基于文本的少样本域自适应视觉融合事件检测

Farhad Moghimifar、Fatemeh Shiri、Reza Haffari、Yuan-Fang Li数据科学与人工智能系 信息技术学院 莫纳什大学，澳大利亚墨尔本
firstname.lastname@monash.edu

Van Nguyen国防科学技术集团
信息科学部澳大利亚阿德莱德
Van.Nguyen5@defence.gov.au

arXiv:2305.03517v2 [cs.CL] 5 Jun 2023

摘要——**近年来，将图像等辅助模态融入事件检测模型的研究日益受到关注。自然语言在描述情境时具有复杂性，这促使研究人员利用相关的视觉上下文来提升事件检测性能。然而，该领域现有的方法面临数据稀缺问题，即模型训练需要大量标注文本图像对。此外，在推理时获取视觉上下文的机会有限，也会对这类模型的性能产生负面影响，使其在现实世界场景中的实际应用效果不佳。本文提出了一种新颖的领域自适应视觉融合事件检测方法，该方法仅需少量标注的图像-文本配对数据即可完成训练。具体而言，我们引入了一种视觉想象生成方法，能够在缺乏视觉上下文的情况下根据文本合成图像。此外，该想象生成器还可以针对特定领域进行定制。通过这种方式，我们的模型能够充分利用预训练视觉语言模型的能力，并在少样本设置下完成训练。这也使模型能够在仅有单模态数据（即文本）可用的情况下进行有效推理。在基准 M2E2 数据集上进行的实验评估表明，与现有最先进模型相比，我们的模型性能最高可提升 11 个百分点。**

索引词—**事件检测、多模态事件检测、少样本学习、图像生成**

I. 引言

事件检测是自然语言处理（NLP）中的事件抽取子任务，旨在从文本中抽取有关事件的信息。这项任务的主要挑战在于自然语言的复杂性。多年来，人们提出了各种各样的方法来解决这一任务，从基于规则的方法 [1]–[4] 到端到端神经网络模型 [5]–[9] 。尽管这些技术在现有标准基准测试中取得了令人鼓舞的结果，但它们严重依赖单一模态，即文本（图 2(a)），未能考虑其他相关模态中的现实世界依据。

事实证明，利用图像等其他模态引入显著信息，以补充文本中包含的内容，在机器翻译 [10]–[12]等其他领域中是有效的。然而，针对事件检测开发多模态方法仍是一个探索不足的领域。以 M2E2 数据集 [13]中的示例图 1 为例。输入文本中的“拘留”一词可能表明发生了逮捕：监禁 事件类型，因为它暗示了逮捕某人的行为。这一解释得到了



图 1. M2E2 数据集 [13]中的一则多媒体 VOA 新闻示例。本文中的文本输入报道了一起“逮捕：监禁”类型的事件，并由视觉上下文加以补充。该事件的触发词是“拘留”，配图中的边界框标出了被带去拘留的男子。

图像中蕴含的视觉特征的补充。通常，将某种情境的底层视觉表示纳入其中，可以提高事件检测方法的数据效率，并有助于发现难以用语言描述或被遗漏的事件。

开发此类多模态事件检测模型的主要挑战之一，是缺乏大量标注数据，这对于基于大型语言模型的微调技术尤为关键 [13] （图 2(b)）。这一问题阻碍了其在现实世界中的实用性，因为在现实场景中，领域特定的多模态数据（即配对文本和图像）可能十分稀缺。此外，在现实场景中获取视觉上下文可能受到限制，这也会影响多模态事件检测模型在推理时的性能。

为克服这些局限，我们提出了一种少样本视觉融合事件检测模型，能够有效

应对训练和推理过程中数据稀缺的问题。在训练期间，针对只能获取数量有限的带标签图文对这一问题，我们采用少样本迁移学习模型。我们利用预训练大型视觉语言模型初始化模型参数，并对其进行更新，使其适应目标数据的分布，从而更有效地完成目标任务。

此外，我们还提出了一种定制化的 *Visual Imaginator*模型，该模型能够仅利用训练期间的少量示例，学习合成文本输入的视觉表征。该方法缩小了预训练模型所基于的数据与我们的目标数据集之间的领域偏移，使我们提出的 Visual Imaginator 能够生成定制化的领域特定图像，以补充文本输入。该方法有助于我们应对只能获取单模态数据（即文本）的挑战。我们设想，在推理时，我们的模型可以利用合成生成图像替代缺失的实际图像，从而执行多模态事件检测。因此，我们提出的事件检测方法能够达到与文本和图像均完整可用的多模态模型相当的准确率。

总而言之，我们的贡献如下：

- 我们将多模态事件检测任务表述为一个小样本学习问题，并提出一种新颖的视觉融合事件检测框架，以缓解数据稀缺问题。
- 我们提出的定制化 Visual Imaginator 模块能够利用文本输入合成领域特定图像，以应对推理时缺失文本-图像对的挑战。
- 在事件检测标准基准数据集上的实验，证实了我们的模型在低资源和少样本场景下优于最先进的模型。

II. 背景

我们的方法与三个研究方向相关。

A. 多模态事件检测

大多数事件抽取方法从单模态数据（主要是文本）中抽取事件信息，并忽略视觉媒体的贡献 [14]–[17]。然而，在现实场景中，数据通常是多模态的。例如，新闻文章、政策/法规文件、情报报告和社交媒体帖子通常包含文本及配套图像。一些事件抽取方法通过检索额外的相关视觉知识，取得了令人鼓舞的成果 [18], [19]。

直到最近，人们才开始研究从多模态数据中进行事件抽取的问题 [13], ，但这仍是一个尚未得到充分探索的领域。 [13] 提出了一种多媒体结构化公共空间构建方法，利用现有的图像-标题对和单模态标注数据进行弱监督训练。 [19] 表明，纳入视觉上下文能够提升抽取性能。 [18] 提出了一种双循环多模态模型，用于深入交互图像与句子，以实现模态特征聚合。 [20]提出了一种基于预训练视觉语言模型的方法 [21]，用于解决多模态事件抽取问题。然而，这些方法都需要完整的文本-图像对集合进行训练。此外，这些模型还受到数据稀缺问题的困扰，在低资源场景下尤为突出。与这些方法不同，我们提出了一种能够合成视觉上下文的模型，以缓解低资源数据问题。为了解决歧义，可以借助视觉线索来澄清含义。我们提出了一种利用视觉信息的自动事件检测方法。

B. 视觉-语言预训练模型

近年来，视觉-语言预训练模型得到了广泛研究 [22]–[28]。研究表明，场景图等视觉结构信息 [29] 对这些模型的预训练很有帮助。然而，事件结构知识在预训练模型中并未得到充分捕捉，导致与动词理解相关的任务存在不足 [30]。基于视觉-语言预训练模型 Clip [21], [20] ，编码结构化的事件语义和知识，以增强视觉-语言预训练。

C. 文本到图像生成

从文本生成图像已得到广泛研究[31]–[33]。具有代表性的研究采用 GAN [33]–[37] 来合成照片级真实感场景，使其在语义层面高度忠实于作为条件输入的文本描述。DALL-E [31]提出了一种结合离散 VAE [38] 的自回归 Transformer，用于根据文本生成图像，涵盖自然语言所能表达的广泛概念。 [39]提出通过建立扩散模型与变分推理之间的联系，利用扩散模型生成高质量图像。尽管我们的方法受到这些研究的启发，但本研究的目标是合成离散视觉表征，以改进事件检测，而非生成高质量的照片级真实感图像。

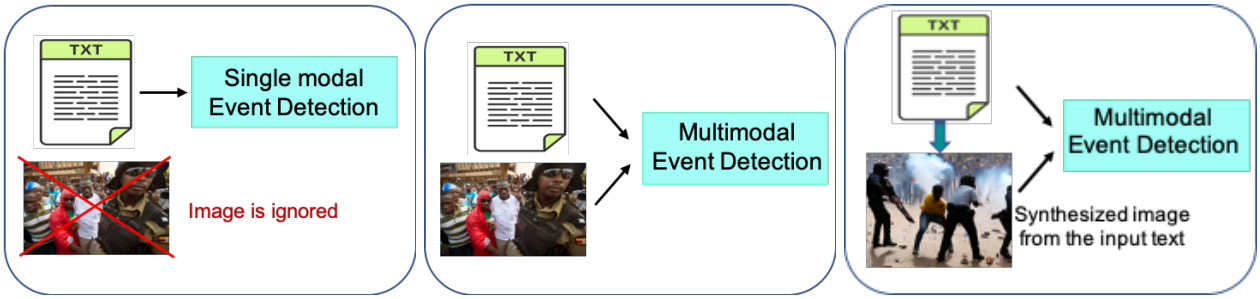
III. 方法

A. 架构

图3展示了我们的视觉融合事件检测（VF-Event）框架的架构，该框架由两个关键组件构成：（1）视觉想象模块；（2）预训练视觉语言模块。定制的想象模块能够解决多模态事件检测的挑战，而无需在推理时使用图像。

B. 任务表述

我们的模型学习在给定支持集 $\mathbb{S}$ 和一组预定义事件类型 $\mathcal{E}$的情况下，从查询实例 s 中提取事件类型；该事件类型也出现在支持集 $\mathbb{S}$中。我们采用 N+1 类、 K 样本的设置，其中有 N 个簇分别表示 N 种事件类型，另有一个簇表示无 事件类型，每个簇包含K 个数据点。此外，还有相应的视觉



(a) (b) (c)

图 2. 对比 (a) 常规纯文本事件检测、(b) 多模态监督事件检测，以及 (c) 我们提出的端到端少样本视觉融合事件检测模型。

上下文 $\mathbb{V}$，它们与 $\mathbb{S}$中的实例配对。支持集表示如下：

$$\begin{aligned}\mathbb{S} = \{&(s_1^1, v_1^1, e_1), \ldots, (s_1^K, v_1^K, e_1),\\ &\ldots\\ &(s_N^1, v_N^1, e_N), \ldots, (s_N^K, v_N^K, e_N),\\ &(s_{N+1}^1, v_{N+1}^1, e_{none}), \ldots, (s_{N+1}^K, v_{N+1}^K, e_{none})\}\end{aligned} \tag{1}$$

其中：

- 每个 $s_i^j = (w_1, w_2, \ldots)$ 表示事件类型 e_i 的支持集中的第 j 个实例，而 w_m 是词汇表 $\mathcal{W}$ 中的一个词元。
- $\mathbb{V} = \{v_1^1, \ldots, v_1^K, \ldots, v_N^K\}$ 是视觉上下文。$v_i^j \in \mathbb{V}$是支持集中与第 j个事件类型实例配对的图像e_i 。
- $\mathcal{E} = \{e_1, \ldots, e_N, e_{none}\}$ 是标签集合，表示所有可能事件类型的集合，而 e_{none} 是非事件的特殊标签。

C. 事件检测

我们将事件检测任务表述为多模态分类任务，在该任务中，我们提出的模型会输出所有可能事件类型 $\mathcal{E}$中概率最高的事件类型 e 。

编码器 **ENC(·)** 由文本编码器（$\mathbf{ENC}_L$）和视觉编码器（$\mathbf{ENC}_V$）组成。$\mathbf{ENC}_L$ 计算输入实例 s的隐藏表示（H_s），而$\mathbf{ENC}_V$计算输入图像 v的视觉隐藏表示（H_v）。

因此，输入（s, v）的隐藏表示（H）可表示为：

$$H = \mathbf{ENC}(s, v) = [\mathbf{ENC}_L(s); \mathbf{ENC}_V(v)] \tag{2}$$

FNN(·) 是一个带有 softmax 层的前馈神经网络，用于计算可能事件类型上的概率分布：

$$P_{\theta_{class}}(e|s, v) = \mathbf{FNN}_{\theta_{FNN}}(w_{class}.\mathbf{ENC}_{\theta_{ENC}}(s, v) + b_{class}), \tag{3}$$

其中，$\mathbf{ENC}_{\theta_{\mathbf{ENC}}} = (\mathbf{ENC}_L, \mathbf{ENC}_V)$ 表示编码器结构，$\mathbf{FNN}_{\theta_{\mathbf{FNN}}}$ 表示前馈神经网络。$\theta_{\mathbf{class}} := \{\theta_{\mathbf{ENC}}, \theta_{\mathbf{FNN}}\}$ 表示我们的分类模型的参数。最后，我们通过以下方式计算预测的事件类型：

$$\widehat{e} = argmax_{e \in \mathcal{E}} P_{\theta_{class}}(e|s, v), \tag{4}$$

给定支持集 $\mathbb{S}$，可通过最小化交叉熵损失来训练 VF-Event 架构：

$$\mathcal{L}_{class} = \mathbb{E}_{s,v,e}[-\log P(e|s, v; \theta_{class})] \tag{5}$$

D. 少样本设置下的训练

训练期间，可获得一个由与图像配对的实例组成的支持集，每种事件类型包含 K 个样例；VF-Event 模型会输出概率最高的事件类型（图3，中间）。

推理时，如果没有视觉输入，VF-Event 会依靠 Visual Imaginator 根据查询文本 x合成一张图像。然后，我们使用图像编码器计算合成图像的视觉隐层表示 $H_{\widehat{v}}$ 。接着，我们基于少样本数据调整视觉想象。

我们的视觉想象模块基于去噪扩散概率模型 [39]（图3，顶部）。该生成模型由一个预训练自编码器、一个对应的解码器和一个扩散模型组成：预训练自编码器将图像映射为空间潜在编码；对应的解码器学习将潜在表示映射回图像；扩散模型则以文本输入 (s) 为条件。受 [40], ，启发，在推理时，我们使用支持集中的 (s，v) 文本输入作为条件，引导模型重新生成v。此外，我们使用输入文本 s 来合成视觉上下文 $\widehat{v}$。随后，我们采用以下基于重建的损失函数，在合成图像 $\widehat{v}$ 上训练视觉想象模块：

$$\mathcal{L}_{visual} = \mathbb{E}_{s,v,\widehat{v},\epsilon}[\omega||F_{\theta_v}(\alpha v + \sigma\epsilon, s) - \widehat{v}||_2^2] \tag{6}$$

其中，F_{θ_v} 表示条件视觉想象。α 和 σ是控制噪声调度的项。v 是来自支持集的真实图像。$\epsilon \sim \mathcal{N}(0, I)$是一个噪声项。视觉想象模块中唯一可训练的参数是文本编码器的参数，其余参数均保持 Frozen。

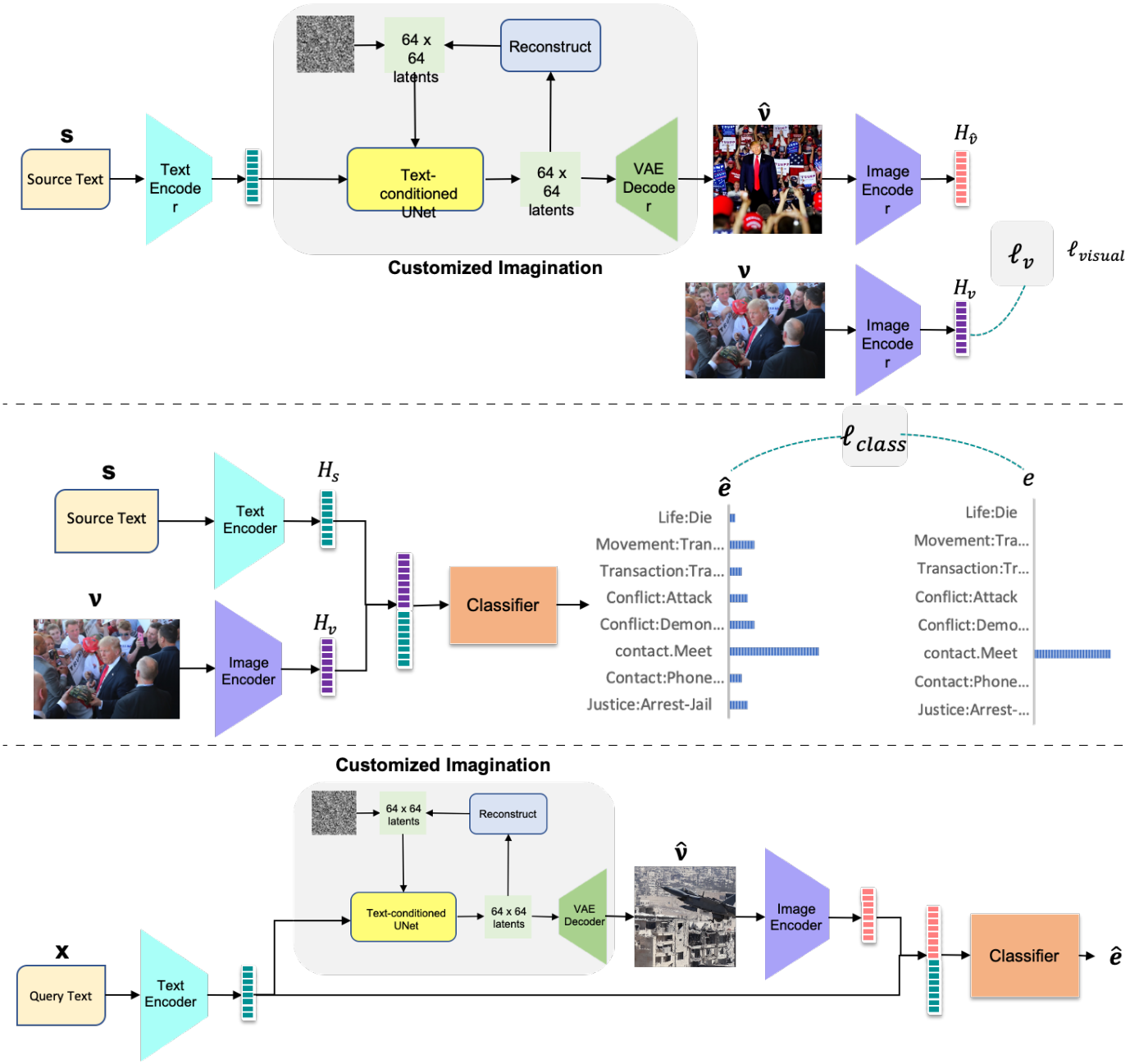


图 3. 我们的 VF-Event 框架示意图。**顶部：**利用支持集中的样本定制想象模块。**中部：**利用支持集对视觉语言模块进行微调。**底部：**在缺少视觉输入的情况下，VF-Event 的推理过程。

在这项工作中，我们结合公式 5 和 6 中的损失函数来训练模型。我们使用超参数 β控制额外视觉损失的贡献。

$$\mathcal{L} = \mathcal{L}_{class} + \beta \mathcal{L}_{visual}. \tag{7}$$

IV. 实验

本节介绍用于评估我们模型 VF-Event 性能的实验设置，并展示将 VF-Event 与基线模型进行比较的实验结果。

A. 实验设置

数据集。 我们在 M2E2 数据集 [13]，上进行了实验；该多模态新闻数据集在 ACE [43]的基础上进行了扩展。该数据集包含 6,167 个句子，其中有 1,297 个事件提及，涵盖 8 种事件类型：Move- ment.Transport、冲突·攻击、冲突·示威、Jus- tice.ArrestJail、联系·电话书信、联系·会面、生命·死亡和交易·转账。为了评估模型在训练数据有限时的性能，我们采用了少样本设置，仅在训练期间使用 K （5、10、15 和 20）个同时包含文本和图像的数据点

表 I M2E 数据集上少样本多模态事件检测的实验结果。结果分别基于 5、10、15 和 20 个训练样本进行报告。

方法	5 样本			10 样本			15 样本			20 样本		
	F1	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率
Text2Event [14]	6.59	27.27	3.75	30.38	30.77	30.00	43.04	43.59	42.5	56.25	54.22	58.44
Valhalla [41]	16.36	20.79	15.55	17.52	18.63	17.77	26.89	37.39	28.88	24.44	29.74	25.28
Ferozen [42]	40.36	43.26	41.11	42.01	43.21	42.22	43.19	44.33	43.33	46.11	47.11	45.97
Clip [21]	47.93	50.15	50	50.38	55.7	**50**	52.06	54.78	52.22	60.06	62.65	60.91
VF-Event	**50.00**	**50.94**	**50**	**50.58**	**60.95**	48.88	**63.79**	**70.14**	**62.22**	**65.23**	**67.6**	**64.36**

阶段。此外，在推理期间，模型仅接收文本输入，配对的图像则被丢弃。

评估指标。与以往处理该任务的工作一致 [13]，，结果采用宏平均精确率（P）、召回率（R）和 F 值（F-1）指标报告。

基线模型。为通过实验比较验证所提出的模型，我们选取了以下多模态事件检测模型作为基线：

- Text2Event [14]: 一种纯文本基线模型，使用预训练语言模型（T5）进行训练，不包含任何视觉信息。
- Valhalla [41]: 一种用于机器翻译的多模态序列到序列模型，在测试时利用视觉幻觉生成。我们通过训练模型仅生成事件类型来训练 Valhalla。
- Frozen [42]: 一种通过扩展前缀微调的软提示理念，将大型语言模型转变为多模态少样本学习系统的方法。
- Clip [21]: 一种近期模型，基于对比语言-图像预训练，能够学习多模态嵌入空间，可用于估算给定文本与图像之间的语义相似度。

实验设置。 我们使用预训练的 Stable Diffusion 模型 [44] 的参数初始化 Visual Imaginator。文本和图像编码采用 CLIP ViT-L/14 [21] ，嵌入维度为 768，并在文本编码器上使用 0.3 的 dropout。我们使用 Hugging Face Transformer 库 [1] 实现 Visual Imaginator 和模型的编码器。优化采用 Adam [45] ，学习率为 $2e-5$。我们使用了一个 1 层前馈神经网络，并将 β 设为 0.01。模型训练了 50 轮，批次大小为 4。对于基线模型，我们使用了作者提供的源代码和实现细节 [2]。所有模型的数据均以相同方式准备。

B. 观察与讨论

表 I 展示了各模型在 M2E2 数据集上的性能，其中训练时仅提供了 K 个数据点同时包含两种模态。通过事件类型检测的 F1 值、精确率和召回率可以看出我们模型的优越性。在 F1 值方面，与纯文本模型 Text2Event [14] 相比，我们的模型在 5 样本设置下提升了 43%，在 20 样本设置下提升了 9%，事件类型检测性能显著提高。我们的 VF-Event 模型在全部四种少样本设置下均显著优于纯文本基线，证明了视觉想象对于纯文本任务的有效性。结果还表明，当训练数据集数量较少（即 5 个）时，单模态纯文本模型无法从事件中提取信息；而在推理过程中融合视觉信息，则会导致事件类型检测效果不佳。如表 I 所示，我们提出的方法在所有对比方法中均表现更优，在不同少样本设置下，除一项指标（10 样本设置下的召回率）外，均取得了最佳精确率、召回率和 F1 值。

总而言之，这些一致的改进清楚地表明，在测试时，视觉想象能够有效利用源句中的视觉表征，从而提升事件检测效果。

C. Visual Imaginator 的定性分析

本节展示由我们模型生成的两组图像样本。我们从 M2E2 的测试集中随机选取了两个数据点，并利用文本信息对视觉想象进行条件控制，从而生成文本的视觉表示（图 4）。从这些示例可以看出，尽管真实图像未能完整呈现文本信息，但我们提出的模型将这些信息融入了视觉表示中。例如，在第一个示例中，原始图像未能提供任何关于投掷石块或火焰/烟雾的视觉线索。相比之下，我们提出的 Visual Imaginator 捕捉到了这些信息（图 4，第一行）。同样，在第二个示例中，由我们模型生成的图像恰当地呈现了空战，而真实图像以及 Stable Diffusion 生成的图像都未能体现这一点（图 4，第二行）。可以观察到，有时合成图像所包含的信息比其原始对应图像更丰富。这表明，我们的 VF-Event 模型能够在推理过程中生成与上下文相关的图像。

[1]https://github.com/huggingface/transformers[2]由于我们无法获取 Frozen 模型的源代码，因此我们严格按照论文中的描述实现了该模型。这确保我们的实现尽可能接近原始模型。

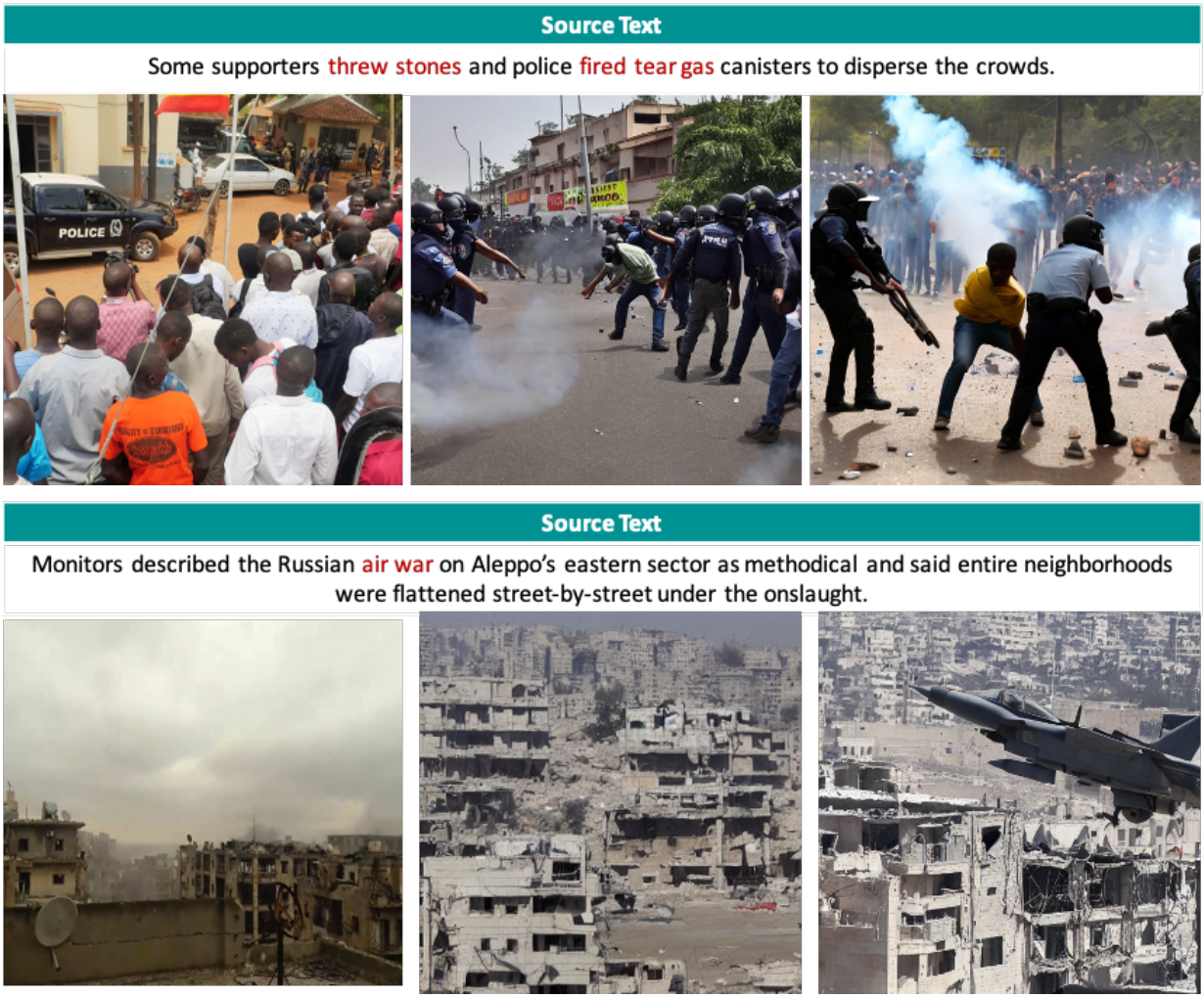


图 4. 真实图像与合成图像的定性比较。**左：** 原始图像。**中：** 预训练 Stable Diffusion 模型 [39] 生成的图像。**右：** 我们的定制 Visual Imaginator 模块 VF-Event 合成的图像。

D. 消融实验

我们进行了消融实验，以研究每种模态对小样本事件检测的影响。我们在两种消融设置下评估了模型的性能：（1）训练期间完全忽略视觉上下文（VF-Event (L)）；（2）训练期间忽略文本上下文（VF-Event (V)）。此外，我们用图像检索方法替换了提出的 Visual Imaginator [46]，并在该设置（VF-Event(RET)）下评估了模型的性能。

表 II 总结了消融实验的结果。如表所示，在所有设置下，VF-Event (L) 的表现均优于 VF-Event (V)，这表明在事件检测中，文本模态比视觉模态包含更丰富的信息。然而，在不同支持集的所有指标上，同时使用两种模态的模型均优于仅使用一种模态（即文本）训练的模型（15 样本设置下的 F1 分数除外）。这一观察结果表明，视觉上下文对事件检测具有互补作用。此外，VF-Event (RET) 的表现逊于但接近 VF-Event (L)，这证明了使用图像检索模型代替 Visual Imaginator 时，我们的模型依然有效。最后，在所有设置下，我们使用 Visual Imaginator 的完整模型均优于 VF-Event (RET)，F1 分数最高相差 12 分。这一观察结果凸显了我们提出的 Visual Imaginator 在合成上下文相关信息方面的重要性。

V. 结论

本文针对模型在训练和推理过程中面对低资源多模态数据时所面临的事件检测挑战。我们提出了一种小样本视觉融合模型，将预训练多模态模块与定制的 Visual Imaginator 模块相结合。通过采用小样本学习方法，并仅利用少量带标签的图文配对样本，我们的模型能够使 Visual Imaginator 适应特定应用场景，从而减少预训练模型的原始领域与目标领域之间的领域偏移。定制的 Visual Imaginator 模块还能够在视觉

表 II M2E2 数据集上的消融实验：视觉模态和文本模态以及 Visual Imaginator 对我们的小样本事件检测模型性能的影响。VF-Event（L）：无图像。VF-Event（V）：无文本。VF-Event（RET）：图像检索。

方法	5 样本			10 样本			15 样本			20 样本		
	F1	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率	F1	精确率	召回率
VF-EVENT (L)	46.19	49.83	45.55	50.26	60.86	47.77	**64.09**	70.02	62.13	62.65	64.85	63.21
VF-EVENT (V)	17.40	16.88	20.0	16.57	22.39	18.88	14.70	18.23	14.44	17.23	29.07	17.24
VF-EVENT (RET)	39.32	40.97	41.11	50.17	57.67	50.0	51.0	59.44	48.88	54.76	54.44	56.32
VF-Event	**50.00**	**50.94**	**50**	**50.58**	**60.95**	**48.88**	63.79	**70.14**	**62.22**	**65.23**	**67.6**	**64.36**

模态缺失的情况下，在推理过程中为模型提供相关的视觉上下文，从而实现有效的事件检测。我们的实验表明，所提出的方法能够有效利用低资源训练数据，并在事件检测方面优于当前最先进的技术。

参考文献

[1] M. A. Valenzuela-Esc´arcega、G. Hahn-Powell、M. Surdeanu 和 T. Hicks，《用于事件抽取的领域无关规则框架》，载于*ACL-IJCNLP 2015*论文集：系统演示，2015年，第127–132页。[2] D. M. H. Cunningham 和 K. Bontcheva，使用*GATE*进行文本处理（第*6*版）。谢菲尔德大学 D，2011年。[3] F. Jenhani、M. S. Gouider 和 L. B. Said，《一种从Twitter中抽取药物滥用事件的混合方法》，*Procedia*计算机科学，第96卷，第1032–1040页，2016年。[4] D. Freitag 和 J. Niekrasz，《通过针对依存句法分析进行模式归纳，推导特征以利用远程标注》，载于第*15*届生物医学自然语言处理研讨会论文集，2016年，第36–45页。[5] I.-H. Hsu、K.-H. Huang、E. Boschee、S. Miller、P. Natarajan、K.-W. Chang 和 N. Peng，《DEGREE：一种数据高效的生成式事件抽取模型》，载于计算语言学协会北美分会人类语言技术*2022*年会议论文集，2022年，第1890–1908页。[6] Z. Zhang、E. Strubell 和 E. Hovy，《通过基于模板的槽位查询，将语义角色标注迁移至事件论元抽取》，载于*2022*年自然语言处理经验方法会议论文集，2022年，第2627–2647页。[7] J. Xu、W. Xu、M. Sun、T. Wang 和 W. Chu，《通过事件矩阵抽取共享触发词的事件》，载于计算语言学协会论文集：*EMNLP 2022*精选论文，2022年，第1189–1201页。[8] Y. Yao、S. Mao、X. Chen、N. Zhang、S. Deng 和 H. Chen，《模式感知的指代作为提示，可改进数据高效的关系三元组与事件抽取》，*arXiv*预印本 *arXiv:2210.10709*，2022年。[9] F. Shiri、T. Y. Zhuo、Z. Li、S. Pan、W. Wang、R. Haffari、Y.-F. Li 和 V. Nguyen，《海事问答系统的释义技术》，载于*2022*年第*25*届国际信息融合会议（*FUSION*）。IEEE，2022年，第1–8页。[10] I. Calixto、Q. Liu 和 N. Campbell，《用于多模态神经机器翻译的双重注意力解码器》，载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），2017年，第1913–1924页。[11] J. Ive、P. S. Madhyastha 和 L. Specia，《借助视觉感知提炼翻译》，载于计算语言学协会第*57*届年会论文集，2019年，第6525–6538页。[12] M. Zhu、P. Pan、W. Chen 和 Y. Yang，《DM-GAN：用于文本到图像合成的动态记忆生成对抗网络》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第5802–5810页。[13] M. Li、A. Zareian、Q. Zeng、S. Whitehead、D. Lu、H. Ji 和 S.-F. Chang，《用于多媒体事件抽取的跨媒体结构化公共空间》，载于计算语言学协会第*58*届年会论文集，2020年，第2557–2568页。[14] Y. Lu、H. Lin、J. Xu、X. Han、J. Tang、A. Li、L. Sun、M. Liao 和 S. Chen，《Text2Event：用于端到端事件抽取的可控序列到结构生成》，*arXiv*预印本 *arXiv:2106.09232*，2021年。

[15] X. Liu、H. Huang、G. Shi 和 B. Wang，“用于生成式模板事件抽取的动态前缀微调”，*arXiv*预印本 *arXiv:2205.06166*，2022年。[16] F. Shiri、T. Wu、Y.-F. Li 和 G. Haffari，“Tcg-event：基于生成的事件抽取的有效任务条件化”，2022年。[17] T. Wu、F. Shiri、J. Kang、G. Qi、G. Haffari 和 Y.-F. Li，“Kc-gee：基于知识的生成式事件抽取条件化 ”，2022 年。[18] M. Tong、S. Wang、Y. Cao、B. Xu、J. Li、L. Hou 和 T.-S. Chua，“新闻文章中的图像增强事件检测”，载于*AAAI*人工智能会议论文集，第34卷，2020年，第9040–9047页。[19] T. Zhang、S. Whitehead、H. Zhang、H. Li、J. Ellis、L. Huang、W. Liu、H. Ji 和 S.-F. Chang，“通过多模态整合改进事件抽取”，载于第*25*届*ACM*国际多媒体会议论文集，2017年，第270–278页。[20] M. Li、R. Xu、S. Wang、L. Zhou、X. Lin、C. Zhu、M. Zeng、H. Ji 和 S.-F. Chang，“Clip-event：通过事件结构连接文本和图像”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第16 420–16 429页。[21] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，“从自然语言监督中学习可迁移的视觉模型”，载于国际机器学习会议。PMLR，2021年，第8748–8763页。[22] Y.-C. Chen、L. Li、L. Yu、A. El Kholy、F. Ahmed、Z. Gan、Y. Cheng 和 J. Liu，“Uniter：通用图像-文本表征学习”，载于计算机视觉——*ECCV 2020*：第*16*届欧洲会议，英国格拉斯哥，*2020*年*8*月*23–28*日，论文集，第*XXX*部分。Springer，2020年，第104–120页。[23] L. Zhou、H. Palangi、L. Zhang、H. Hu、J. Corso 和 J. Gao，“用于图像描述和视觉问答的统一视觉-语言预训练”，载于*AAAI*人工智能会议论文集，2020年，第13 041–13 049页。[24] H. Tan 和 M. Bansal，“Lxmert：从变换器中学习跨模态编码器表征”，*arXiv*预印本 *arXiv:1908 07490*，2019年。。[25] Z. Huang、Z. Zeng、Y. Huang、B. Liu、D. Fu 和 J. Fu，“ 跳出框外看：视觉 - 语言表征学习的端到端预训练 ”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第12 976–12 985页。[26] C. Jia 、 Y. Yang 、 Y. Xia 、 Y.-T. Chen 、 Z. Parekh 、 H. Pham 、 Q. Le、Y.-H. Sung、Z. Li 和 T. Duerig，“利用带噪文本监督扩展视觉和视觉-语言表征学习”，载于国际机器学习会议。PMLR，2021年，第4904–4916页。[27] X. Li、X. Yin、C. Li、P. Zhang、X. Hu、L. Zhang、L. Wang、H. Hu、L. Dong、F. Wei 等，“Oscar：面向视觉-语言任务的物体语义对齐预训练”，载于计算机视觉——*ECCV 2020*：第*16*届欧洲会议，英国格拉斯哥，*2020*年*8*月*23–28*日，论文集，第*XXX 16*部分。Springer，2020年，第121–137页。[28] T. Y. Zhuo、Z. Li、Y. Huang、Y.-F. Li、W. Wang、G. Haffari 和 F. Shiri，“大型预训练语言模型中基于提示的语义解析的鲁棒性：关于Codex的实证研究”，*arXiv*预印本 *arXiv:2301.12868*，2023年。[29] F. Yu、J. Tang、W. Yin、Y. Sun、H. Tian、H. Wu 和 H. Wang，“Ernie-vil：通过场景图增强知识的视觉-语言表征”，载于*AAAI*人工智能会议论文集，第35卷，2021年，第3208–3216页。[30] L. A. Hendricks 和 A. Nematzadeh，“探测图像-语言变换器对动词的理解”，*arXiv*预印本 *arXiv:2106.09141*，2021年。[31] A. Ramesh、M. Pavlov、G. Goh、S. Gray、C. Voss、A. Radford、M. Chen 和 I. Sutskever，“零样本文本到图像生成”，载于国际机器学习会议。PMLR，2021年，第8821–8831页。

[32] P. Esser、R. Rombach 和 B. Ommer，“驯服变换器以实现高分辨率图像合成”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第12 873–12 883页。 [33] S. Reed、Z. Akata、X. Yan、L. Logeswaran、B. Schiele 和 H. Lee，“生成式对抗文本到图像合成”，载于国际机器学习会议。PMLR，2016年，第1060–1069页。 [34] T. Xu、P. Zhang、Q. Huang、H. Zhang、Z. Gan、X. Huang 和 X. He，“Attngan：利用注意力生成对抗网络进行细粒度文本到图像生成”，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第1316–1324页。 [35] H. Zhang、J. Y. Koh、J. Baldridge、H. Lee 和 Y. Yang，“用于文本到图像生成的跨模态对比学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第833–842页。 [36] X. Yu、F. Shiri、B. Ghanem 和 F. Porikli，“我们能看到更多吗？未对齐的微小人脸的联合正面化与幻觉生成”，*IEEE*模式分析与机器智能汇刊，第42卷，第9期，第2148–2164页，2019年。 [37] F. Shiri、X. Yu、F. Porikli、R. Hartley 和 P. Koniusz，“从风格化肖像中恢复保留身份的人脸”，国际计算机视觉期刊，第127卷，第863–883页，2019年。 [38] A. Van Den Oord、O. Vinyals 等，“神经离散表示学习”，神经信息处理系统进展，第30卷，2017年。[39] J. Ho、A. Jain 和 P. Abbeel，“去噪扩散概率模型”，神经信息处理系统进展，第33卷，第6840–6851页，2020年。 [40] N. Ruiz 、 Y. Li 、 V. Jampani 、 Y. Pritch 、 M. Rubinstein 和 K. Aberman，“Dreambooth：针对主题驱动生成微调文本到图像扩散模型”，*arXiv*预印本 *arXiv:2208.12242*，2022年。 [41] Y. Li、R. Panda、Y. Kim、C.-F. R. Chen、R. S. Feris、D. Cox 和 N. Vasconcelos，“Valhalla：用于机器翻译的视觉幻觉”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第5216–5226页。 [42] M. Tsimpoukelli、J. L. Menick、S. Cabi、S. Eslami、O. Vinyals 和 F. Hill，“利用冻结语言模型进行多模态少样本学习”，神经信息处理系统进展，第34卷，第200–212页，2021年。 [43] G. R. Doddington、A. Mitchell、M. A. Przybocki、L. A. Ramshaw、S. M. Strassel 和 R. M. Weischedel，“自动内容抽取（ACE）项目——任务、数据与评估。”载于*LREC*，第2卷，第1期。里斯本，2004年，第837–840页。 [44] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，“利用潜在扩散模型进行高分辨率图像合成”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2022年6月，第10 684–10 695页。 [45] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，*arXiv*预印本 *arXiv:1412.6980*，2014年。 [46] Z. Wu、L. Kong、W. Bi、X. Li 和 B. Kao，“因误解的理由而表现良好：对多模态机器翻译中视觉上下文必要性的实证再审视”，载于第*59*届计算语言学协会年会和第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），2021年，第6153–6166页。