Computer Science and Application 计算机科学与应用, 2026, 16(6), 140-154
Published Online June 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.166215

Hans汉斯

面向BEVFormer时空信息融合的层级化可解释性分析

禹　鑫

西华大学汽车与交通学院，四川 成都

收稿日期：2026年5月10日；录用日期：2026年6月15日；发布日期：2026年6月24日

摘　要

针对BEVFormer在多摄像头空间信息与历史BEV时序信息融合过程中的决策依据不透明问题，本文基于Generic Attention-model Explainability (GAE)的相关性传播思想，构建面向BEVFormer的层级化归因解释方法。由于BEVFormer具有不同于标准Transformer的时空多阶段信息传播结构，本文分别建立检测查询到当前BEV、当前BEV到图像特征以及当前BEV到历史BEV的归因链路。针对可变形注意力难以直接形成规则注意力矩阵的问题，结合采样点权重及其梯度，将采样点级正贡献映射至规则BEV网格或图像特征空间。基于nuScenes数据集的实验结果表明，本文方法能够定位目标相关BEV区域、关键摄像头视角和局部历史BEV区域；忠诚度实验和单分支全遮挡实验进一步表明，图像特征主要支撑类别与几何属性估计，历史BEV特征对速度估计和运动连续性保持具有更明显作用。

关键词

三维目标检测，可解释性，层级化归因，可变形注意力，BEVFormer

Hierarchical Explainability Analysis of Spatiotemporal Information Fusion in BEVFormer

Xin Yu

School of Automobile and Transportation Engineering, Xihua University, Chengdu Sichuan

Received: May 10, 2026; accepted: June 15, 2026; published: June 24, 2026

Abstract

To address the opaque decision basis of BEVFormer in fusing multi-camera spatial information and

文章引用: 禹鑫. 面向 BEVFormer 时空信息融合的层级化可解释性分析[J]. 计算机科学与应用, 2026, 16(6): 140-154.
DOI: 10.12677/csa.2026.166215

historical BEV temporal information, this paper constructs a hierarchical attribution explanation method for BEVFormer based on the relevance propagation idea of Generic Attention-model Explainability (GAE). Since BEVFormer has a spatiotemporal multi-stage information propagation structure that differs from standard Transformers, this paper establishes attribution paths from detection queries to current BEV features, from current BEV features to image features, and from current BEV features to historical BEV features. To overcome the difficulty that deformable attention cannot directly form regular attention matrices, sampling-point-level positive contributions are mapped to regular BEV grids or image feature spaces by combining sampling weights and their gradients. Experimental results on the nuScenes dataset show that the proposed method can locate target-related BEV regions, key camera views, and local historical BEV regions. Faithfulness experiments and single-branch full-masking experiments further indicate that image features mainly support category and geometric attribute estimation, while historical BEV features play a more significant role in velocity estimation and motion continuity preservation.

Keywords

3D Object Detection, Explainability, Hierarchical Attribution, Deformable Attention, BEVFormer

1. 引言

三维目标检测是自动驾驶环境感知中的关键任务，其目的在于从复杂道路场景中准确感知周围目标的空间位置、尺度与运动状态，为行为预测、路径规划和决策控制提供基础支撑[1]。相较于激光雷达方案，基于多摄像头图像的感知方法具有成本较低、语义信息丰富等优势，因此近年来受到广泛关注。但受限于透视成像机制，直接在图像视角下进行目标预测，难以充分实现多视角信息的有效对齐与融合，于是构建统一空间表征成为提升相机三维感知性能的重要研究方向。

围绕这一思路，相关研究主要从视角变换、深度估计和时序融合等方面展开。其中，OFT [2]、Lift-Splat-Shoot [3]和 BEVDet [4]通过不同形式的视角变换实现多摄像头图像特征到 BEV 空间的映射；BEVDepth [5]和 BEVStereo [6]则从深度监督和时序立体匹配角度提升 BEV 特征构建质量；BEVDet4D [7]进一步引入多帧时序信息以增强动态场景感知能力。作为时空 BEV 感知模型的代表，BEVFormer [8]利用空间交叉注意力聚合多摄像头图像特征，并通过时间自注意力融合历史 BEV 信息。尽管 BEVFormer 取得了较好检测性能，但其结果由空间观测信息与历史时序信息共同决定，内部传播路径较长，决策依据缺乏直观解释。因此，有必要围绕 BEVFormer 的时空信息融合过程开展可解释性研究。

针对深度模型的黑箱性问题，已有研究提出了多种后验可解释性方法，主要包括梯度归因、相关性传播和局部代理解释等方向。梯度类方法利用模型输出对输入或中间特征的敏感性刻画特征重要性，代表性工作包括 saliency map [9]、Grad-CAM [10]和 Integrated Gradients [11]；相关性传播类方法将模型输出逐层分解回输入空间，典型方法包括 LRP [12]和 DeepLIFT [13]；局部代理和加性归因方法则通过可解释模型或特征加性分解解释局部预测，如 LIME [14]和 SHAP [15]。这些方法为理解深度模型预测依据提供了基础，但主要面向卷积网络或一般深度模型，难以充分刻画 Transformer 中多头注意力、残差连接和跨层信息传播过程。

针对 Transformer 模型，相关研究逐渐从直接注意力可视化转向跨层信息传播和相关性分解分析。早

期工作常将注意力图作为解释依据[16] [17]，但后续研究指出，原始注意力权重与特征重要性及模型输出之间并不总能保持稳定一致的对应关系[18]。为刻画跨层信息流动，Abnar 等[19]提出了 attention rollout 与 attention flow；Chefer 等[20]将相关性传播思想引入 Transformer，并进一步提出 GAE [21]，将解释范围扩展到双模态和编码器–解码器结构。随后，相关研究从守恒传播[22]、logits 更新[23]、注意力层 LRP [24] 及分解式方法比较[25]等角度进一步提升 Transformer 解释的一致性与适用性。

尽管现有 Transformer 可解释性方法已能够刻画标准注意力结构中的信息传播关系，但对于 BEVFormer 这类多摄像头时空 BEV 感知模型仍难以直接适用。一方面，BEVFormer 并非标准 Transformer 结构，其检测结果涉及检测查询、当前 BEV 特征、多视角图像特征和历史 BEV 特征之间的多阶段传播，难以通过单一输入特征解释其决策依据；另一方面，BEVFormer 的空间交叉注意力和时间自注意力均采用可变形注意力机制，其注意力关系由连续采样位置和采样权重共同决定，无法直接形成规则稠密注意力矩阵。此外，当前 BEV 特征同时融合图像观测与历史时序信息，使空间来源与时序来源在最终检测结果中相互耦合，难以直接区分不同输入分支对类别、几何属性和运动状态估计的支撑作用。因此，如何在可变形注意力结构下建立层级化归因路径，并分析图像特征与历史 BEV 特征的差异化贡献，是 BEVFormer 可解释性分析中需要关注的问题。

为此，本文基于 Generic Attention-model Explainability (GAE)的相关性传播思想，构建面向 BEVFormer 的层级化归因解释方法。本文主要工作如下：

1) 构建面向 BEVFormer 时空信息融合过程的层级化归因框架，将解释过程划分为检测查询到当前 BEV、当前 BEV 到多视角图像特征以及当前 BEV 到历史 BEV 特征三个阶段，实现检测结果空间来源与时序来源的联合分析。

2) 构建适配可变形注意力结构的采样点级正贡献重建方法，结合采样点权重及其梯度计算正贡献，并依据采样位置将其映射至规则 BEV 网格或图像特征空间，使相关性传播能够适配 BEVFormer 的稀疏采样注意力结构。

3) 设计面向图像特征分支与历史 BEV 分支的忠诚度验证和贡献分析实验，通过特征删除式忠诚度实验和单分支全遮挡实验，分析两类特征在类别、几何属性和速度估计中的差异化支撑作用。

2. 方法

本文围绕 BEVFormer 检测结果的空间来源与时序来源分析，构建面向时空信息融合过程的层级化归因解释方法。该方法以 GAE 的相关性传播思想为基础，针对 BEVFormer 中检测查询、当前 BEV 特征、多视角图像特征和历史 BEV 特征之间的多阶段传播关系，依次建立检测查询到当前 BEV、当前 BEV 到多视角图像特征以及当前 BEV 到历史 BEV 特征的归因链路。同时，针对 BEVFormer 中可变形注意力难以直接形成规则注意力矩阵的问题，本文将采样点级正贡献重建到规则 BEV 网格或图像特征空间，使相关性传播能够适配其稀疏采样注意力结构。

2.1. 总体框架与相关性传播基础

BEVFormer 的检测结果由检测查询、当前 BEV 特征、多视角图像特征和历史 BEV 特征共同决定，单一输入空间的归因难以区分空间观测信息与历史时序信息的贡献。为此，本文将解释过程划分为三个阶段：首先建立待解释检测查询到当前 BEV 特征的归因关系，以确定检测结果在 BEV 空间中的关键支撑区域；随后以当前 BEV 特征为中间变量，分别回溯其对应的多视角图像证据和历史 BEV 时序贡献。本文层级化归因与正贡献重建的总体流程如图 1 所示。

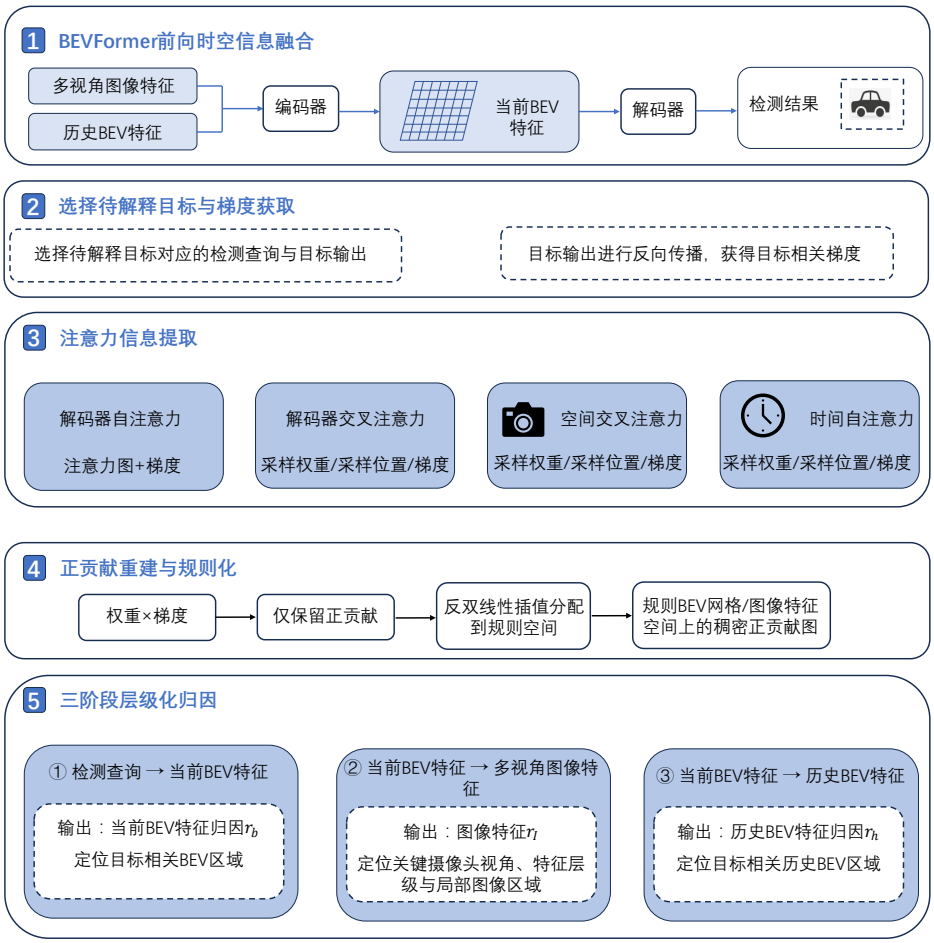


Figure 1. Overall workflow of three-stage hierarchical attribution and positive contribution reconstruction for BEVFormer
图 1. BEVFormer 三阶段层级化归因与正贡献重建总体流程

如图 1 所示，本文利用解码器自注意力与交叉注意力建立检测查询到当前 BEV 特征的相关性传播关系；进一步结合编码器中的空间交叉注意力，将目标相关 BEV 区域回溯至不同摄像头视角和特征层级；同时利用时间自注意力中的当前分支与历史分支，分析历史 BEV 特征对当前检测结果的贡献。由此，本文形成检测查询到当前 BEV、当前 BEV 到图像特征、当前 BEV 到历史 BEV 的层级化归因框架。

本文以 GAE 的相关性传播思想作为基础。对于任一注意力层，GAE 不直接采用原始注意力图，而是结合注意力权重及其梯度构造正贡献注意力图：

$$\bar{A} = E_h\left((\nabla A \odot A)^+\right) \tag{1}$$

其中，A 表示注意力图，$\nabla A = \partial y^* / \partial A$ 表示待解释目标得分 y^* 对注意力图 A 的梯度，$\odot$ 表示 Hadamard 乘积，$(\cdot)^+$ 表示仅保留正贡献项，$E_h(\cdot)$ 表示沿注意力头维度取平均。

设 R^{xx} 表示特征空间 x 内部的自相关矩阵，R^{xy} 表示特征空间 x 到特征空间 y 的跨空间相关矩阵，初始化时令 $R^{xx} = \mathrm{I}$、$R^{xy} = 0$。对于自注意力层，相关性递推可写为：

$$R^{xx} = R^{xx} + \bar{A} R^{xx} \tag{2}$$

$$R^{xy} = R^{xy} + \bar{A} R^{xy} \tag{3}$$

对于交叉注意力层，在对累积自相关矩阵进行行归一化后，跨空间相关性更新为：

$$R^{xy} = R^{xy} + \left(\bar{R}^{xx}\right)^{\mathrm{T}} \cdot \bar{A} \cdot \bar{R}^{xy} \tag{4}$$

上述传播规则为不同特征空间之间的相关性传播提供了统一形式。需要说明的是，解码器交叉注意力直接访问编码器最终输出的当前 BEV 特征；若继续沿 BEV 空间向编码器早期层传播，归因对象将由最终决策依据转向中间 BEV 表示，容易造成解释语义偏移。因此，本文采用分阶段归因策略，先获得检测查询对最终当前 BEV 特征的依赖关系，再分别回溯图像分支和历史 BEV 分支的来源贡献。

2.2. 正贡献选择的理论依据与局限性

在注意力相关性构造中，原始注意力权重主要反映信息聚合强度，不能直接表明其对待解释输出的目标相关贡献。对于 BEVFormer 的可变形注意力，较大的采样权重仅表示对应采样点在前向聚合中占比较高，并不必然说明其对当前检测结果具有正向作用。因此，仅依据原始采样权重归因，容易将被模型访问但贡献有限甚至具有抑制作用的区域误判为关键证据。

为增强归因结果与待解释目标之间的相关性，GAE 采用注意力权重与目标输出梯度的逐元素乘积构造目标相关注意力贡献。该乘积同时包含注意力连接的聚合强度和目标输出对该连接变化的局部敏感性，能够反映其对当前输出的影响方向与影响强度。当乘积为正时，表示该连接对待解释输出具有促进作用；当乘积为负时，则表示该连接对当前输出具有反向或抑制作用。若在面向支持性证据的可视化中直接混合正、负贡献，负贡献可能削弱或掩盖正向支持区域的显著性，从而不利于突出模型形成当前检测结果时依赖的支持性证据。因此，本文仅保留正贡献项，用于刻画对当前检测结果具有正向支撑作用的 BEV 区域、图像特征和历史 BEV 区域。

从相关性传播角度看，仅保留正贡献并非否定负贡献的解释意义，而是与 LRP 中区分正向证据和反向证据的思想一致。LRP 可根据解释目标采用不同传播规则，既可侧重传播支持目标输出的正向证据，也可同时考虑正、负贡献以获得更完整的相关性分解[12]。需要说明的是，本文并未对 BEVFormer 的全部模块建立严格守恒式相关性传播规则，因此更适合回答哪些区域支持当前检测结果，而不能完整刻画抑制性负贡献及正负因素之间的全部作用关系，后续可结合正负贡献分解和守恒式传播规则进一步完善。

2.3. 可变形注意力的稠密正贡献重建

由于可变形注意力仅在连续采样点处计算注意力权重，无法直接得到规则的 token-to-token 注意力矩阵。为使 2.1 节中的相关性传播规则适用于 BEVFormer，本文将采样点级正贡献重建为规则 BEV 网格或图像特征空间上的稠密正贡献矩阵。以下以解码器交叉注意力为例说明，重建流程如图 2 所示。

设第 m 个检测查询在第 h 个注意力头、第 k 个采样点处的注意力权重为 $a_{m,h,k}$，其对应梯度为 $\partial y^{*} / \partial a_{m,h,k}$，即待解释目标得分 y^{*} 对该采样点注意力权重的梯度。参照 2.1 节中正贡献注意力图的构造方式，可将采样点级正贡献定义为：

$$\bar{a}_{m,h,k} = \left(\frac{\partial y^{*}}{\partial a_{m,h,k}} \cdot a_{m,h,k} \right)^{+} \tag{5}$$

其中，y^{*} 表示待解释目标对应的置信度得分，$(\cdot)^{+}$ 表示仅保留正贡献项。

设该采样点的连续坐标为 $p_{m,h,k} = \left(x_{m,h,k}, y_{m,h,k}\right)$，其相邻离散网格单元集合为 $N\left(p_{m,h,k}\right)$，第 n 个相邻网格单元的中心坐标记为 $\left(x_n, y_n\right)$，为将采样点贡献映射到规则网格，本文采用反双线性分配方式，其分配系数为：

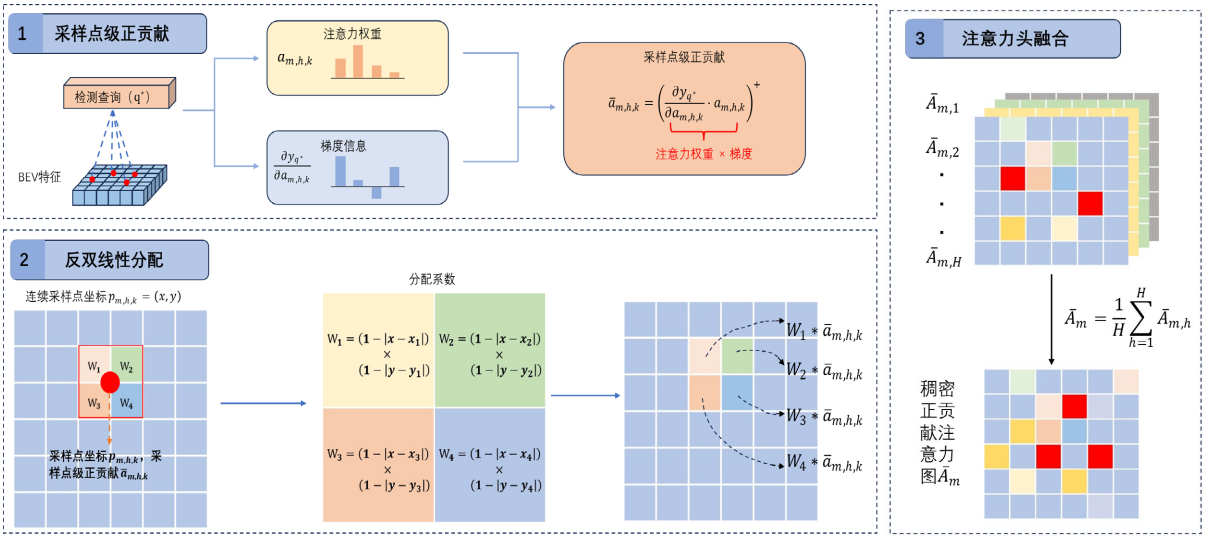


Figure 2. Schematic diagram of deformable attention-dense positive contribution reconstruction process
图 2. 可变形注意力稠密正贡献重建过程示意图

$$\omega\left(p_{m,h,k},n\right)=\begin{cases}\left(1-\left|x_{m,h,k}-x_n\right|\right)\left(1-\left|y_{m,h,k}-y_n\right|\right),n\in N\left(y_{m,h,k}\right)\\0,n\in N\left(p_{m,h,k}\right),n\notin\left(y_{m,h,k}\right)\end{cases} \tag{6}$$

由此，第 m 个检测查询在第 h 个注意力头上对应第 n 个规则网格单元的稠密正贡献为：

$$\bar{A}_{m,h,n}=\sum_{k=1}^{K}\omega\left(p_{m,h,k},n\right)\bar{a}_{m,h,k} \tag{7}$$

进一步沿注意力头维度进行聚合，便可得到检测查询与当前 BEV 网格之间的稠密正贡献注意力表示：

$$\bar{A}_{m,n}=\frac{1}{H}\sum_{h=1}^{H}\bar{A}_{m,h,n} \tag{8}$$

其中，K 为采样点数量，H 为注意力头数量。该重建过程将可变形注意力中定义于连续采样点上的正贡献映射为规则网格上的稠密表示，从而可与 2.1 节的相关性传播规则衔接，并作为后续层级化归因计算的基础。

2.4. 检测查询到当前 BEV 特征的归因

BEVFormer 的检测结果由解码器中的目标查询生成，因此，对单个检测结果进行解释时，首先需要确定该查询对编码器最终输出当前 BEV 特征的依赖关系。设待解释目标对应的查询索引为 q^*，解码器查询空间记为 d，BEV 空间记为 b。定义检测查询的自相关性矩阵为 $R^{dd}\in\mathbb{R}^{N_d\times N_d}$，检测查询与 BEV 特征之间的跨空间相关性矩阵为 $R^{db}\in\mathbb{R}^{N_d\times N_b}$，其中 N_d、N_b 分别表示检测查询和 BEV 特征的个数。初始化方式沿用 2.1 节，即 $R^{dd}=I$，$R^{db}=0$。

在解码器相关性传播过程中，解码器自注意力用于更新查询空间内部的相关性 R^{dd}，并使检测查询与当前 BEV 特征的相关性 R^{db} 在查询空间中继续更新；解码器交叉注意力则用于更新检测查询与当前 BEV 特征之间的跨空间相关性。由于该交叉注意力采用可变形注意力，本文根据 2.3 节将其采样点级正贡献重建为稠密正贡献矩阵 $\bar{A}^{db}$，并按照交叉注意力传播规则更新 R^{db}：

$$R^{db}=R^{db}+\left(\bar{R}^{dd}\right)^{\mathrm{T}}\bar{A}^{db} \tag{9}$$

式中 $\bar{R}^{dd}$ 表示经查询自注意力累积并进行行归一化后的查询空间自相关矩阵。该式表明，经自注意力上下文化后的检测查询可通过交叉注意力进一步将相关性传递至当前 BEV 空间。与标准 GAE 中继续向编码器内部传播不同，本阶段仅解释检测查询对编码器最终输出 BEV 特征的直接依赖，不再沿 BEV 空间回溯至编码器早期层。

在实际计算过程中，本节沿 6 层解码器层对检测查询到 BEV 特征的跨空间相关性矩阵进行逐层递推更新与累计，从而得到最终的 R^{db}。对于待解释目标，取其对应查询 q^* 对应的相关性分布作为 BEV 归因向量：

$$r_b = R^{db}\left[q^*,:\right] \tag{10}$$

其中，$r_b \in \mathbb{R}^{1\times N_b}$ 表示待解释检测结果对当前 BEV 各空间单元的贡献强弱。该向量既用于定位目标相关 BEV 区域，也作为后续图像分支和历史 BEV 分支归因的输入。

2.5. 当前 BEV 特征到多视角图像特征的归因

基于 2.4 节得到的 BEV 归因向量 r_b，本文进一步分析目标相关 BEV 区域的图像来源。由于 BEVFormer 通过空间交叉注意力从多摄像头、多尺度图像特征中聚合空间观测信息，因此，本节建立当前 BEV 特征到图像特征的跨空间归因关系，用于确定待解释目标所依赖的关键摄像头视角和图像区域。

图像特征来源于 6 个摄像头视角的 4 个 FPN 特征层级，需分别建立 BEV 特征对各相机视角、各特征层级的相关性矩阵。记第 c 个摄像头、第 m 个特征层级上的图像特征空间为 $I_{c,m}$，其中，$c=1,2,\cdots,6$，$m=1,2,3,4$。设 $\bar{A}^{bI}_{(c,m)}$ 与 $R^{bI}_{(c,m)}$ 分别为 BEV 特征到第 c 个相机视角的第 m 个特征层级图像特征空间的稠密正贡献注意力矩阵与跨空间相关性矩阵；设 $\bar{A}^{bb}$ 与 R^{bb} 分别为时间自注意力当前分支对应的稠密正贡献注意力矩阵与自相关矩阵。其中 $\bar{A}^{bI}_{(c,m)}$ 与 $\bar{A}^{bb}$ 均依据 2.3 节提出的稠密正贡献重建方法获得。由于图像特征与历史 BEV 特征分别经由空间交叉注意力和时间自注意力汇入当前 BEV，二者之间不存在直接注意力交互，因此本节仅沿当前 BEV 到图像特征的路径建立归因关系。

在编码器中，当前 BEV 特征先通过时间自注意力当前分支更新 BEV 空间内部的相关性，并使已有的当前 BEV 到图像特征的跨空间相关性在当前 BEV 空间中继续更新；随后，通过空间交叉注意力继续更新当前 BEV 与多视角图像特征之间的跨空间关联。本文使用 $\bar{A}^{bb}$ 更新 R^{bb}，并结合由空间交叉注意力重建得到的 $\bar{A}^{bI}_{(c,m)}$，计算当前 BEV 到第 c 个相机视角第 m 层图像特征的跨空间相关性矩阵：

$$R^{bI}_{(c,m)} = R^{bI}_{(c,m)} + \left(\bar{R}^{bb}\right)^{\mathrm{T}} \bar{A}^{bI}_{(c,m)} \tag{11}$$

式中，$\bar{R}^{bb}$ 表示由时间自注意力当前分支累积并行归一化后的当前 BEV 空间自相关矩阵。

结合检测目标对应的 BEV 相关性向量 r_b，可得到待解释目标在第 c 个相机视角第 m 层图像特征上的图像归因结果：

$$r_I^{(c,m)} = r_b R^{bI}_{(c,m)} \tag{12}$$

其中，$r_I^{(c,m)}$ 表示对所有相机和特征层级分别计算该结果，即可分析检测结果所依赖的关键视角、尺度层级和局部图像区域。

2.6. 当前 BEV 特征到历史 BEV 特征的归因

基于 2.4 节得到的 BEV 归因向量 r_b，本文进一步分析待解释目标对历史 BEV 特征的依赖关系。由于 BEVFormer 通过时间自注意力融合历史 BEV 信息，仅分析图像特征来源难以说明历史状态和运动连续性在当前预测中的作用。同时，历史 BEV 的贡献不能直接等同于时间自注意力历史分支的原始注意力

权重，而需结合当前 BEV 分支中的累计相关性共同确定。

设历史 BEV 特征空间 H 、当前 BEV 特征到历史 BEV 特征的稠密正贡献矩阵和跨空间相关矩阵分别记为 A^{bh} 和 R^{bh} 。同时，沿用 2.5 节中当前 BEV 空间自相关矩阵 R^{bb} 及其对应的稠密正贡献矩阵 A^{bb} 。其中， A^{bh} 和 A^{bb} 均由 2.3 节的采样点级正贡献重建方法得到。沿用 2.1 节的初始化方式，令 $R^{bb}=I$ ，$R^{bh}=0$ 。

在时间自注意力中，当前分支用于更新当前 BEV 空间内部的相关性，并使已有的当前 BEV 到历史 BEV 的跨空间相关性在当前 BEV 空间中继续更新；历史分支则用于建立当前 BEV 与历史 BEV 特征之间的跨空间关联。本文使用 A^{bb} 更新 R^{bb} 及 R^{bh} ，并结合历史分支重建得到的 A^{bh} ，计算当前 BEV 到历史 BEV 的跨空间相关性矩阵：

$$R^{bh}=R^{bh}+\left(\bar{R}^{bb}\right)^{\mathrm{T}}\bar{A}^{bh} \tag{13}$$

式中， $\bar{R}^{bb}$ 表示由时间自注意力当前分支累积并行归一化后的当前 BEV 空间自相关矩阵。该式表明，历史 BEV 归因并非仅由历史分支原始注意力决定，而是由当前 BEV 分支中的累计相关性与历史分支正贡献共同确定。

结合检测目标对应的 BEV 相关性向量 r_b ，可得到待解释目标在历史 BEV 空间中的归因结果：

$$r_h=r_bR^{bh} \tag{14}$$

其中， r_h 表示历史 BEV 各特征单元对当前待解释检测结果的贡献强弱分布。

3. 实验

本章从定性可视化和定量忠诚度两个方面验证所构建层级化归因方法的有效性。定性实验围绕检测查询到当前 BEV、当前 BEV 到多视角图像特征以及当前 BEV 到历史 BEV 特征三类归因结果展开，用于观察本文方法能否定位目标相关 BEV 区域、关键摄像头视角和局部历史 BEV 区域。定量实验则通过历史 BEV 分支和图像特征分支上的特征删除式忠诚度实验，以及单分支全遮挡实验，分析归因结果与模型输出变化之间的一致性，并进一步比较图像特征与历史 BEV 特征在类别、几何属性和速度估计中的差异化支撑作用。

3.1. 实验设置

本文的可解释性实验基于 BEVFormer 检测框架开展。基础模型采用 BEVFormer-base 配置，并加载公开预训练权重进行推理与归因分析。实验数据来自 nuScenes 数据集，每个样本包含 6 个摄像头视角图像。BEV 空间分辨率设置为 200 × 200，共包含 40000 个 BEV 单元；解码器包含 900 个目标查询；时序输入队列长度设置为 3，以支持历史 BEV 的构建与时序信息融合。本文实验基于 PyTorch 深度学习框架，在配置 RTX A6000 GPU 的 Linux 平台上完成。

本文选取 Raw、Rollout 和 Grad-CAM 作为对比方法，并从可视化分析、特征删除忠诚度实验和单分支全遮挡实验三个方面进行验证。忠诚度实验选取 nuScenes 数据集中的 111 个待解释目标。对于历史 BEV 分支，按照归因得分逐步删除历史 BEV 特征；对于图像分支，将当前帧 6 个摄像头视角、4 个特征层级上的图像特征统一展开，并按归因得分排序后逐步删除。每一步扰动后均重新前向推理，并根据目标重匹配分数选择与原始目标最相似的检测结果，最终统计各一致性指标随删除比例变化的曲线及其 AUC。

3.2. 评价指标

参考文献[26]中的删除式评价思想，本文通过扰动前后检测结果的一致性变化评估归因结果的忠诚

度。正向删除优先移除高归因特征，用于检验高归因区域是否对应模型关键决策依据；负向删除优先移除低归因特征，用于检验归因方法对非关键区域的区分能力。考虑到三维检测结果同时包含类别、位置、尺寸、朝向和速度等属性，本文构建类别一致性 S_{cls}、中心一致性 S_{center}、尺寸一致性 S_{size}、朝向一致性 S_{yaw} 和速度一致性 S_{vel}，并进一步定义目标重匹配分数与综合一致性指标，用于完成扰动后目标跟踪与整体评价。

设原始待解释目标与扰动后候选检测框的类别概率向量分别为 P^* 和 P，中心坐标分别为 $c^*=\left(x^*,y^*,z^*\right)$ 和 $c=\left(x,y,z\right)$，尺寸分别为 $d^*=\left(w^*,l^*,h^*\right)$ 和 $d=\left(w,l,h\right)$，偏航角分别为 θ 和 θ^*，速度分别为 $v^*=\left(v_x^*,v_y^*\right)$ 和 $v=\left(v_x,v_y\right)$。各一致性指标定义为：

$$S_{cls}=\frac{p\cdot p^*}{\left\|p\right\|_2\left\|p^*\right\|_2} \tag{15}$$

$$S_{center}=\max\left(0,1-\frac{\left\|c-c^*\right\|_2}{\left\|d^*\right\|_2}\right) \tag{16}$$

$$S_{size}=\frac{\prod_{u\in\{w,l,h\}}\min\left(d_u^*,d_u\right)}{\prod_{u\in\{w,l,h\}}d_u^*+\prod_{u\in\{w,l,h\}}d_u-\prod_{u\in\{w,l,h\}}\min\left(d_u^*,d_u\right)} \tag{17}$$

$$S_{yaw}=\max\left(0,1-\frac{\min\left(\left|\theta-\theta^*\right|,2\pi-\left|\theta-\theta^*\right|\right)}{\pi}\right) \tag{18}$$

$$S_{vel}=\max\left(0,1-\frac{\left\|v-v^*\right\|_2}{\max\left(\left\|v^*\right\|_2,\tau_v\right)}\right) \tag{19}$$

其中，τ_v 为速度下界阈值，本文取 0.1 m/s。

为在每一步扰动后稳定匹配原始待解释目标，本文采用类别、中心和尺寸一致性的调和平均作为目标重匹配分数；在完成目标重匹配后，进一步采用类别、中心、尺寸、朝向和速度一致性的调和平均作为综合一致性指标：

$$S_{track}=\frac{3}{\frac{1}{S_{cls}}+\frac{1}{S_{center}}+\frac{1}{S_{size}}},S_{overall}=\frac{5}{\frac{1}{S_{cls}}+\frac{1}{S_{center}}+\frac{1}{S_{size}}+\frac{1}{S_{yaw}}+\frac{1}{S_{vel}}} \tag{20}$$

其中，重匹配阶段主要依据类别、中心和尺寸等目标身份相关属性，朝向和速度用于匹配后的属性一致性评价。本文统计各指标随删除比例变化的曲线，并计算曲线下面积(Area Under Curve, AUC)进行定量比较。对于正向删除，AUC 越小表示高归因特征被删除后检测结果下降越快；对于负向删除，AUC 越大表示低归因特征被删除后模型输出越稳定。

3.3. 可视化结果分析

3.3.1. 检测查询到 BEV 特征的归因结果

图 3 展示了检测查询到当前 BEV 特征的归因结果，包含真实检测框以及不同方法得到的 BEV 归因热力图。可以看出，不同样例中的高响应区域均主要集中在目标框及其邻域，而远离目标的背景区域响应较弱，说明 BEVFormer 在形成检测结果时主要依赖目标相关的局部 BEV 单元，而非均匀利用整个 BEV 空间。

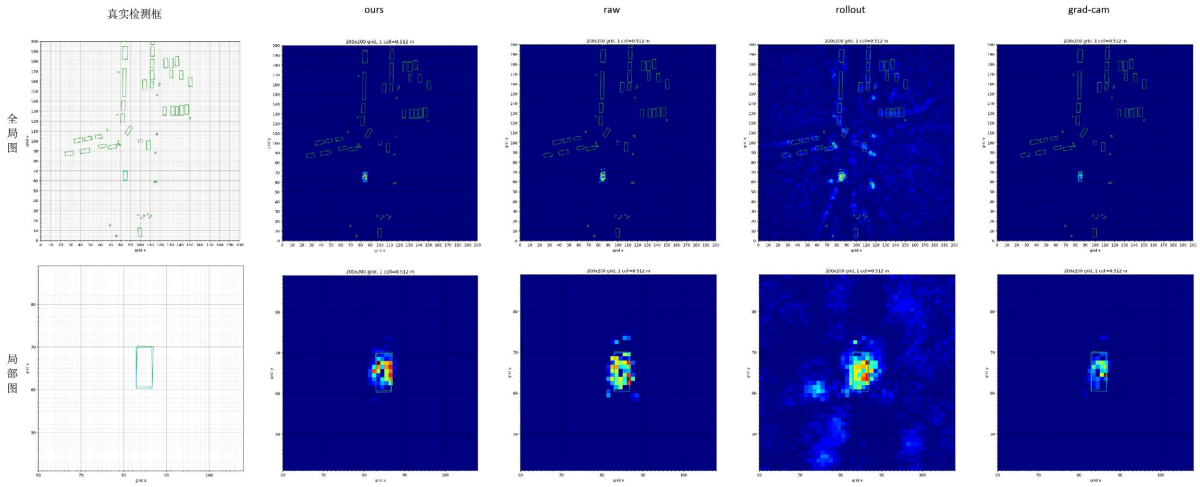


Figure 3. Visual comparison of attribution results for the current BEV features detected by the query
图 3. 检测查询到当前 BEV 特征的归因结果可视化对比

由图 3 可见，不同方法均能在一定程度上定位目标邻域，但可视化形态存在差异。Rollout 的响应范围较大，在背景区域存在明显扩散；Raw 和 Grad-CAM 能够定位目标位置，但仍存在一定局部离散或背景响应。相比之下，本文方法的热点更集中于目标框及其邻近区域，目标相关区域与非关键区域之间的区分更加清晰，表明其能够较有效地刻画检测结果在当前 BEV 空间中的关键支撑区域。

3.3.2. 当前 BEV 到多视角图像特征的归因结果

图 4 展示了当前 BEV 到多视角图像特征的归因结果。可以看出，当前检测结果对多视角图像信息的利用具有明显的视角选择性：当目标在某一视角中清晰可见时，该视角上的归因响应更显著；而在目标不可见或仅局部可见的视角中，响应明显较弱。这说明 BEVFormer 并非平均依赖所有摄像头视角，而是主要利用能够提供目标直接观测信息的关键视角。

Figure 4. Visual comparison of current BEV attribution results to multi-view image features
图 4. 当前 BEV 到多视角图像特征的归因结果可视化对比

Figure 5. Enlarged view of image feature attribution results from a key perspective
图 5. 键视角下图像特征归因结果局部放大图

图 5 给出了左后视角上的局部放大结果。由图 5 可见，不同方法均能反映关键视角对检测结果的作用，但响应范围存在差异。Rollout 在多个视角中均出现较明显激活，表现出更强的跨视角扩散；Raw 和 Grad-CAM 能够关注到目标所在视角，但仍存在一定背景响应。相比之下，本文方法的归因响应更集中于目标清晰可见的关键视角及其邻近区域，更有利于揭示当前检测结果所依赖的核心图像证据。

3.3.3. 当前 BEV 到历史 BEV 特征的归因结果

图 6 展示了当前 BEV 到历史 BEV 特征的归因结果。可以看出，不同样例中的高响应区域主要集中在当前目标对应的局部历史 BEV 区域，而非均匀分布于整幅历史 BEV 空间，说明 BEVFormer 对历史信息的利用具有明显的目标相关性和空间选择性。该结果表明，历史 BEV 分支并非简单提供全局背景补充，而是主要通过目标邻域的历史表征辅助当前检测结果形成。

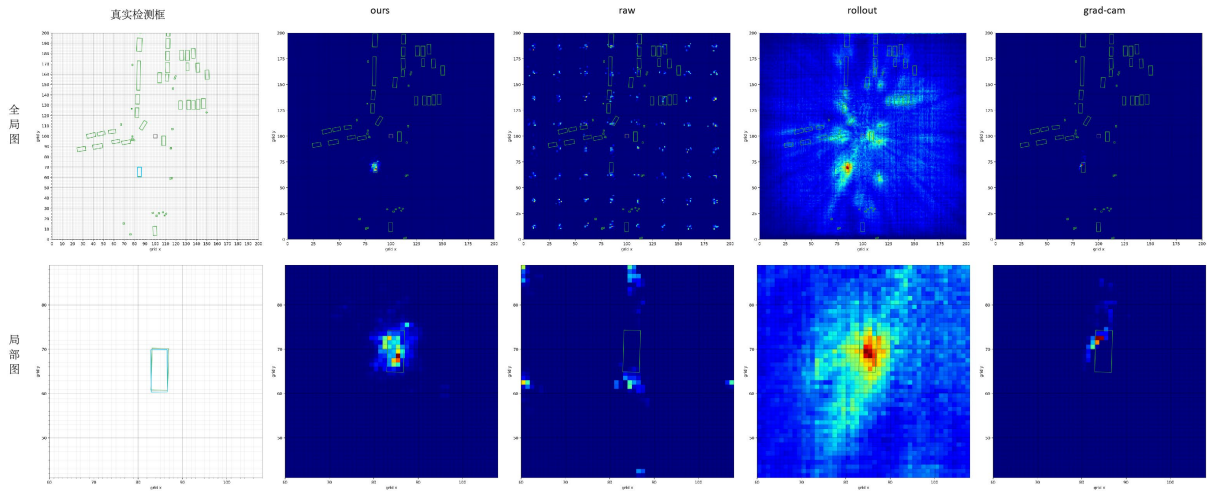


Figure 6. Visual comparison of attribution results from current BEV features to historical BEV features
图 6. 当前 BEV 特征到历史 BEV 特征的归因结果可视化对比

从不同方法对比来看，各方法均能反映当前检测结果对目标邻域历史 BEV 特征的依赖，但响应形态存在差异。Rollout 的归因范围更大，存在明显扩散；Raw 结果相对离散；Grad-CAM 更偏向局部敏感区域。相比之下，本文方法在目标邻域内响应更集中且局部结构更连续，说明其能够更稳定地刻画历史 BEV 对当前检测结果的局部支撑作用。

3.4. 忠诚度验证与分支贡献分析

为定量验证归因结果的忠诚度，本文分别在历史 BEV 特征分支和图像特征分支上进行删除式扰动实验。历史分支实验仅扰动历史 BEV 特征，图像分支实验仅扰动多视角图像特征；正向删除优先移除高归

因特征，负向删除优先移除低归因特征。本文对 111 个待解释目标的一致性曲线逐点求均值，并计算平均曲线的 AUC 作为定量结果。

Table 1. Historical BEV branch and image feature branch positive deletion experiment AUC results
表 1. 历史 BEV 分支与图像特征分支正向删除实验 AUC 结果

扰动特征	方法	类别一致性↓	中心一致性↓	尺寸一致性↓	偏航角一致性↓	速度一致性↓
历史 BEV 特征	Ours	0.9908	**0.8620**	**0.9101**	**0.9484**	**0.4242**
	Raw	0.9940	0.9185	0.9486	0.9708	0.6899
	Rollout	**0.9896**	0.8682	0.9132	0.9516	0.4273
	Grad-cam	0.9933	0.9034	0.9388	0.9641	0.6340
图像特征	Ours	**0.9956**	**0.8853**	**0.9464**	0.9645	0.8890
	Raw	0.9963	0.8934	0.9495	0.9710	0.8977
	Rollout	0.9971	0.8970	0.9476	0.9690	0.8905
	Grad-cam	0.9970	0.9025	0.9508	**0.9638**	**0.8789**

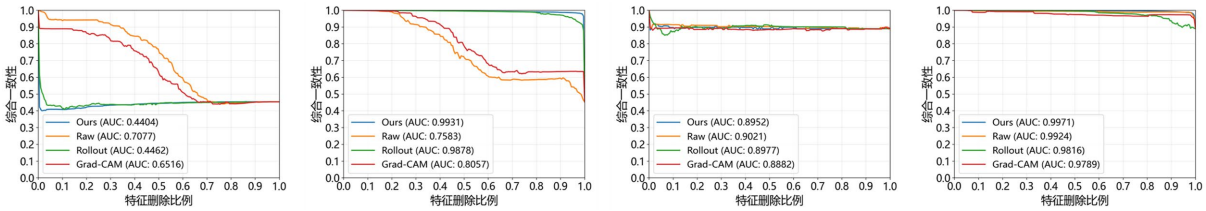
Table 2. Historical BEV branch and image feature branch negative deletion experiment AUC results
表 2. 历史 BEV 分支与图像特征分支负向删除实验 AUC 结果

扰动特征	方法	类别一致性↑	中心一致性↑	尺寸一致性↑	偏航角一致性↑	速度一致性↑
历史 BEV 特征	Ours	**1.0000**	**0.9960**	**0.9924**	**0.9966**	**0.9870**
	Raw	0.9970	0.9268	0.9559	0.9740	0.7449
	Rollout	0.9999	0.9922	0.9903	0.9959	0.9814
	Grad-cam	0.9996	0.9501	0.9654	0.9844	0.7929
图像特征	Ours	1.0000	**0.9974**	**0.9962**	**0.9983**	**0.9950**
	Raw	1.0000	0.9924	0.9913	0.9954	0.9862
	Rollout	0.9999	0.9821	0.9891	0.9958	0.9777
	Grad-cam	1.0000	0.9866	0.9885	0.9897	0.9704

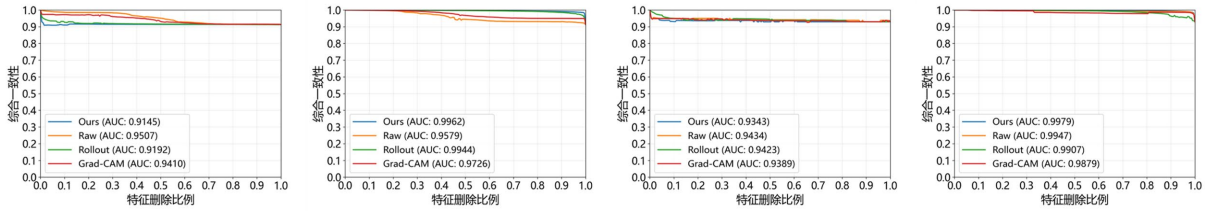
从表 1 可以看出，在正向删除实验中，本文方法在历史 BEV 特征扰动下的中心、尺寸和速度一致性 AUC 整体较低，说明其高归因历史 BEV 区域对检测结果具有较强影响。从表 2 可以看出，在负向删除实验中，本文方法在历史 BEV 和图像特征扰动下均保持较高 AUC，说明低归因区域被删除时模型输出整体较稳定。正向删除下降快、负向删除保持稳定，表明本文方法能够较好地区分关键特征与非关键特征。

Table 3. Consistency statistics under single-branch feature preservation conditions
表 3. 单分支特征保留条件下的一致性统计

保留特征	类别一致性	中心一致性	尺寸一致性	偏航角一致性	速度一致性
图像特征	0.9893	0.8618	0.9095	0.9408	0.4323
历史 BEV 特征	0.9952	0.8854	0.9449	0.9586	0.8839



(a) 引入速度一致性项后的综合一致性曲线



(b) 未引入速度一致性项的综合一致性曲线

Figure 7. The overall consistency curve under the perturbation of historical BEV branch and image feature branch
图 7. 历史 BEV 分支与图像特征分支扰动下的综合一致性曲线

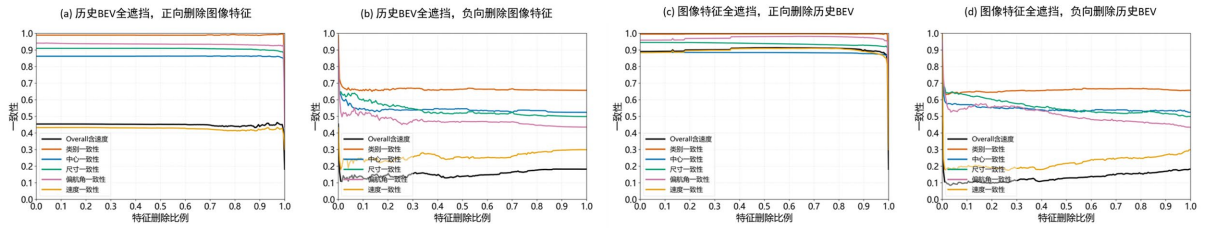


Figure 8. Consistency curves for positive and negative deletion under single-branch retention conditions
图 8. 单分支保留条件下的正负向删除一致性曲线

由表 3 可见，当仅保留图像特征时，类别、中心、尺寸和偏航角一致性仍保持较高水平，但速度一致性明显降低；当仅保留历史 BEV 特征时，速度一致性显著提高，且几何一致性整体更稳定。进一步结合图 7 可知，当综合一致性包含速度项时，历史 BEV 分支正向删除曲线下降更明显；而不包含速度项时，各曲线整体保持较高水平，说明速度一致性是区分历史 BEV 分支与图像特征分支作用差异的重要指标。

图 8 进一步表明，在一个分支被完全遮挡后，继续扰动另一个保留分支时，多项一致性曲线的下降幅度较图 7 更明显。这说明在双分支同时存在时，图像特征与历史 BEV 特征之间存在一定跨分支补偿；而当另一分支被完全遮挡后，保留分支中的关键特征被进一步删除，模型输出更容易受到影响。同时，图 8 中正向删除曲线整体低于负向删除曲线，说明保留分支中的高归因区域仍比低归因区域对检测结果具有更大影响。由此可知，BEVFormer 对图像特征和历史 BEV 特征的利用并非简单替代关系，而是在类别、几何属性和运动状态估计中表现出一定信息冗余与互补分工。

需要说明的是，当完全删除输入特征时，一致性并不必然降为 0。这是因为本文并非以二值方式判断原始目标是否完全消失，而是在每次扰动后重新前向推理，并基于扰动后的候选检测结果与原始目标之间的相似度进行评价。由于扰动后的预测集合中仍可能存在与原始目标在类别、位置、尺寸或运动属性上部分相似的候选框，因此各一致性指标可能保留非零值。

4. 结论

针对 BEVFormer 在多摄像头空间信息与历史 BEV 时序信息融合过程中的决策依据不透明问题，本

文基于 GAE 相关性传播思想，构建了面向 BEVFormer 的层级化归因解释方法，并通过采样点级正贡献重建使相关性传播能够适配其可变形注意力结构。主要结论如下。

1) 通过建立当前 BEV、图像特征和历史 BEV 三个层面的归因关系，并将可变形注意力中的采样点级正贡献映射至规则 BEV 网格或图像特征空间，本文方法能够为 BEVFormer 检测结果的空间来源与时序来源分析提供可行方式。

2) 可视化结果表明，BEVFormer 的检测结果具有明显的目标相关性和空间选择性。其关键响应主要集中于目标邻域 BEV 区域、目标可见性较强的摄像头视角以及目标相关的局部历史 BEV 区域。

3) 分支扰动实验与单分支全遮挡实验表明，图像特征与历史 BEV 特征在 BEVFormer 中并非简单替代关系，而是表现出一定的信息冗余与功能分工。两类特征均能够为目标类别、位置、尺度和朝向等属性提供支撑，但历史 BEV 特征在速度估计和运动连续性保持方面表现出更突出作用。BEVFormer 的检测结果并非由单一分支独立决定，而是由当前图像观测与历史 BEV 时序信息共同支撑。

4) 本文方法可为其他基于可变形注意力的感知模型提供参考，但其推广需要结合具体模型结构进行调整。对于能够获得采样权重、采样位置，并能够计算目标输出对采样权重梯度的模型，可参照本文方法构建采样点级正贡献，并映射至相应规则特征空间；但不同模型的查询定义、特征组织方式和跨空间融合路径存在差异，不能直接套用 BEVFormer 中的层级归因链路。此外，本文主要关注注意力模块中的正向支持性贡献，尚未完整刻画非注意力模块及抑制性负贡献，后续可进一步结合正负贡献分解和守恒式传播规则加以扩展。

参考文献

[1] Mao, J.G., Shi, S.S., Wang, X.G., *et al*. (2023) 3D Object Detection for Autonomous Driving: A Comprehensive Survey. *International Journal of Computer Vision*, **131**, 1909-1963. https://doi.org/10.1007/s11263-023-01790-1

[2] Roddick, T., Kendall, A. and Cipolla, R. (2019) Orthographic Feature Transform for Monocular 3D Object Detection. *Proceedings of the British Machine Vision Conference*, Cardiff, 9-12 September 2019, Article No. 285.

[3] Philion, J. and Fidler, S. (2020) Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In: Vedaldi, A., *et al*., Eds., *Proceedings of the European Conference on Computer Vision*, Springer International Publishing, 194-210. https://doi.org/10.1007/978-3-030-58568-6_12

[4] Huang, J.J., Huang, G., Zhu, Z., *et al*. (2021) BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View. https://arxiv.org/abs/2112.11790

[5] Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., *et al*. (2023) BEVDepth: Acquisition of Reliable Depth for Multi-View 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 1477-1485. https://doi.org/10.1609/aaai.v37i2.25233

[6] Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J. and Li, Z. (2023) BEVStereo: Enhancing Depth Estimation in Multi-View 3D Object Detection with Temporal Stereo. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 1486-1494. https://doi.org/10.1609/aaai.v37i2.25234

[7] Huang, J.J. and Huang, G. (2022) BEVDet4D: Exploit Temporal Cues in Multi-Camera 3D Object Detection. https://arxiv.org/abs/2203.17054

[8] Li, Z.Q., Wang, W.H., Li, H.Y., *et al*. (2022) BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In: Avidan, S., *et al*., Eds., *Proceedings of the European Conference on Computer Vision*, Springer, 1-18. https://doi.org/10.1007/978-3-031-20077-9_1

[9] Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Workshop at International Conference on Learning Representations*, Banff, 14-16 April 2014, 1-8. https://doi.org/10.48550/arXiv.1312.6034

[10] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/iccv.2017.74

[11] Sundararajan, M., Taly, A. and Yan, Q.Q. (2017) Axiomatic Attribution for Deep Networks. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 3319-3328.

[12] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. and Samek, W. (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, **10**, e0130140. https://doi.org/10.1371/journal.pone.0130140

[13] Shrikumar, A., Greenside, P. and Kundaje, A. (2017) Learning Important Features through Propagating Activation Differences. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 3145-3153.

[14] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[15] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. In: *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Curran Associates, 4765-4774.

[16] Xu, K., Ba, J., Kiros, R., *et al*. (2015) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the* 32*nd International Conference on Machine Learning*, Lille, 6-11 July 2015, 2048-2057.

[17] Choi, E., Bahadori, M.T., Sun, J., *et al*. (2016) RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. In: *Proceedings of the* 30*th International Conference on Neural Information Processing Systems*, Curran Associates, 3504-3512.

[18] Jain, S. and Wallace, B.C. (2019) Attention Is Not Explanation. In: *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Association for Computational Linguistics, 3543-3556.

[19] Abnar, S. and Zuidema, W. (2020) Quantifying Attention Flow in Transformers. In: *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 4190-4197. https://doi.org/10.18653/v1/2020.acl-main.385

[20] Chefer, H., Gur, S. and Wolf, L. (2021) Transformer Interpretability Beyond Attention Visualization. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 782-791. https://doi.org/10.1109/cvpr46437.2021.00084

[21] Chefer, H., Gur, S. and Wolf, L. (2021) Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), 11-17 October 2021, 397-406. https://doi.org/10.1109/iccv48922.2021.00045

[22] Ali, A., Schnake, T., Eberle, O., *et al*. (2022) XAI for Transformers: Better Explanations through Conservative Propagation. *Proceedings of the* 39*th International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 435-451.

[23] Ferrando, J., Gállego, G.I., Tsiamas, I. and Costa-Jussà, M.R. (2023) Explaining How Transformers Use Context to Build Predictions. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics*, Volume 1, 5486-5513. https://doi.org/10.18653/v1/2023.acl-long.301

[24] Achtibat, R., Hatefi, S.M.V., Dreyer, M., *et al*. (2024) AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for Transformers. *Proceedings of the* 41*st International Conference on Machine Learning*, Vienna, 21-27 July 2024, 135-168.

[25] Arras, L., Puri, B., Kahardipraja, P., *et al*. (2025) A Close Look at Decomposition-Based XAI-Methods for Transformer Language Models. https://arxiv.org/abs/2502.15886

[26] Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V.I., Mehra, A., Ordonez, V., *et al*. (2021) Black-Box Explanation of Object Detectors via Saliency Maps. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 11443-11452. https://doi.org/10.1109/cvpr46437.2021.01128