

提高大型视觉语言模型对现场数据的理解

Xiaomei Zhang, Hanyu Zheng, Xiangyu Zhu, *Senior Member, IEEE*, Jinghuan Wei, Junhong Zou, Zhen Lei, *Senior Member, IEEE* and Zhaoxiang Zhang

Abstract—大型视觉语言模型 (LVLMs) 在人像描述和视觉问答等综合视觉和文本理解的多个任务中表现出色。这些模型是在配有文本的大规模图像和视频数据集上训练的, 使它们能够连接视觉感知和自然语言处理。然而, 它们在科学领域的应用, 特别是在解释常用于自然科学的复杂现场数据方面, 仍然少有探索。在这项工作中, 我们介绍了 FieldLVLM, 一种新颖的框架, 旨在提高大型视觉语言模型对现场数据的理解。FieldLVLM 由两个主要组件组成: 现场感知语言生成策略和数据压缩的多模态模型调优。现场感知语言生成策略利用专门的机器学习流水线从现场数据中提取关键物理特征, 如流动分类、雷诺数和涡流模式。然后将这些信息转换为作为数据集的结构化文本描述。数据压缩的多模态模型调优集中在具有这些生成数据集的 LVLMs 上, 使用数据压缩策略来减少现场输入的复杂性, 仅保留最具信息量的值。这确保了与模型的语言解码器的兼容性, 并更有效地指导其学习。在新提出的基准数据集上的实验结果表明, FieldLVLM 在涉及科学现场数据的任务中显著优于现有方法。我们的研究结果表明, 这种方法为将大型视觉语言模型应用于科学研究开辟了新的可能性, 帮助缩小大型模型与领域专属发现之间的差距。

Index Terms—Field data, field-aware language generation strategy, data-compressed multimodal model tuning, large vision-language models

视觉和语言是人类感知和解释世界的两种基本模式。视觉提供空间和结构理解, 而语言传达语境和抽象信息。最近在大型视觉-语言模型 (LVLMs) 方面的进展使得视觉输入 (如图像、视频) 与自然语言的整合成为可能, 使这些模型能够捕捉空间关系和语义意义。该融合显著增强了它们的表现能力并实现了全面的场景理解。LVLMs 已被证明能够从数据中学习大量的深入知识。经过预训练的 LVLMs 在一系列开放世界视觉任务中表现出色, 包括图像分类 [1], [2]、目标检测 [3], [4] 和语义分割 [5], [6]。然而, LVLMs 在领域科学数据上的表现效果不佳。然而, 将其应用于科学实地数据时, 其性能仍然有限。这一局限性主要来源于两个挑战。首先, 适合训练 LVLMs 的高质量、大规模科学数据集的可用性有限。其次, 实地数据往往在长度和复杂性上有所不同, 可能超过当前模型的最大输入标记限制, 使得有效的编码和推理变得困难。

在本文中, 我们提出了一个新的框架, 用于增强大型视觉语言模型对科学领域数据的理解。我们的方法整合了两个关键组件: 领域感知的语言生成策略和数据压缩的多模态模型调优。所提出的领域感知语言生成策略结合了专业模型的高准确性和大型语言模型的一致性作为我们的数据源。虽然领域专业模型在特定任务的准确性方面表现卓越, 但在不同数据类型的泛化和一致性方面表现不足。相比之下, 大型语言模型提供了广泛且一致的解释, 但在领域特定分析的精度上可能表现不佳。具体来说, 给定领域数据, 领域感知语言生成策略使用领域专业模型来生成流动分类、雷诺数和涡旋检测。然后, 将结果和原始领域数据输入到一个大型语言模型中, 以生成一致的领域语言表示用于下游学习和模型调优。

The authors are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences (Beijing 100190).

Manuscript received Jul. xx, 2025.

基于最近提出的 Qwen2.5-VL [7], 我们开发了一种数据压缩的多模态模型, 专为现场数据的独特特性而设计。为了适应大型语言模型的输入限制, 我们引入了 VQGAN [8] 来压缩现场数据。具体来说, 我们首先将速度和压力场映射到 RGB 图像的三个通道上, 以生成原始场的 256×256 表示。然后, 该图像由 VQGAN 编码成 256 个离散标记, 并传递给语言解码器。同时, 我们从原始数据中提取几个代表性的物理值来指导模型的学习, 确保关键定量特征的保留。为丰富输入的语义结构, 我们还将生成的场的文本描述转换为图像表示, 并发送到图像解码器。压缩标记、代表性关键值和语义图像表示随后被输送到语言增强的多模态模型中。在我们提出的基准数据集上的实验结果表明, 我们的方法在现场数据理解任务中优于现有的最新方法, 在多个评估指标上实现了卓越的性能。特别是, 我们的论文做出了以下贡献:

1. 一个新颖的框架 (FieldLVLM), 它将视觉-语言建模与科学领域数据理解相结合。
2. 我们提出了一个数据重整流程, 一种领域感知的语言生成策略, 通过将高精度的专用模型与大型语言模型结合, 实现一致且可解释的领域描述。
3. 我们通过压缩原始数据以满足语言解码器的令牌限制, 开发了一种数据压缩多模态模型调优方法。同时, 我们选择具有代表性的关键值来指导网络的学习。
4. 大量实验表明, 我们的方法在具有挑战性的科学数据集上取得了优异的结果。

I. 相关工作

A. 大型语言模型

近年来, 数据可用性和计算资源的进步显著推动了大型语言模型的发展。早期的模型, 如基于编码器-解码器结构的 BERT [9] 和 T5 [10], 以及以解码器为中心的架构如 GPT [11], 利用 Transformer 框架在一系列自然语言处理任务中取得了显著成功。GPT3 [12] 的突破帮助确立了仅使用解码器的架构, 该架构通过自回归方法生成输出。在此基础上, 诸如 PaLM [13] 等模型在参数和训练数据方面进行了扩展, 而像 InstructGPT [14] 和 ChatGPT [15] 则通过使用微调和强化学习等技术提高了对话能力。

B. 大型视觉语言模型

随着大型语言模型 (LLMs) 的迅速发展, 出现了一组新兴的研究人员开始关注将视觉信息整合到这些模型中。这一方向的基础性工作来自视觉语言学习领域, 特别是模态对齐 [1], [16]。一个突出的例子是 CLIP [1], 该模型展示了如何通过在大规模图像-文本数据集上的对比学习有效地对齐视觉和文本表征。像 LLaVA [17] 和 InstructBLIP [18] 这样的模型进一步推动了该领域的发展, 通过细化指令调优来更好地解释复杂查询。尽管取得了这些进展, 但当前领域的的数据仍然缺乏高质量图像-文本对, 限制了大型视觉语言模型的应用范围。



Fig. 1. (a) 各种方法在实地数据中的视觉语言响应的定量比较。(b) 关键评估指标的性能，包括流动分类、雷诺数计算、涡流识别和综合实地数据解释。

C. 科学发现

科学发现传统上依赖于假设驱动的实验、统计分析和专业化模拟等方法。早期的计算工具，包括符号推理系统和基于专家的框架 [19]，在增强科学探究中发挥了基础性作用。然而，这些传统技术通常需要大量人工输入和深入的领域知识，使得它们难以在各个领域中扩展或适应。最近，数据中心方法的兴起，特别是深度学习的应用，重塑了发现的过程。大型语言模型 (LLMs) 和通用基础模型在自动假设生成 [20]、预测材料属性 [21] 和建模蛋白质结构 [22] 等领域表现出色。通过利用科学文本和实验结果的海量数据集，这些模型可以揭示隐藏的模式，提供可能超越人类直觉的见解。尽管它们具有潜力，但研究仍在解决关键问题，包括可解释性、对特定科学领域的泛化，以及生成内容的准确性。

在视觉-语言学习领域，已经进行了各种努力来提高图像-文本对字幕的质量。例如，LaCLIP 利用大型语言模型来优化原始字幕，尽管其性能往往因有限的视觉上下文和原始注释质量差而导致的幻觉而受到影响。其他研究探索了结合和筛选原始和合成字幕以提升 CLIP 性能的策略。一个较新的方法，VeCLIP，使用大型语言模型结合来自两个字幕来源的见解。然而，由于合成数据的低保真度，结果字幕仅整合了最少的视觉内容。迄今为止，LLaVA 在多模态模型领域表现突出，它通过将简短的人工编写字幕和边界框输入到 GPT-4 语言模型，使其能够在生成更丰富的描述之前模拟视觉感知。尽管如此，该方法仍然严重依赖劳动密集型的人工注释且缺乏真实的视觉输入，经常导致对主要对象（包括边缘区域中的那些）的过于详细的描述，仅依靠边界框提示进行引导，从而增加了幻觉的风险。相比之下，我们的工作采用了 GPT-4 Vision，这是一种最先进的多模态模型，能够直接从精心设计的提示和图像输入中生成丰富、准确的字幕，从而显著增强视觉知识的整合。

II. 方法

虽然像来自 CC [23] 和 LAION [24] 的图文对这样的大型多模态数据集的可用性极大地推动了视觉语言研究，但这一进展并没有延伸到科学领域。科学领域的多模态

数据集仍然稀缺，主要是因为获取这些数据存在困难，以及精确标注需要大量的专业知识和时间。为了应对这一限制，我们从 DeepSeek 模型 [25] 中获得灵感，该模型在自动文本标注方面表现出色。利用其功能，我们提议使用 DeepSeek 基于现有领域数据生成特定领域的文本描述。通过这种计算机化的方法，可以进行大规模的领域语言收集，减少对人工标注的依赖，同时提高生成的多模态数据集的质量和一致性。

给定一个数据域 X ，自然可以创建一组问题 X_q ，以描述该域内容。我们利用 Deepseek 来制作这样的问题列表。因此，将一个数据域扩展为其数据语言版本的简单方法是人： X ， $X_q < STOP >$ 助理： X_l ，其中 X_l 代表生成的语言描述。尽管这种简单的扩展版本容易构建，但缺乏多样性和深入的推理。

为了解决基于简单提示生成的浅层推理的局限性，我们将领域特定的模型作为强大的教师来丰富领域数据的语言表示。具体来说，我们利用高精度模型来完成分类 [26] 和检测 [27] 等任务，这些模型针对视觉输入进行操作。为此，我们首先将原始现场数据转换为图像表示。在专家知识的指导下，我们开始根据环境和几何特征对流场进行分类，例如区分腔体驱动流动和外部流动。例如，在顶盖驱动的腔体流动的情况下，我们识别出其决定性特征以支持更有针对性的分析。接下来，我们通过计算雷诺数来分类流动状态。这个回归模型提供了对潜在动力学的基本见解。此外，我们还考虑了罕见但关键的现象——如突发的流动不稳定或异常——这些可以作为低概率事件出现。通过使用专业的模型，可以早期检测到这些现象，并采取积极的缓解策略。

A. 数据压缩的多模态模型调整

1) 数据：大型语言模型已展示出从数据中获取大量领域知识的能力。然而，当它们应用于科学领域的数据时，面临着两个主要限制。首先，领域数据的纯文本表示通常缺乏明确的语义结构，使得模型难以提取有意义的模式或关系。因此，重要的物理特征可能被忽视或理解不佳。其次，这些模型受到固定最大输入长度的限制。领域数据——通常是高维和详细的——容易超过这个令牌限制，导致关键信息被截断或丢失。

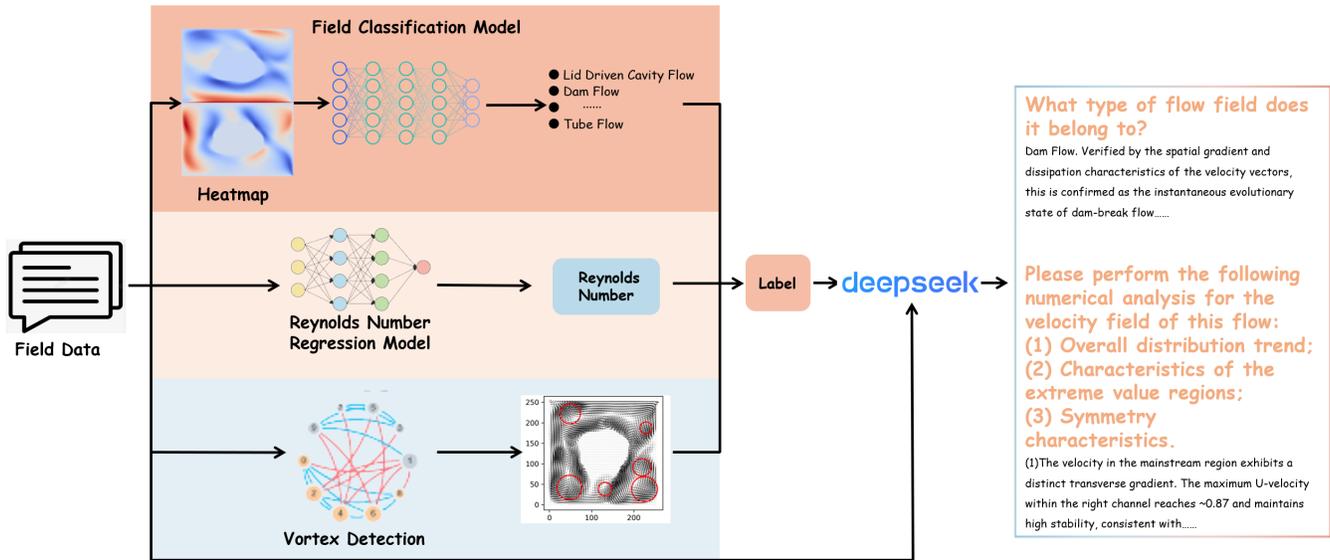


Fig. 2. 一种领域感知语言数据生成策略的流程，该策略集成了用于领域分类、雷诺数估算和涡旋检测的专用机器学习模型。

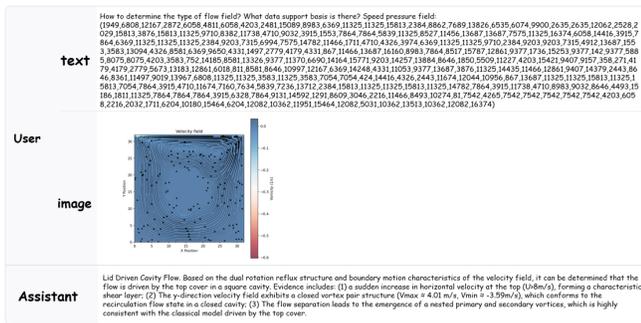


Fig. 3. 生成的场语言表示的示例，展示了从速度和压力数据得出的流场分析的结构化文本输出。

图像比文本更有利于语义信息的发现。因此，我们将字段数据的文本表示转换为图像表示。接着，字段数据的图像表示被输入到多模态大型模型的图像编码器中。

为了解决在领域科学数据中过多的 token 长度问题（例如速度-压力矩阵超过 65,536 个 token），我们实现了一种两阶段压缩流程，能够在保留关键物理特征的同时将输入维度减少 99.6%。在第一阶段，水平速度 (u)、垂直速度 (v) 和压力 (p) 的 256×256 标量场通过线性归一化并分别通过 $R \propto u_{\text{norm}}$, $G \propto v_{\text{norm}}$, $B \propto p_{\text{norm}}$ 映射到 3 通道 RGB 图像。然后，这种 RGB 表示通过一个预训练的 VQGAN 模型编码为 256 个离散 token，有效地将原始 65,536 字符序列压缩成与 Qwen2.5-VL 的 token 限制兼容的紧凑潜在表示。

数据中的局部关键数据对于分析整个数据具有指导意义。因此，基于这个原则，选择一个关键数据进行重点分析，以提高整体数据分析的有效性。

2) 训练：我们在 LLaMA-Factory 框架内使用低秩适应 [28] (LoRA) 进行参数高效微调，重点是将 Qwen2.5-VL-7B 模型适应于实地数据任务。该架构保持了两个关键设计原则：冻结视觉编码器，即 CLIP-ViT 主干在训练过程中完全锁定，以保留预训练的视觉表示并防止灾难性遗忘；

选择性参数更新，仅更新 LoRA 适配器（秩 = 32，缩放因子 = 128）和多模态投影器，与完全微调相比，减少了 98.7% 的可训练参数。

该模型接收包含两个组成部分的多模态输入：(1) 一个提出分析问题的文本提示，与表示速度-压力场的压缩符号序列配对；(2) 流场的热图可视化，为其提供附加的视觉上下文。模型经过训练生成与四个预定义分析类别之一对的结构化输出。例如，当负责涡流分析时，输出包括详细参数如涡流位置、大小、环流强度和旋转方向。

III. 实验

A. 实现细节和数据集统计

实现细节。我们使用 Qwen2.5-VL-7B 架构作为基本模型来评估 FieldLVLM。训练进行 4 个周期，学习率为 $5e-5$ ，最大梯度范数为 1.0，文本长度为 4096 个标记。每个设备处理一个包含 4 个样本的批次，进行 8 次梯度累积，从而实现一个有效的批量大小为 32。使用 10% 的验证拆分，并在前 100 步中应用学习率预热。

数据集统计。我们的实验基于 FlowBench [30] 和 CFD Bench [31] 的扩展版本，其中每个样本都配有由我们的领域感知策略生成的结构化语言描述。

B. 评估基准和指标

我们设计了四个基准任务来评估科学领域数据理解：流动分类、雷诺数估计、涡旋识别和综合领域分析。每个任务针对解释速度-压力场所需的特定分析能力，并使用上述数据集中的超过 70,000 个样本进行基准测试。

分类任务侧重于基于速度-压力场的形态和动态特征对流动类型进行分类，采用精度 (%) 定义为正确识别流动类型（例如，顶盖驱动的空腔流与绕流钝体的流动）与真实注释的比率。流体状态表现出独特的空间特征，可以通过视觉-语言融合进行检测——例如，在绕流钝体中卡门涡街或在顶盖驱动腔中的剪切层——高精度表明 FieldLVLM 能够将视觉模式映射到典型流动类别，对于初步诊断至关重要。

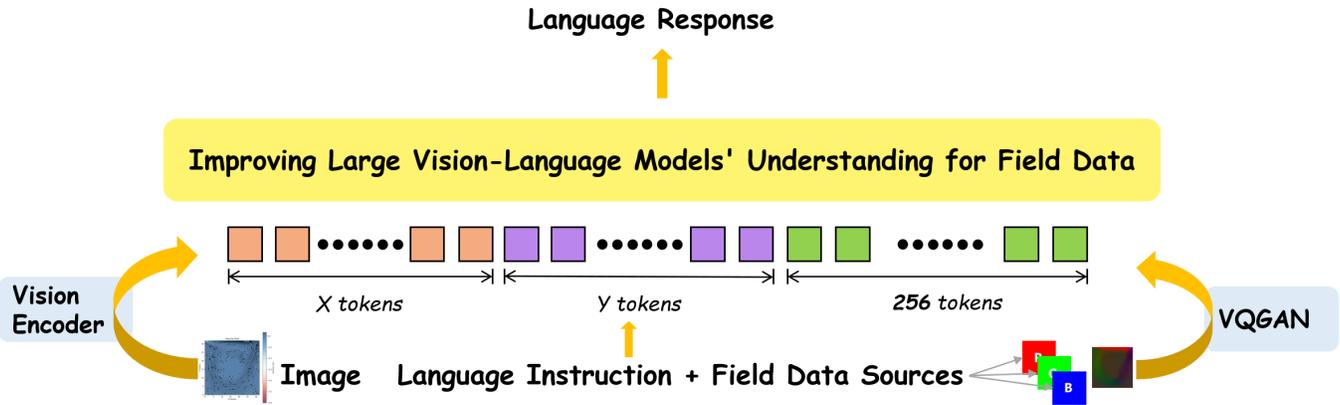


Fig. 4. 数据压缩多模态模型的输入输出架构，具备基于 VQGAN 的令牌压缩、键值选择和图像表示转换，以增强领域数据语义分析。

TABLE I

与一些视觉语言模型进行比较评估，包括 DEEPSEEK-VL-7B [25]、LLaVA-v1.6 [17] 和 LLAMA-3.2 [29] 在分类任务、雷诺数计算、涡流识别和综合场数据分析中的表现。

Method	Categorize Task	Reynolds Number	Vortex Identification	Field Data Analysis
DeepSeek-VL-7b-chat [25]	0/NA	0/NA	0/NA	0/NA
LLaVA-v1.6-vicuna-7b [17]	0/NA	0/NA	0/NA	0/NA
Llama-3.2-11B-Vision-Instruct [29]	0/NA	0/NA	0/NA	0/NA
FieldLVM (ours)	100	99.79	97.23	85.41

TABLE II

在 FIELDLVM 框架中，对数据压缩对涡旋识别准确性的影响进行消融分析：评估 QWEN2.5-VL-7B 基线、基础微调以及基于 VQGAN 的令牌压缩增强性能。

Method	Vortex Identification
Qwen2.5-VL-7B (Baseline) [7]	0/NA
Base Fine-tuning	82.28
+ Compress data	85.41

TABLE III

FIELDLVM 中关于关键数据选择用于现场数据分析的消融研究：包含 QWEN2.5-VL-7B 基线、基础微调 and 通过代表值引导优化准确性应对语义结构挑战的性能比较。

Method	Field Data Analysis
Qwen2.5-VL-7B (Baseline) [7]	0/NA
Base Fine-tuning	53.94
+ Key data	100

雷诺数计算从现场数据中计算出无量纲参数 ($Re = \frac{\rho U L}{\mu}$)，其精度使用了%，当相对误差为 $\leq 10\%$ 时预测被认为是正确的。这结合了数值精度与基于热力学的验证，以减轻先前 LVM 中普遍存在的幻觉风险，确保物理上合理的输出。

涡旋识别在检测相关结构的同时量化位置、半径和旋转方向。我们定义准确性 (%) 为成功，当且仅当：(i) 涡旋中心坐标偏离注释的流动域大小为 $\leq 25\%$ ，并且 (ii) 旋转方向与真实值匹配。这验证了 FieldLVM 从复杂拓扑

中提取几何特征的能力，解决了传统 LVM 在空间推理方面的局限性。

现场数据分析确定最大速度 (U_{max})，并通过自然语言 Q & A 解释现场语义。指标包括用于 U_{max} 定位 ($\leq 10\%$ 误差容限) 的准确性 (%), 以及在不同流动状态下进行速度-压力分析时衡量语义一致性的交叉熵损失。 U_{max} 评估测试细粒度回归能力，而交叉熵量化上下文推理，这对于从多模态输入中得出符合物理规律的见解至关重要。

我们系统地评估了所提出的 FieldLVM 框架与最新的基线模型，如 DeepSeek-vl-7b-chat、LLaVA-v1.6-vicuna-7b、Llama-3.2-11B-Vision-Instruct。我们在四个标准化任务上进行比较：分类、雷诺数计算、涡流识别和现场数据分析。如表所示，我们改进的大规模视觉-语言模型对现场数据的理解 (FieldLVM) 在科学现场数据处理方面显著优于现有的通用模型。FieldLVM 在雷诺数计算中实现了 99.79 的准确率，而基线模型 (deepseek-vl-7b-chat、Llava-v1.6-vicuna-7b、Llama-3.2-v1B-Vision-Instruct) 由于与科学数据结构的体系结构不兼容，结果为 0/NA。同样地，FieldLVM 在涡流识别中达到 97.23 的准确率，在现场数据分析中达到 85.41 的准确率，超越了所有对比模型，而这些模型在任务中得分为 0/NA。这些结果突显了传统架构在处理具有语义不规则性的长序列科学数据方面的局限性，同时证实了我们的方法在科学发现任务中增强了稳健性。

消融研究。表 II 详细描述了我们的组件级别调查。使用 Qwen2.5-VL-7B 作为基线 (0/NA 准确率)，仅通过基本微调，性能就提升到了 82.28%，这是通过初步适应现场数据特征实现的。随后整合我们的压缩策略 (+ 压缩数据) 使得准确率提升到 85.41%，验证了我们的两阶段管道 (RGB 映射 + VQGAN 编码) 在解决因数据长度过长 (如 256×256 矩阵) 导致的标记溢出方面的效率。这种 99.6%

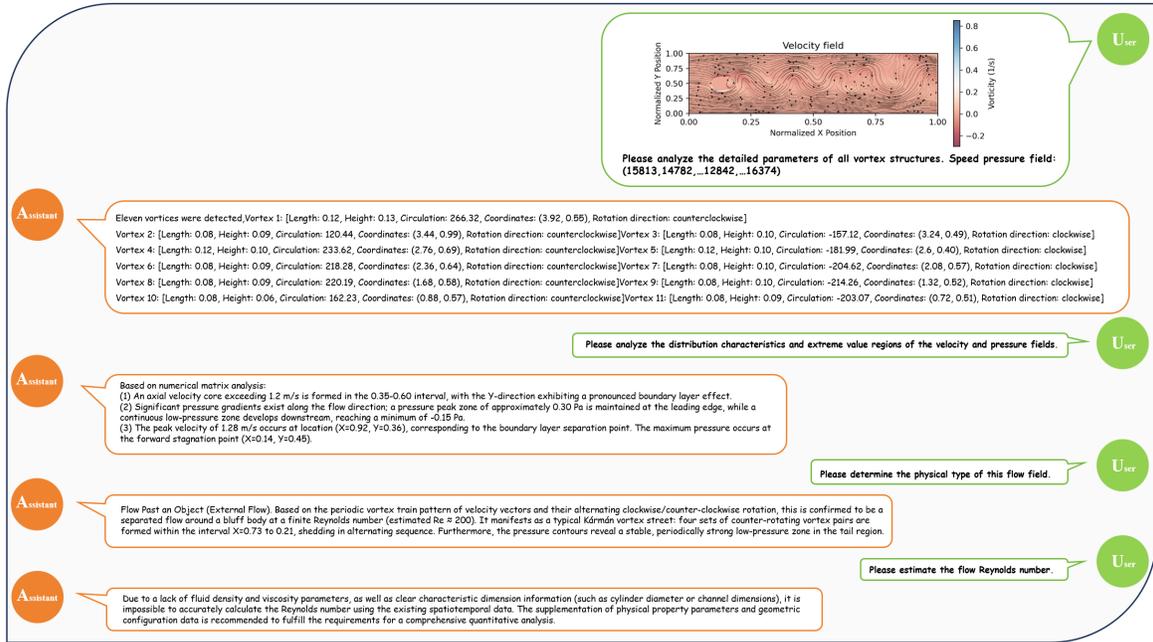


Fig. 5. 对流过钝体的涡旋脱落动态和压力分布进行分析，突出卡门涡街特征及三阶段流动结构。

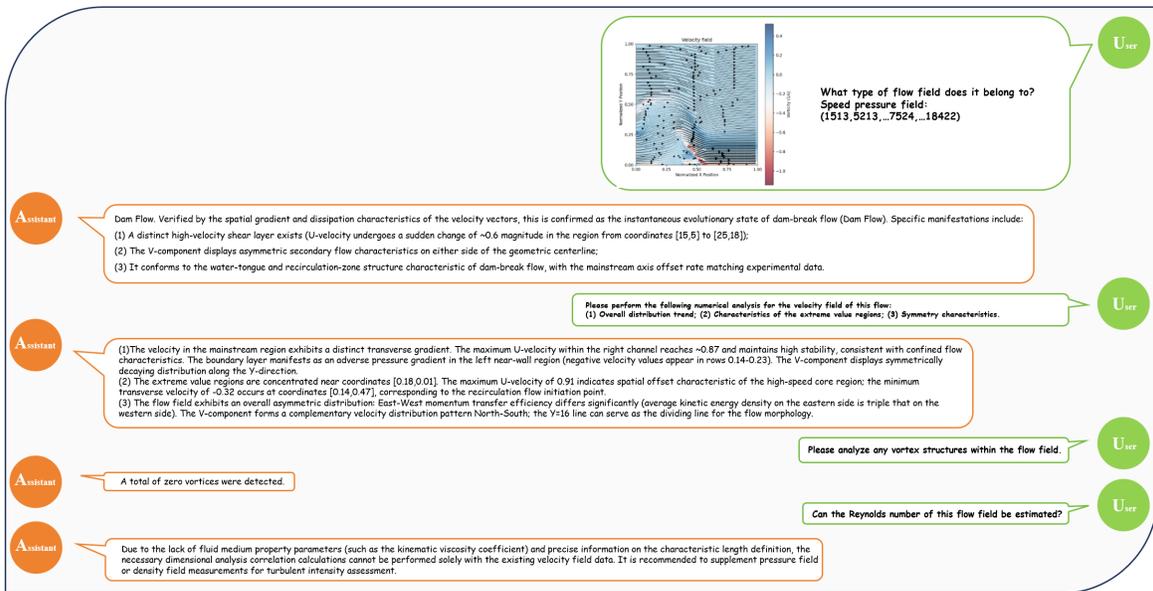


Fig. 6. Q & A 一场关于速度分布和涡度分析的会议，用于表征坝崩溃流场，展示剪切层动力学和关键流动状态。

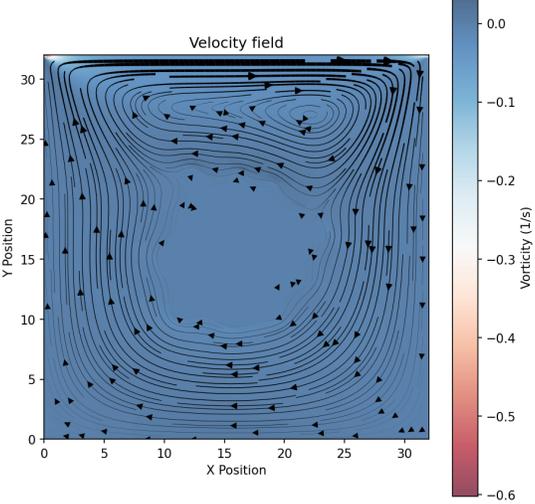
的维度缩减在提高计算效率的同时保留了关键物理特征，确立了数据压缩在资源密集型科学应用中的必要性。表 III 显示了不同的性能模式。相同的基线 (Qwen2.5-VL-7B) 得分为 0/NA，而基本微调实现了 53.94 % 的准确率——明显低于其他任务，反映了高维矩阵和语义稀疏的内在复杂性。值得注意的是，实施关键数据选择 (+ 关键数据) 后，准确率达到 100 %，这表明对高价值信号区域（如关键流动特征）的优先分析能够实现高效的全局模式提取和噪声消除。这与我们的核心设计原则一致：现场数据解释需要战略性地关注局部重要信息。

C. 定性结果

图 5 展示了 FieldLVLm 在钝体流动场景中对四个关键问题的响应：流场分类、涡旋结构分析、雷诺数估算和速度场解释。该测试验证了模型从速度-压力场数据中提取科学见解的能力。FieldLVLm 通过识别交替的涡旋脱落特性，准确识别了卡门涡街动态机制。在涡旋分析中，模型量化了 11 个检测到的涡旋的结构参数（包括半径、尺寸和旋转方向），展示了优异的涡旋检测能力。在速度场分析中，FieldLVLm 确定了峰值速度 (X=0.92, Y=0.36 时为 1.28 m/s) 和最大压力点 (X=0.14, Y=0.45)，并进行了速度-压力变化的详细三个方面的特征描述。

图 6 展示了 FieldLVLm 在瞬态溃坝流动四个任务上

TABLE IV
在以盖板驱动腔流为情景的大型视觉-语言模型中，流场分类性能的比较分析。

Visual input example, Extreme Ironing:	
	
Source: FlowBench/LDC_NS_2D/256x256/harmonics_lid_driven_cavity_Y/frame_376_vortex.png	
User	Judge the type of flow field based on the coupled effect of velocity field and pressure field.
FieldLVM	Lid-Driven Cavity Flow. Through comprehensive analysis, this flow is identified as a lid-driven cavity flow. Supporting evidence includes: The velocity vector matrix reveals characteristic structures: a strong shear layer at the top and recirculation zones at the bottom; Pressure gradients intensify significantly along the normal direction to the moving wall, aligning with the mixed Couette-Poiseuille flow characteristics; A secondary separation point induced by the primary vortex is detected at coordinate [0.71, 0.59] in the bottom-right corner, with its low-pressure core intensity matching the typical value at Reynolds number $Re = 37$.
User	Judge the type of flow field based on the coupled effect of velocity field and pressure field.
LLaVA	This chart shows a distorted flow field, presenting a very complex shape that includes many distorted curves and points. This type of flow field may be a special type of flow field, such as a twisted flow field or other types of nonlinear flow fields.
User	Judge the type of flow field based on the coupled effect of velocity field and pressure field.
Llama	In this chart, we can see the coordinate axis of a velocity field and a curved curve. This curve represents the motion of a fluid in space. Based on the characteristics of the chart, we can determine the type of motion of this fluid: 1. Shape of Curve: The curve presents a curved shape, indicating that the fluid exhibits rotational motion in space. This type of motion is commonly referred to as rotational motion. 2. Curvature of the curve: The curvature of the curve is large, indicating that the fluid is moving at a fast speed in space. 3. The degree of curvature of the curve: The degree of curvature of the curve is significant, indicating that the fluid is moving at a fast speed in space.
User	Judge the type of flow field based on the coupled effect of velocity field and pressure field.
DeepSeek	Based on the provided data, which includes position coordinates (X, Y) and vorticity values, the flow field exhibits rotational characteristics due to the presence of non-zero vorticity (ranging from 0.1 to 0.4 1/s). The flow is rotational and viscous, likely a vortex or shear-driven flow. The pressure field would exhibit radial gradients to balance the rotational motion, but additional pressure data would be needed for precise classification (e.g., Rankine vortex vs. free vortex).

的表现：速度分布分析、涡度特征化、流动分类和雷诺数估计。该模型准确定位了高速剪切层（从 [0.23,0.07] 到 [0.39,0.28]），并识别了流体冲击前沿的涡度集中区。FieldLVM 将流动分类为“溃坝流动的瞬时演化状态”，通过速度梯度分析匹配了物理特征，同时利用热力图可视化功能从 RGB 映射中提取语义信息。

表格 IV 比较了 FieldLVM、LLaVA、Llama 和 DeepSeek 在基于速度和压力场耦合效应分类盖板驱动空腔流动时的表现。当被要求“使用耦合速度-压力场效应判断流动类型”时，FieldLVM 准确地识别出流动为盖板驱动空腔流动，同时描述了剪切层的形成和角涡结构，与经典流体力学原理相符。相反，LLaVA 显示出泛化错误（例如，将其标记为“畸变流场”）并忽视空腔特有的特征。Llama 则过度强调曲线几何（描述“旋转运动”）而不关联压力梯度，DeepSeek 仅提供部分正确的响应（提到“剪切驱动流”）但缺乏定量参数。至关重要，FieldLVM 生

成的结构化响应富含领域特定术语（例如，“剪切层”，“再循环区域”），揭示其他模型的知识差距，并验证了其 RGB 映射加 VQGAN 压缩机制的语义结构保留能力。

表格 V 比较了 FieldLVM、LLaVA、Llama 和 DeepSeek 在包含相干结构的流场上涡旋特征化能力。当提示分析涡旋参数时，FieldLVM 能够准确识别出两个涡旋，并报告完整的物理描述。涡旋 1 表现出逆时针旋转，环量为 +168.36，核心坐标为 (0.38, 0.24)，尺寸为 0.40×0.40 。涡旋 2 顺时针旋转，环量为 -142.15，位置为 (0.51, 0.72)，大小为 0.45×0.45 。这表明 FieldLVM 可以精确地整合位置、涡量和几何指标。相比之下，LLaVA 提供模糊的几何描述（例如，“弯曲的曲线”），未能捕捉到旋转机制。Llama 提供非物理解释，将矢量场视为标量函数，忽略涡旋结构。DeepSeek 检测到两个涡旋，但引入了关键错误：核心位置偏离超过 20%，环量大小被高估（例如， -5.0 s^{-1} 实际上是 -4.73 s^{-1} ），并且旋转方向颠倒。只有 FieldLVM 使

TABLE V
在腔体流动中使用视觉语言模型生成结构化输出的涡旋参数提取精度比较。

Visual input example, Extreme Ironing:	
Source: CFDBench/cavity/geo/case0011/frame_18_vortex.png	
User	Please analyze the detailed parameters of all vortex structures.
FieldLVM	Two vortices were detected, vortex 1: [length 0.40, height 0.40, circulation 168.36, coordinates (0.38,0.24), rotation direction: counterclockwise] vortex 2: [length 0.45, height 0.45, circulation -142.15, coordinates (0.51,0.72), rotation direction: clockwise]
User	Please analyze the detailed parameters of all vortex structures.
LLaVA	This chart displays the velocity field of a vortex structure, which includes some parameters and distribution patterns. 1. Shape of velocity field: The velocity field of the vortex structure presents a curved curve, which can help us understand the shape of the vortex structure. 2. Distribution of velocity field: The distribution of vortex structures in the velocity field presents a curved shape, which can help us understand the distribution of vortex structures. 3. Strength of velocity field: The vortex structure in the velocity field has a certain strength, which can be measured by the strength value of the velocity field.
User	Please analyze the detailed parameters of all vortex structures.
Llama	This chart shows the parameters and spatial distribution of a vortex structure. This structure is called the 'Velocity field', representing the velocity field. The following is an analysis of the parameters and their spatial distribution patterns in the chart: 1. X Position: This parameter represents the position of the velocity field in the X-axis direction. The chart shows that the velocity field presents a curved curve on the X-axis, indicating that the variation of the velocity field in the X-axis direction is nonlinear. 2. Y Position: This parameter represents the position of the velocity field in the Y-axis direction. The chart shows that the velocity field presents a curved curve on the Y-axis, indicating that the variation of the velocity field in the Y-axis direction is also nonlinear. 3. Velocity: This parameter represents the size of the velocity field. The chart shows that the velocity field presents a curved curve throughout the entire space, indicating that the magnitude of the velocity field also varies nonlinearly in space.
User	Please analyze the detailed parameters of all vortex structures.
DeepSeek	Vortex 1 (Upper Vortex)Location (Core Center): Vortex 1 (Upper Vortex)Location (Core Center): Approximate coordinates: (X: 0.50, Y: 0.70)Positioned slightly above the vertical midline. Rotation Direction: Anticyclonic/Counter-Clockwise (inferred from red/yellow vorticity → positive vorticity).Vorticity Magnitude: Peak vorticity: +1.5 to +2.0 s^{-1} (centered in red region).Core region: Strong vorticity ($> +1.0 s^{-1}$) extends 0.15 normalized units radially. Size (Radius): Core radius: 0.10–0.15(defined by closed streamlines and peak vorticity). Vortex 2 (Lower Vortex)Location (Core Center): Approximate coordinates: (X: 0.50, Y: 0.30)Positioned symmetrically below Vortex 1. Rotation Direction: Cyclonic/Clockwise (inferred from blue vorticity → negative vorticity).Vorticity Magnitude: Peak vorticity: -4.0 to -5.0 s^{-1} (deep blue core). Core region: Strong vorticity ($< -4.0 s^{-1}$) extends 0.10–0.12 normalized units radially. Size (Radius): Core radius: 0.08–0.12.

TABLE VI
场数据分析中涡旋检测能力的定量指标: 视觉-语言模型在核心参数方面的性能比较

Capability	FieldLVM	LLaVA	Llama	DeepSeek
Vortex detection count	✓ (2)	✗ (0)	✗ (0)	✓ (2)
Core position error	0 %	N/A	N/A	> 20 %
Circulation quantification	✓	✗	✗	△ (sign error)
Rotation direction accuracy	✓	✗	✗	✗ (inverted)

用“环量”、“核心坐标”和“旋转方向”等术语生成完整

的、与领域相关的回答,验证了其 VQGAN-RGB 压缩在保持涡量拓扑方面的有效性。

D. 定性评估

表 VI 进一步量化了模型的表现。FieldLVM 在涡旋计数、核心定位、环流量大小和旋转方向上实现了完美的准确性,突显了它对涡旋动力学的深刻理解。相比之下,DeepSeek 错误定位了涡旋中心,错误估计了环流,并且经常误分类旋转方向。这些结果证实了 FieldLVM 在提取和推理流动场中的结构化物理参数方面的卓越能力。

本文提出了一种新方法，旨在提高大型视觉语言模型对现场数据的理解。我们的方法首先设计了一种领域感知的语言生成策略，将特殊模型的高精度优势与大型语言模型的一致性相结合，作为我们的数据来源。然后，提出了数据压缩的多模态模型，以压缩数据来适应 LVLMM 的输入大小。它在现场数据上实现了优异的准确性。此外，我们建立了一个专用的评估基准，用于评估 LVLMM 在这一领域的的能力。虽然这项工作专注于四个具有代表性的任务，但它作为一个初步步骤，为更广泛地将 LVLMM 整合到科学领域分析中铺平了道路。我们希望这项研究能激发进一步探索将视觉语言模型应用于复杂科学数据，并支持开发更健壮、更具广泛适用性的多模态系统，以推动科学发现。

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. Pmlr, 2021, pp. 8748–8763.
- [2] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, “Florence: A new foundation model for computer vision,” *arXiv preprint arXiv:2111.11432*, 2021.
- [3] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, “Grounded language-image pre-training,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10965–10975.
- [4] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, “Regionclip: Region-based language-image pretraining,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16793–16803.
- [5] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan *et al.*, “Generalized decoding for pixel, image, and language,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15116–15127.
- [6] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Rantftl, “Language-driven semantic segmentation,” *arXiv preprint arXiv:2201.03546*, 2022.
- [7] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, “Qwen2. 5-vl technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [8] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [11] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
- [14] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.
- [15] S. S. Biswas, “Role of chat gpt in public health,” *Annals of biomedical engineering*, vol. 51, no. 5, pp. 868–869, 2023.
- [16] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.
- [17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.
- [18] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” 2023.
- [19] E. A. Feigenbaum *et al.*, “The art of artificial intelligence: Themes and case studies of knowledge engineering,” 1977.
- [20] Y. Zheng, H. Y. Koh, J. Ju, A. T. Nguyen, L. T. May, G. I. Webb, and S. Pan, “Large language models for scientific discovery in molecular property prediction,” *Nature Machine Intelligence*, pp. 1–11, 2025.
- [21] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, “Unsupervised word embeddings capture latent knowledge from materials science literature,” *Nature*, vol. 571, no. 7763, pp. 95–98, 2019.
- [22] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko *et al.*, “Highly accurate protein structure prediction with alphafold,” *nature*, vol. 596, no. 7873, pp. 583–589, 2021.
- [23] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.
- [24] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *Advances in neural information processing systems*, vol. 35, pp. 25278–25294, 2022.
- [25] Q. Zhu, D. Guo, Z. Shao, D. Yang, P. Wang, R. Xu, Y. Wu, Y. Li, H. Gao, S. Ma *et al.*, “Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence,” *arXiv preprint arXiv:2406.11931*, 2024.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [27] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.
- [28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models,” *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [29] Meta, “Llama,” 2024. [Online]. Available: <https://www.llama.com/>
- [30] R. Tali, A. Rabeh, C.-H. Yang, M. Shadkhah, S. Karki, A. Upadhyaya, S. Dhakshinamoorthy, M. Saadati, S. Sarkar, A. Krishnamurthy *et al.*, “Flowbench: A large scale benchmark for flow simulation over complex geometries,” *arXiv preprint arXiv:2409.18032*, 2024.
- [31] Y. Luo, Y. Chen, and Z. Zhang, “Cfdbench: A large-scale benchmark for machine learning methods in fluid dynamics,” *arXiv preprint arXiv:2310.05963*, 2023.