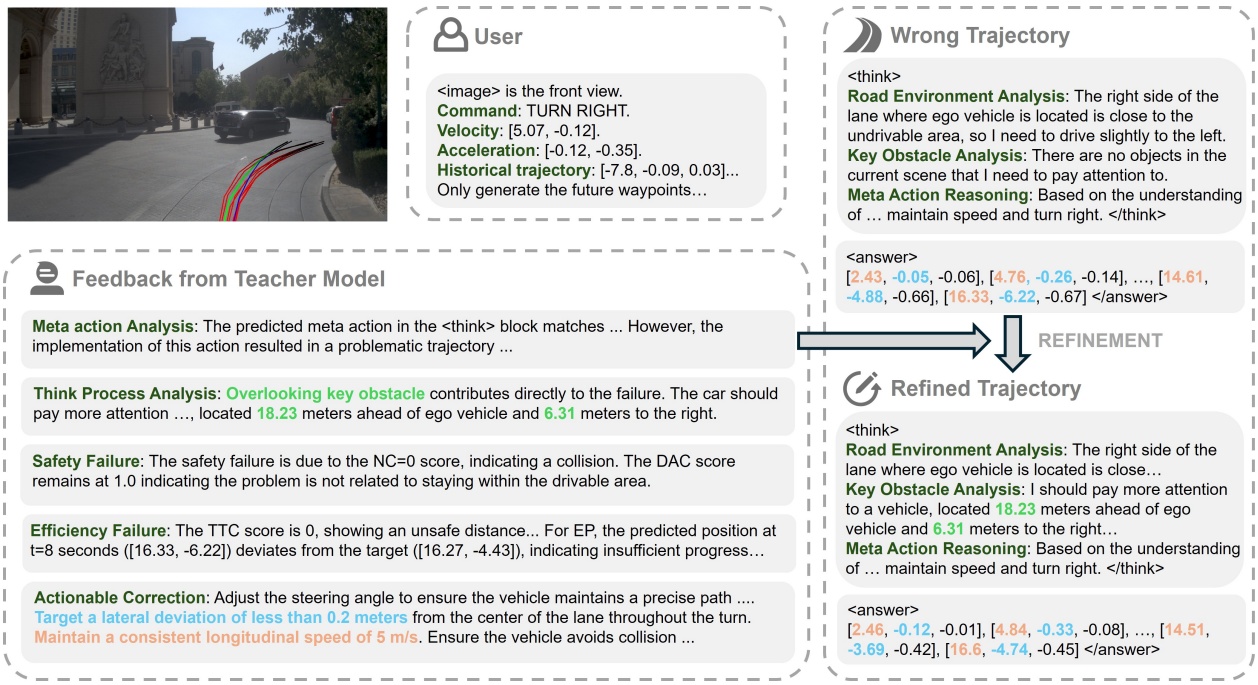


图11. **ELF-VLA在NAVSIM数据集上的轨迹优化过程可视化。** 初始错误轨迹（红色）、真实值（绿色）以及最终优化后的轨迹（蓝色）的可视化。