时尚领域中的掩码视觉-语言Transformer模型

季葛鹏†[1], 诸葛鸣晨†[1], 高德宏[1], 范登平✉[2], Christos Sakaridis[2] and Luc Van Gool[2]

[1]国际技术部, 阿里巴巴集团, 杭州, 中国.
[2]计算机视觉实验室, 苏黎世联邦理工学院, 苏黎世, 瑞士.

Abstract

本文设计了一个掩码视觉-语言Transformer模型（**MVLT**），旨在学习时尚领域中的跨模态表征。从技术层面来看，本文使用基于视觉Transformer模型针对BERT [1]结构进行重构，使其成为时尚领域中第一个可端到端训练的多模态框架。此外，本文设计了掩码图像重建（Masked Image Reconstruction，MIR）预训练策略，用于时尚领域中的细粒度理解。MVLT模型的易用性高且扩展性强，接收原始多模态数据作为输入，可以进行隐式的视觉-语言对齐，而无需引入额外的预处理模型（例如：ResNet）。更重要的是，MVLT模型能够轻易泛化到各种匹配任务和生成任务中。实验结果表明，与Fashion-Gen 2018数据集获胜者Kaleido-BERT相比，在检索任务rank@5指标和识别任务精度上分别提高了**17%**和**3%**。代码获取链接：https://github.com/GewelsJI/MViLT。

Keywords: 视觉-语言预训练任务、掩码图像重建、Transformer、时尚领域、电子商务。

1 引言

Transformer模型的出现引起了学术界的广泛关注，并促进了计算机视觉（CV）[3, 4]和自然语言处理（NLP）[5, 6]的发展。由于Transformer模型的卓越表现，研究者们也不断探索其在视觉-语言（VL）领域的作用 [7–11]。为更好地利用CV 和NLP 领域中的预训练模型，现有的通用视觉-语言模型主要使用预训练后的BERT模型 [1]、视觉特征提取器 [14, 15]或者同时使用两者。然而，通用的视觉-语言方法 [16–18]仍难以被应用于电商中的时尚领域，主要因为以下两个问题：**a）粒度不足：**不同于具有复杂背景的通用场景，若模型仅关注时尚产品的粗粒度语义是远远不足的 [19–21]，因为这种方式将导致

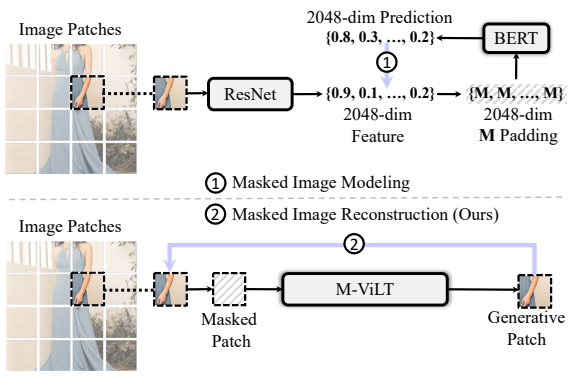


图1 用于视觉-语言预训练 [12, 13]的视觉重建任务使用了随机掩码策略（即：使用M填充去替换原始向量）的掩码图像建模（上图），其用于在特征层级重建预提取的视觉语义（向量）。本文引入基于掩码图像重建的生成式任务（下图），其直接重建像素层级的原始图像。

† 同等贡献。✉ 通讯作者。本工作由季葛鹏在阿里巴巴集团实习期间完成。本文为MIR2023 [2]的中文翻译稿，由季葛鹏翻译，诸葛鸣晨、高德宏、范登平校稿。

网络收敛于次优解。反之，面向时尚领域的模型往往需要更细粒度的表征，例如：一件具有不同材质（例如：羊毛、亚麻、棉）或衣领（例如：立领、古巴领、温莎领）的西装。**b）迁移性差:** 就时尚领域任务而言，当前预提取的视觉特征缺乏针对性，从而限制了跨模态表征的能力。

为解决上述问题，本文提出了一个新颖的视觉-语言框架，称为掩码视觉-语言Transformer（Masked Vision-Language Transformer, MVLT）。具体而言，本文首先针对时尚领域的VL框架引入了一个生成式任务，即：掩码图像重建（Masked Image Reconstruction，MIR）。相比于之前的预训练任务，例如：掩码图像建模（回归任务）或者掩码图像分类（分类任务），MIR使网络通过像素级视觉信息来习得更多细粒度表征（请参见图1）。此外，受金字塔视觉Transformer（PVT）[22]的启发，本方法使用金字塔结构作为视觉-语言Transformer。所引入的MIR任务显著增强了模型对特定时尚领域理解和生成式任务的适应能力，并且能够以端到端的方式训练。为此，MVLT模型可直接处理原始的稠密形式的多模态输入，即：语言词例（token）和图像块（patch），而无需额外的预处理模型 [23, 24]（如使用ResNet）。**本文的主体贡献总结如下:**

- 本文提出一种全新的掩码图像重建（MIR）任务，这是在视觉-语言预训练中第一个采用像素级生成的方案。
- 基于MIR任务，本文提出了一个用于时尚领域的端到端视觉-语言框架MVLT，极大地提高了下游任务和大规模网站应用的可迁移性。
- 广泛实验表明，MVLT模型在匹配式和生成式任务上的表现明显优于同期的前沿模型。

2 研究背景

近年来，基于BERT模型的预训练模型在视觉-语言任务中被广泛研究。之前的若干尝试都在诸多下游应用中获得了成功，例如：LXMERT [25]、VL-BERT [26]和FashionBERT [12]。现有研究及结果表明，BERT模型是一种用于学习多模态表征的强大方法，其性能优于先前基于CNN模型 [27]或基于LSTM模型 [28]的方法。与以往研究工作相比，本文旨在开发一种更为高效的自监督目标任务，其可以轻易地部署于预训练过程，并为实际应用提供更好的表征。因此，本节回顾了给予我们最大启发的有关掩码式学习策略和端到端多模态框架的文献。

2.1 掩码式学习策略

掩码建模策略是BERT模型 [1]中重要的自监督任务，并在自然语言处理领域中崭露头角。因为该策略在多模态和视觉任务中同样具有实用性，所以研究人员迁移了其在语言模型中的优势。大多数视觉-语言方法 [17, 26, 29]将掩码建模策略迁移到视觉词例（token）中，并使用回归任务从随机替换中构建词例（token）特征或者使用分类任务来预测词例（token）的属性。为了降低学习的难度，Kaleido-BERT模型 [13]通过采用万花筒策略来优化掩码式建模，进而促进多粒度语义下的联合学习。虽然该方法提高了时尚领域中视觉-语言相关任务的性能，但是词例-图像块（token-patch）的预对齐方案仍然很复杂，阻碍了实际场景中的下游应用。与本文思想类似的是另一工作 [30]引入了MLIM方法，使用图像重建任务增强了掩码图像的建模。然而，本文的实验表明在没有任何提示的情况下，要求一个模型重建整张图像过于困难。最近，BEiT模型 [31]和MAE模型 [32]利用BERT风格的预训练任务作为视觉学习的一部分，并发现在该方案下能够有效地学习

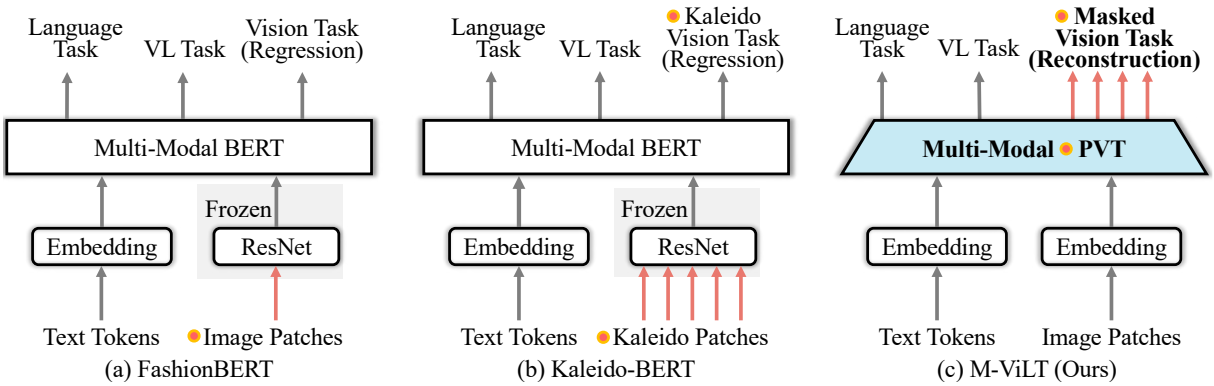


图2 MVLT与现有的时尚领域视觉-语言框架的比较。FashionBERT模型（a）利用基于语言的编码器（即：BERT模型）来提取单尺度视觉输入（即：图像块）的视觉-语言表示。Kaleido-BERT模型（b）基于上述模型进行了如下两种扩展：增加了五个固定尺度的输入（即：Kaleido图像块）以获得层次化的视觉特征，并设计了Kaleido视觉预任务以充分学习视觉-语言表征。然而，这些模型的视觉编码是固定的（即：不可训练的权重），因此缺乏特定领域的视觉知识，严重阻碍了它们的迁移性。不同的是，本文所提出的MVLT模型（c）通过在端到端框架中引入掩码视觉任务，以一种自适应的方式学习层次化特征，大大促进了与视觉-语言相关的理解和生成任务。

视觉语义信息。这两项工作进一步表明，将原始的掩码图像建模（即回归任务）转换为掩码图像重建任务是具备可能性的。然而，本文的主要目标是设计一个生成式的代理（pretext）任务，在消除对先验知识需求的同时，让视觉-语言预训练中图像建模过程更为容易，这将使得拥有十亿级数据容量的实际下游应用极大受益。

2.2 端到端多模态框架

Pixel-BERT模型 [33]是第一个考虑端到端预训练的方法。其采用了2×2的最大池化层来减少图像特征的空间维度，每张图像被降采样64倍。虽然这项工作开创了端到端预训练的先例，但是这种初步方法在实际环境中不能很好地发挥其效力，因其仍然使用了ResNet模型 [14]作为联合预训练的一部分，并没有考虑到速度和性能的损失。最近，VX2TEXT模型 [34]提出将所有模态转换到语言空间，然后使用松弛方案实现端到端预训练。虽然将所有的模态转化为统一隐空间的做法是令人振奋的，但是忽略了其仍然使用预训练方法所提取的数据作为模型输入，因此不能被视为一个真正意义上的端到端的框架。

根据时间线梳理，ViLT模型 [35]是第一个使用基于图像块（patch）的映射取代基于区域或网格特征来设计端到端框架的研究方法。然而，如果没有其他设计并不具有竞争力的性能，因为其仅是ViT模型 [3]的一个简单扩展。Grid-VLP模型 [36]与ViLT模型相似，但其进一步证明了使用预训练的CNN作为视觉骨干网络可以提高下游任务的性能。SOHO模型 [37]将整个图像作为输入，并创建一个视觉字典来进行局部区域的仿射变换。然而，由于缺乏可靠的对齐信息，这种方法并不适合时尚领域的应用。因此视觉字典可能仅学习到背景或前景的定位，而不是复杂的语义。此外，还有一些为端到端预训练设计的方法 [38–40]，但它们是用于特定的任务，并不直接适用于本文的研究领域。

尽管现有的研究工作都采用了不同的策略来构建端到端的方案，但丢弃使用预训练模型的方法（如：ResNet模型、BERT模型）和使用原始数据（即：文本、图像）作为输入的解决方案仍然没有得到充分的探索，而这一特性对于多模态应用的可拓展性十分重要。

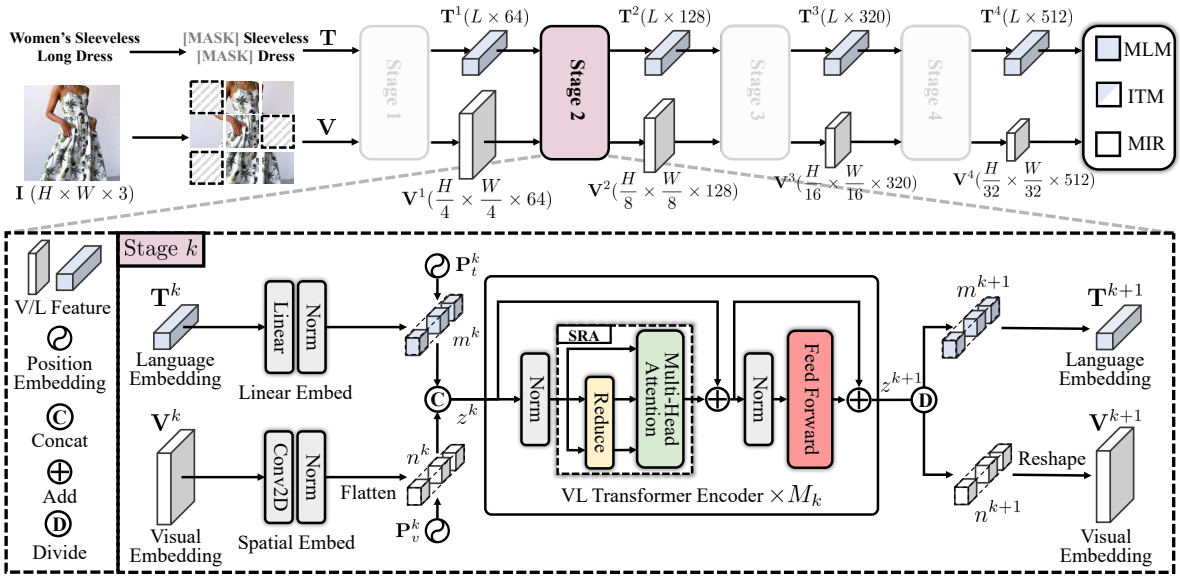


图3 MVLT的框架示意图。MVLT的整体结构由四个阶段组成，每个阶段均包含语言和视觉嵌入过程以及M_k个Transformer编码器。通过在三个子任务中引入掩码策略，即：掩码图像重建（MIR）、图像-文本匹配（ITM）和掩码语言建模（MLM），MVLT以端到端的方式进行训练。详细描述请参见第3节。

评注： 如图2所示，与现有的两种时尚领域的方法相比，即FashionBERT模型（a）和Kaleido-BERT模型（b），本文提出的MVLT（c）也是一个基于图像块（patch）的视觉-语言学习框架。本文所提出的MVLT将PVT模型 [22]进行了扩展，成为一个用于解决时尚领域跨模态任务的自适应提取层次表征架构。它也是第一个用于解决时尚领域视觉-语言预训练的端到端模型，这也使得在使用双塔结构时，可以简化MVLT模型在时尚领域任务中复现流程。

3 掩码式视觉-语言Transformer模型

本文旨在为时尚领域建立一个端到端的视觉-语言预训练框架。图3展示了MVLT模型的整体框架。与PVT模型类似，本文提出的模型也采用四阶段结构生成不同尺度的特征。其中，多模态编码器（第3.1节）和预训练目标（第3.2节）为MVLT模型的两个关键部分。

3.1 多模态编码器

如图3所示，MVLT模型可同时接受视觉和语言的输入。对于语言侧，本文首先对时尚产品描述文字进行语言词例（token）的生成，并使用特定的`[MASK]`词例以r_l的比例[1]对描述词例进行随机掩码。随后，在掩码后的词例序列头部插入一个特定的`[CLS]`词例。此外，如果词例序列长度小于128，使用`[PAD]`词例将该序列填充到一个统一的长度L。这会生成语言的输入id成为$\mathbf{T} \in \mathbb{R}^L = \langle t_1; \cdots; t_L \rangle$。在视觉侧，本文将$\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$作为视觉输入，其中$H$和$W$代表给定输入图像的高度和宽度。该输入被切分成多个网格状的图像块（patch）$\mathbf{V} \in \mathbb{R}^{N \times P \times P \times 3} = \langle v_1; \cdots; v_N \rangle$，其中$N = \frac{HW}{P^2}$为图像块的总数，$P$为图像块的长宽尺寸。同样地，分割后图像块以$r_v$的比例进行掩码操作。关于上述语言和视觉部分的掩码策略，在第3.2节给出了详细描述。

[1]这里遵循了BERT模型 [1]的默认设置。

表1 MVLT模型中多模态编码器的超参数设置。

超参数	$k=1$	$k=2$	$k=3$	$k=4$
层数量M_k	2	2	2	2
隐维度D_k	64	128	320	512
降采样尺寸R_k	4	8	16	32
卷积核尺寸K_k	4	2	2	2
步长S_k	4	2	2	2

上述的多模态输入数据经编码后被送入四个连续的视觉-语言交互阶段（即：$k \in \{1,2,3,4\}$）。在第一阶段，本文通过给定的输入（$\mathbf{T}$和$\mathbf{V}$），分别生成视觉嵌入特征$\mathbf{T}^1$和语言嵌入特征$\mathbf{V}^1$。为了简化描述，在后文中将默认考虑第k个阶段。如图3底部所示，本文首先将语言输入$\mathbf{T}^k \in \mathbb{R}^{L\times D_k}$ 编码为语言隐特征$m^k \in \mathbb{R}^{L\times D_{k+1}}$，该过程可以被表述为：

$$m^k = \mathbf{T}^k * \mathbf{W}_t^k + \mathbf{P}_t^k, \tag{1}$$

其中$\mathbf{W}_t^k \in \mathbb{R}^{D_k\times D_{k+1}}$ 和$\mathbf{P}_t^k \in \mathbb{R}^{L\times D_{k+1}}$分别代表可学习的线性编码矩阵和位置编码矩阵。$D_k$代表隐特征编码的维度。

每阶段输入的视觉编码为$\mathbf{V}^k \in \mathbb{R}^{\frac{H}{R_k}\times\frac{W}{R_k}\times D_k}$，其中$R_k$表示视觉编码的空间降采样尺寸系数。为获得金字塔式的视觉特征，首先对$\mathbf{V}^k$进行二维投影（即二维卷积模块），再展平为视觉隐特征$n^k \in \mathbb{R}^{(HW/R_{k+1}^2)\times D_{k+1}}$。具体而言，该投影过程使网络通过参数为$\mathbf{W}_v^k \in \mathbb{R}^{D_k\times K_k\times K_k\times D_{k+1}}$ 的卷积核将等效空间维度从$\mathbb{R}^{HW/R_k^2}$ 减少到$\mathbb{R}^{HW/R_{k+1}^2}$，其中$K_k$ 为卷积核尺寸，S_k 为步长，公式描述为：

$$n^k = \mathbf{Flatten}(\mathbf{V}^k * \mathbf{W}_v^k) + \mathbf{P}_v^k, \tag{2}$$

其中，$\mathbf{P}_v^k \in \mathbb{R}^{N\times D_{k+1}}$ 表示位置编码矩阵。随后将视觉-语言隐特征进行拼接$z^k = \langle m^k;\ n^k\rangle$，并将其送入$M_k$个视觉-语言Transformer编码器。每个编码器都包含具有空间降采样（即：降采样模块）的多头自注意层、多层感知机（MLP）和层归一化模块。最后，得到编码后的多模态特征$z^{k+1} = \langle m^{k+1};\ n^{k+1}\rangle$，并将其拆分为语言部分$\mathbf{T}^{k+1} = m^{k+1}$和视觉部分$\mathbf{V}^{k+1} = \mathbf{Reshape}(n^{k+1})$，其中$\mathbf{Reshape}(\cdot)$操作用于恢复指定特征的空间维度。

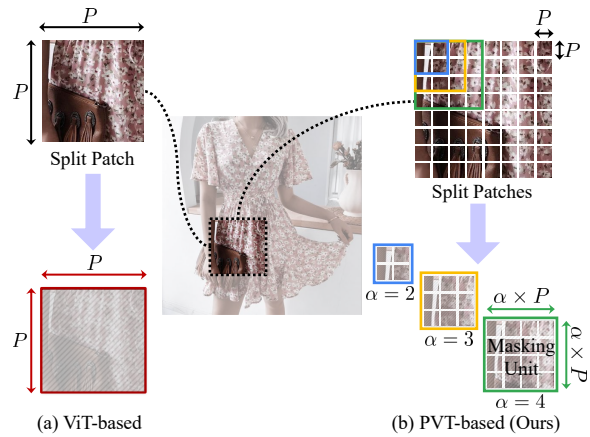


图4 基于PVT的结构为设计掩码策略提供了更多的选择。基于ViT模型[35]的传统方法（a）只选择一个固定尺度，即P^2的图像块（patch）来进行掩码。然而，基于PVT的方法（b）更加灵活，因为它结合了更多细粒度的图像块（patch）作为掩码的基本单元，即$(\alpha \times P)^2$，其中$\alpha \in \{1,2,..,8\}$。该特性提供了一种灵活的方式，即通过不同α值学习合适的语义信息。

经过四个视觉-语言交互阶段，分别生成了四个文本编码特征$\{\mathbf{T}^k\}_{k=1}^4$和四个具有金字塔形态的视觉编码特征$\{\mathbf{V}^k\}_{k=1}^4$。表1中展示了更为详细的超参数设置。

3.2 预训练任务

为了获得具有鉴别性的多模态表征，本文采用了三个预训练任务，从最原始的模态输入中建立视觉-语言之间的交互关系和内在关系，包括：视觉模态交互任务（掩码图像重建，MIR）、语言模态交互任务（掩码语言建模，MLM）以及视觉-语言模态交互任务（图像-文本匹配，ITM）。

预训练目标1：掩码图像重建（MIR）：通用领域模型可以从基于图像块或基于区域的预训练目标中学习粗粒度语义，并获得令人满意的结果。但是，时尚领域的模型需要更为细粒度的

表征，例如：具有不同材料（例如：羊毛）或衣领（例如：温莎）的西装，这就需要建立像素到像素关系的视觉预训练目标。受掩码式语言建模 [1]的启发，本文试图从生成任务的角度建立像素与像素之间的关联，进而提高视觉表征的可延展性。因此，本文设计了掩码图像重建（MIR）预训练目标。为提高MVLT模型在MIR任务上的学习效果，本文基于PVT [22]的金字塔结构设计了一个灵活的掩码策略。不同于图4（a）中基于ViT模型的掩码策略，本文的架构（b）设计了包含更高细粒度图像块（patch）的掩码单元矩阵，对输入图像进行掩码操作。给定图像块（patch）序列$\mathbf{V}=\{v_n\}_{n=1}^{N}\in\mathbb{R}^{N\times P\times P\times 3}$，掩码后序列$\mathbf{V}_{\setminus\Phi}$的定义如下：

$$\begin{aligned}\mathbf{V}_{\setminus\Phi}&=\mathcal{F}_M(\{\mathbf{M}(q;\ \alpha;\ \Phi)\}_q^Q,\{v_n\}_{n=1}^N)\\&=\begin{cases}\texttt{[ZERO]}, & \mathbf{M}(q;\ \alpha;\ \Phi)=1\\ v_n, & \mathbf{M}(q;\ \alpha;\ \Phi)=0,\end{cases}\end{aligned}\tag{3}$$

其中，$\mathcal{F}_M(\cdot;\ \cdot)$代表掩码函数（或过程），$q$代表随机选择的掩码单元区域，[ZERO]表示使用像素值0来填充所选掩码区域[2]。其中，掩码单元$\{\mathbf{M}(q;\ \alpha;\ \Phi)\}_{q=1}^{Q}$由指示函数得来：

$$\mathbf{M}(q;\ \alpha;\ \Phi)=\mathbf{1}(q)=\begin{cases}1, & q\in\Phi,\\ 0, & q\notin\Phi,\end{cases}\tag{4}$$

其中，整数集Φ中的每个值都是以r_v的比例在$[1,Q]$的范围内随机选择的，$Q=\frac{H\times W}{(\alpha\times P)^2}$为掩码单元的总数量。如图4（b）所示，$\alpha$的取值范围为$\{1,2,..,8\}$。为捕捉更高细粒度的语义，本文选择$\alpha=4$作为默认设置[3]。

[2]在具体实现中，本文选择[ZERO] $=10^{-6}$以带来更好的优化稳定性和更少的模式退化。

[3]图4（a）中传统掩码策略（$P=32$）为本文提出的掩码策略图4（b）的一个特例（即：当$\alpha=8, P=4$时）。

由于smooth-$\ell 1$损失函数对离群值不敏感特性，本文将其作为预训练优化目标，使用掩码后的序列$\mathbf{V}_{\setminus\Phi}$重建整个图像，公式定义如下：

$$\mathcal{L}_{\text{MIR}}=\begin{cases}0.5\times(\mathbf{I}'_{(x,y)}-\mathbf{I}_{(x,y)})^2, & \text{if } \mathbf{I}'_{(x,y)}-\mathbf{I}_{(x,y)}<1\\ |\mathbf{I}'_{(x,y)}-\mathbf{I}_{(x,y)}|-0.5, & \text{otherwise,}\end{cases}\tag{5}$$

其中，$\mathbf{I}'_{(x,y)}$和$\mathbf{I}_{(x,y)}$分别表示重建图像$\mathbf{I}'$和输入图像$\mathbf{I}'$，(x,y)代表像素坐标。$\mathbf{I}'=\mathcal{F}_{\text{MIR}}(\mathbf{V}_{\setminus\Phi};\ \mathbf{W}_{\text{MIR}})$由可学习的权重$\mathbf{W}_{\text{MIR}}$进行参数化。函数$\mathcal{F}_{\text{MIR}}(\cdot;\ \mathbf{W}_{\text{MIR}})$代表一个标准的四层U-Net [41]形状解码器，其接受四个金字塔式视觉编码特征$\{\mathbf{V}^k\}_{k=1}^4$作为输入。

预训练目标2: 图像-文本匹配（ITM）： 在最后一个语言编码$\mathbf{T}^4$中所附加的分类编码被用来耦合来自视觉-语言模态输入的表征。然后利用函数$\mathcal{F}_{\text{ITM}}(\cdot;\ \mathbf{W}_{\text{ITM}})$来表示权重为$\mathbf{W}_{\text{ITM}}$的全连接层（FC）和softmax层。$\mathcal{F}_{\text{ITM}}$输出二分类概率向量$\mathbf{p}_{\text{ITM}}=\mathcal{F}_{\text{ITM}}(\langle\mathbf{T},\mathbf{V}\rangle;\ \mathbf{W}_{\text{ITM}})$，用于表示所输入时尚领域的图像和描述是否匹配（即：正、负样本）。其中，正样本是从同一时尚产品类别中选择的，而负样本则是从不同条目中随机选择的。该预训练目标使用了二元交叉熵损失函数，

$$\begin{aligned}\mathcal{L}_{\text{ITM}}=&-\mathbb{E}_{\langle\mathbf{T},\mathbf{V}\rangle}[\mathbf{y}_{\text{ITM}}\log(\mathbf{p}_{\text{ITM}})\\&+(1-\mathbf{y}_{\text{ITM}})\log(1-\mathbf{p}_{\text{ITM}})],\end{aligned}\tag{6}$$

其中$\mathbf{y}_{\text{ITM}}$用于表示真实标签，1表示正样本，0表示负样本。

预训练目标3: 掩码语言建模（MLM）： 根据文献 [42]，本文使用特定的[MASK]词例来随机替换原始文本词例。MLM预训练目标是利用未掩码的词例和图像块来预测被掩码的文本内容。给定一个词例序列$\mathbf{T}=\{t_1,\ldots,t_L\}$，掩码后的序列可以表示为$\mathbf{T}_{\setminus i}=\{t_1,\ldots,\texttt{[MASK]}_i,\ldots,t_L\}$。本文使

表2 Fashion-Gen数据集的检索任务（TIR和ITR）和识别任务（M-CR和S-CR）上的性能展示。↑表示数值越大表现越好。其中，Sum$\mathcal{R}$=($\mathcal{R}$@5+$\mathcal{R}$@10) ×100 和Sum$\mathcal{C}$=($\mathcal{A}$ + macro-$\mathcal{F}$) × 100。“N/A”表示无法获取分数。“Diff”是指排名第二的方法与MVLT模型之间性能的数值差异。

任务	指标	VSE arXiv$_{14}$	VSE++ BMVC$_{18}$	SCAN ECCV$_{18}$	PFAN arXiv$_{19}$	ViLBERT NeurIPS$_{19}$	ImageBERT arXiv$_{20}$	FashionBERT SIGIR$_{20}$	VL-BERT ICLR$_{20}$	OSCAR ECCV$_{20}$	Kaleido-BERT CVPR$_{21}$	MVLT OUR$_{22}$	MVLT Diff
TIR	$\mathcal{R}$@5 ↑	12.76%	16.89%	13.00%	20.79%	37.23%	45.20%	46.48%	36.48%	49.14%	60.60%	**78.00%**	+17.40%
	$\mathcal{R}$@10 ↑	20.91%	28.99%	22.30%	31.52%	50.11%	55.90%	55.74%	48.52%	56.68%	68.59%	**89.50%**	+20.91%
	Sum$\mathcal{R}$ ↑	33.67	45.88	35.30	52.31	87.34	101.10	102.22	85.00	105.82	129.19	**167.50**	+38.31
ITR	$\mathcal{R}$@5 ↑	11.03%	14.99%	16.50%	14.90%	40.49%	41.89%	46.31%	39.90%	44.67%	60.09%	**77.20%**	+17.11%
	$\mathcal{R}$@10 ↑	22.14%	24.10%	26.60%	24.20%	48.21%	50.77%	52.12%	46.05%	52.55%	68.37%	**91.10%**	+22.73%
	Sum$\mathcal{R}$ ↑	33.17	39.09	43.10	39.10	88.70	92.66	98.43	85.95	97.22	128.46	**168.30**	+39.84
M-CR	$\mathcal{A}$ ↑	N/A	N/A	N/A	N/A	N/A	90.77%	91.25%	N/A	91.79%	95.07%	**98.26%**	+3.19%
	macro-$\mathcal{F}$ ↑	N/A	N/A	N/A	N/A	N/A	0.699	0.705	N/A	0.727	0.714	**0.896**	+0.169
	Sum$\mathcal{C}$ ↑	N/A	N/A	N/A	N/A	N/A	160.67	161.75	N/A	164.49	166.47	**187.86**	+21.39
S-CR	$\mathcal{A}$ ↑	N/A	N/A	N/A	N/A	N/A	80.11%	85.27%	N/A	84.23%	88.07%	**93.57%**	+5.50%
	macro-$\mathcal{F}$ ↑	N/A	N/A	N/A	N/A	N/A	0.575	0.620	N/A	0.591	0.636	**0.829**	+0.193
	Sum$\mathcal{C}$ ↑	N/A	N/A	N/A	N/A	N/A	137.61	147.27	N/A	143.33	151.67	**176.47**	+24.80

用交叉熵损失对该目标进行建模:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{\mathbf{T}}[\log(\mathbf{p}_{\text{MLM}})], \tag{7}$$

其中，$\mathbf{p}_{\text{MLM}} = \mathcal{F}_{\text{MLM}}(\mathbf{T}_{\setminus i};\ \mathbf{W}_{\text{MLM}})$ 表示使用$\mathbf{T}_{\setminus i}$ 对每个被掩码词例$\texttt{[MASK]}_i$ 的预测概率。函数$\mathcal{F}_{\text{MLM}}(\cdot;\ \mathbf{W}_{\text{MLM}})$ 代表参数为$\mathbf{W}_{\text{MLM}}$ 的分类器。MVLT模型最终预训练目标采用了如下三个预训练目标的组合:

$$\mathcal{L}_{\text{total}} = w_1 \times \mathcal{L}_{\text{MIR}} + w_2 \times \mathcal{L}_{\text{ITM}} + w_3 \times \mathcal{L}_{\text{MLM}}. \tag{8}$$

3.3 下游任务

为公平起见，本文遵循了与文献 [12, 13]相同的标准，在实验中采用了Fashion-Gen 2018 [43]测评基准。该数据集包含67,666个时尚产品及其相关产品描述（60,147个训练样本和7,519个测试样本），每个产品包含了不同视角下的产品图像（包括1 ∼ 6 个样本）。因此，本文最终使用了260,480和35,528个图像-文本对作为训练集合和测试集合。为了公平对比，本文在Fashion-Gen数据集上测试了MVLT模型并在如下四个时尚领域相关的视觉-语言下游任务和其他模型进行对比。

下游任务1: 文本-图像检索（TIR）： TIR任务要求模型在不同查询图像中检索出一个最高相似度的文本。本文将一个产品的标题和其对应的图像，作为一个图像-文本的正样本对，而图像-文本的负样本对则是从不匹配的图像池中随机选择而来。为增加实验的难度，本文将同一组图像-文本输入（1个正样本和100个负样本）限制在同一子类别中，使得它们尽可能相似。

下游任务2: 图像-文本检索（ITR）： 作为TIR任务的逆过程，ITR任务旨在从所给定一串时尚相关的描述文本中检索出与之最匹配的图像，上述双向的检索任务（即：TIR任务和ITR任务）为跨模式研究中重要的研究方向。与上述TIR任务中的样本挑选策略相类似，本文准备了101组候选图像-文本对，包括同一子类别中的1个正样本对和100个负样本对。MVLT模型在没有进一步微调的情况下，评估了对上述两项检索任务的零样本学习能力。本文使用准确率（即：$\mathcal{R}$@5指标和$\mathcal{R}$@10指标），其通过对一组预测概率的排序来进行性能评估。

下游任务3: 类别识别（M-CR和S-CR）： 该任务分为两个部分：品类识别（M-CR）和子品类识别（S-CR）。作为电商业务的基础，这些任务被用于提供所查询产品的具体类别。本文希望模型应具备在不同粒度下的识别能力，包括48 个品类和122 个子品类，例如：{M-CR = $\texttt{SWEATERS}$, S-CR = $\texttt{CREWNECKS}$}。在最后一个语言

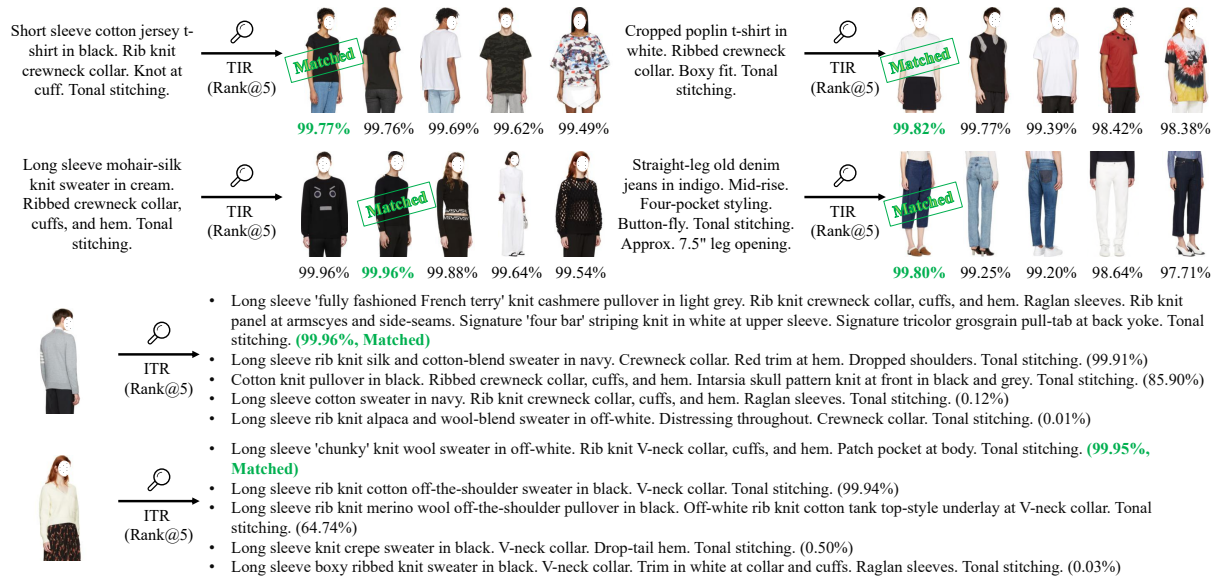


图5 TIR和ITR任务的可视化结果，即由 MVLT模型所预测概率中的前五名。“Matched” 表示图像-文本对的真实标注。

编码特征$\mathbf{T}^4$中的类别词例编码后端，本文增加了两个相互独立的全连接层，用于分别生成两个不同的产品类别分类概率，这个过程需要使用对应标注进行额外的微调。最后，采用两个分类相关的指标来评估性能：准确率（$\mathcal{A}$）和macro F-measure（macro-$\mathcal{F}$）。

任务4: 掩码图像生成（MIG）： MIG任务可以被视为一个像素级别重建任务，即对图像中以概率r_v对图像块（patch）进行随机掩码（请参见第3.2节中所描述的MIR预训练目标）。然后以未掩码的区域作为视觉线索，要求模型重建出完整的图像。

4 实验

本节将详细介绍实验部分，以验证MVLT模型中各个子模块的有效性。

4.1 设定

本小节给出了训练过程中的超参数设置。**i）预训练过程：** 使用PyTorch框架在8个Tesla V100 GPU下进行训练。采用AdamW优化器，动量值设定为0.9，批大小设定为1200（即：每个GPU为150），权重衰减设定为10^{-4}。为避免过拟合，MVLT模型使用了ImageNet预训练的权重 [22]初始化。初始学习率设定为2.5×10^{-3}，并使用余弦衰减策略进行学习率调整。对于视觉侧，输入图像的尺寸被统一调整为$H = W = 256$，并被分割成多个尺寸为$P = 4$的图像块（patch）。对于语言侧，所有产品的文本描述都被词例化，并填充为统一长度$L = 128$的词例（token）序列，其中包含了分类词例、描述词例和填充词例。视觉和语言的掩码概率分别被设定为$r_v = 0.5$ 和$r_l = 0.15$。本文根据经验设置了权重系数$\{w_1 = 10, w_2 = 1, w_3 = 1\}$来平衡不同损失数值的数量级。**ii）微调过程：** 本文采用端到端的微调方式将预训练的视觉-语言表征迁移到不同的下游应用，其设置与预训练过程保持一致。

4.2 实验结果

第3.3节中给出了四个时尚相关下游任务的细节。实验结果表明，MVLT模型优于所有的对比方法，包括：VSE [44]、VSE++ [45]、SCAN [27]、PFAN [46]、ViLBERT

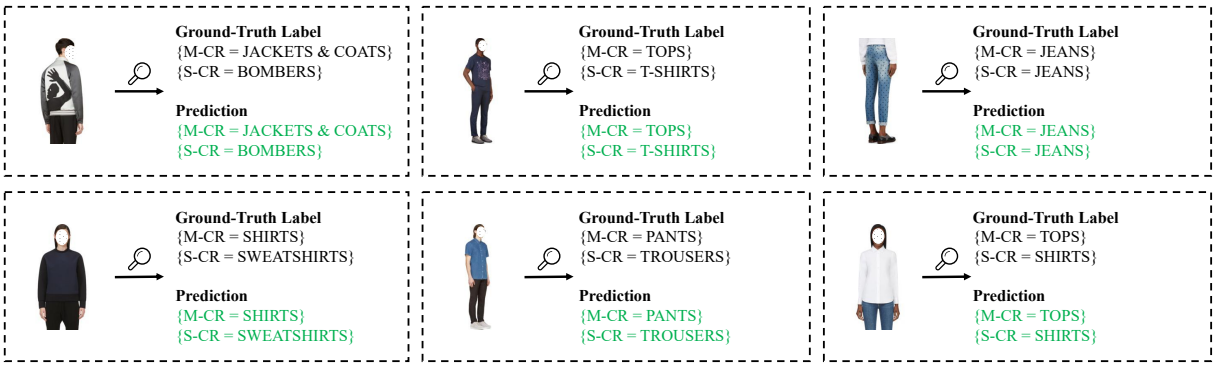
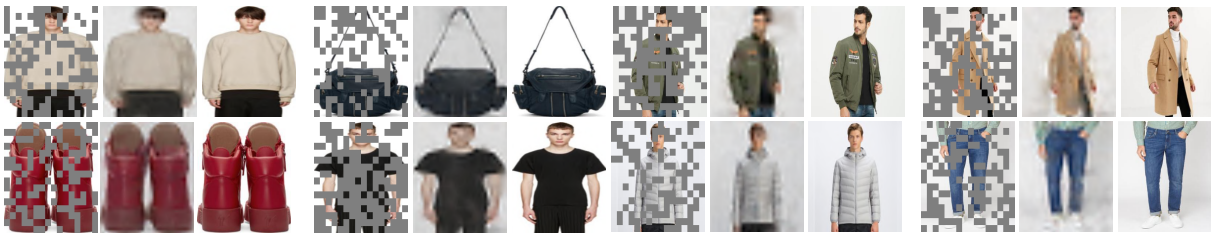


图6 在Fashion-Gen上父类/子类识别的可视化结果。绿色字体表示与真实标签匹配。



(a) Fashion-Gen (In-domain) (b) Data from e-commercial website (Out-of-domain)

图7 MVLT模型生成样本的可视化。灰色代表被掩码的区域。

[17]、ImageBERT [16]、FashionBERT [12]、VL-BERT [26]、OSCAR [29]和Kaleido-BERT [13]，这充分展现了MVLT模型在处理视觉-语言理解和生成任务上的优越性。

TIR 和ITR任务： 如表2 所示，MVLT模型在TIR任务上超过了最优方法（即：Kaleido-BERT-CVPR$_{21}$），在$\mathcal{R}$@5指标、$\mathcal{R}$@10指标上的提升分别为**17.40%**和**20.91%**。对于ITR任务，本文提出的方法取得了更具竞争力的结果，在$\mathcal{R}$@5指标和$\mathcal{R}$@10指标上分别提高了**17.11%**和**22.73%**。在这些情况下，结果都有力地证明了本文所提出模型具有匹配视觉和语言的能力，并展示出**a）MIR预训练目标**和**b）端到端预训练**在时尚领域中的效果。得益于简单、高效、强大的结构设计，MVLT模型能够应用于许多相关领域。此外，图5 中展示了这两个检索任务的可视化结果。

M-CR任务和S-CR任务： 与基于BERT架构的模型 [12, 13, 16, 29] 相比，本文在这两个任务中

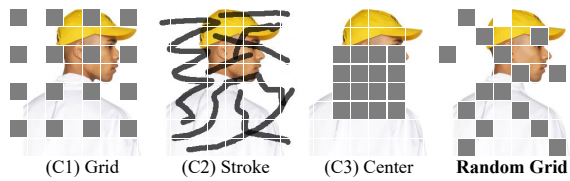


图8 本文设计了四种不同的图像掩码策略，实验结果表明使用随机网格策略的表现最好。

也取得了最好的表现，验证了MVLT模型具有出色的图文理解能力。与现有的最优方法Kaleido-BERT相比，所提出的模型在S-CR 任务中的macro-$\mathcal{F}$ 指标提高了**0.193**。此外，在Sum$\mathcal{C}$指标方面的平均提升（即M-CR：**21.39** 和S-CR：**24.80**）均十分明显。由于该指标对数据分布非常敏感，进一步表明MVLT模型具有很强的鲁棒性。我们还在图6中展示了M-CR和S-CR的识别结果。

MIG任务： 图7展示了在Fashion-Gen 2018数据集验证子集（a）和我们的电商网站（b）上的图像重建效果，图像对比表明本文的重建性能十分显著。由于这个任务需要模型真正地学习到与

表3 本文的MVLT模型中五个关键预训练因素的消融实验。更多相关分析请参见第4.3节。

		(a) 掩码比例 (r_v)				(b) 掩码单元尺寸 (α)				(c) 掩码方式			(d) 预训练目标			(e) 预训练	MVLT
		(A1)	(A2)	(A3)	(A4)	(B1)	(B2)	(B3)	(B4)	(C1)	(C2)	(C3)	(D1)	(D2)	(D3)	(E1)	(Final)
应用	指标	0.10	0.30	0.70	0.90	1	2	8	16	Grid	Stroke	Center	ITM	ITM+MIR	ITM+MLM	*w/o* PVT	
TIR	$\mathcal{R}$@1	31.10%	33.50%	30.50%	30.70%	31.90%	30.30%	30.00%	32.20%	32.20%	31.40%	30.40%	30.40%	32.20%	32.90%	29.00%	**34.60%**
	$\mathcal{R}$@5	75.70%	76.00%	75.50%	73.80%	75.30%	75.60%	73.90%	76.90%	75.30%	76.10%	75.10%	74.10%	76.00%	76.20%	72.20%	**78.00%**
	$\mathcal{R}$@10	88.60%	88.70%	88.00%	88.60%	89.60%	88.60%	88.20%	88.60%	88.50%	89.20%	87.20%	83.50%	87.20%	88.60%	86.60%	**89.50%**
	Sum$\mathcal{R}$	195.40	198.20	194.00	193.10	196.80	194.50	192.10	197.70	196.00	196.70	192.70	188.00	195.40	197.70	187.80	**202.10**
	Diff	−6.70	−3.90	−8.10	−9.00	−5.30	−7.60	−10.00	−4.40	−6.10	−5.40	−9.40	−14.10	−6.70	−4.40	−14.30	–
ITR	$\mathcal{R}$@1	30.00%	29.90%	29.90%	28.50%	29.00%	29.70%	29.00%	28.90%	31.40%	31.10%	30.10%	29.30%	30.40%	28.40%	25.60%	**33.10%**
	$\mathcal{R}$@5	75.70%	74.90%	76.50%	75.00%	76.90%	77.10%	74.20%	77.30%	77.40%	74.50%	73.90%	70.80%	75.50%	76.30%	71.50%	**77.20%**
	$\mathcal{R}$@10	88.80%	89.00%	89.20%	88.20%	89.40%	87.70%	88.00%	89.90%	89.60%	88.50%	87.80%	86.80%	87.80%	88.80%	85.90%	**91.10%**
	Sum$\mathcal{R}$	194.50	193.80	195.60	191.70	195.30	194.50	191.20	196.10	198.40	194.10	191.80	186.90	193.70	193.50	183.00	**201.40**
	Diff	−6.90	−7.60	−5.80	−9.70	−6.10	−6.90	−10.20	−5.30	−3.00	−7.30	−9.60	−14.50	−7.70	−7.90	−18.40	–
M-CR	$\mathcal{A}$	98.16%	97.87%	98.09%	98.06%	98.03%	98.04%	98.11%	98.01%	98.12%	98.07%	98.04%	96.49%	97.11%	98.08%	97.92%	**98.26%**
	macro-$\mathcal{F}$	0.870	0.860	0.890	0.870	0.870	0.880	0.850	0.870	0.869	0.877	0.870	0.806	0.853	0.876	0.879	**0.896**
	Sum$\mathcal{C}$	185.16	183.87	187.09	185.06	185.03	186.04	183.11	185.01	185.02	185.77	185.04	177.09	182.41	185.68	185.82	**187.86**
	Diff	−2.70	−3.99	−0.77	−2.80	−2.83	−1.82	−4.75	−2.85	−2.84	−2.09	−2.82	−10.77	−5.45	−2.18	−2.04	–
S-CR	$\mathcal{A}$	93.10%	93.34%	93.36%	93.23%	93.29%	93.34%	93.32%	93.32%	93.37%	93.21%	93.59%	89.64%	90.87%	93.29%	92.90%	**93.57%**
	macro-$\mathcal{F}$	0.800	0.810	0.820	0.810	0.810	0.810	0.800	0.799	0.794	0.814	0.830	0.703	0.728	0.809	0.790	**0.829**
	Sum$\mathcal{C}$	173.10	174.34	175.36	174.23	174.29	174.34	173.32	173.22	172.77	174.61	176.59	159.94	163.67	174.19	171.90	**176.47**
	Diff	−3.37	−2.13	−1.11	−2.24	−2.18	−2.13	−3.15	−3.25	−3.70	−1.86	+0.12	−16.53	−12.80	−2.28	−4.57	–

时尚相关的语义信息，从而可验证MVLT的生成能力。

4.3 消融实验

掩码比例： 表3（a）给出了具有不同掩码概率r_v的四个变体模型，包括0.10（A1）、0.30（A2）、0.70（A3）、0.90（A4）和本文的设定0.50。$\mathcal{R}$@5指标随着掩码概率的增加而稳步上升，直到最优点（75.70% → 78.00%），然后性能出现急剧下降（73.80%）。本文认为，增加r_v会使MIR任务变得更加困难，因此使得MVLT模型能够在受限情况下更好地学习语义信息。然而，掩盖过多的视觉区域意味着失去更多的有效信息，从而导致表征学习变差。

掩码单元尺寸： 得益于PVT模型的灵活性，我们可以轻松地尝试不同尺寸下的掩码图像块设计。表3（b）中给出了四种不同尺寸α设定的掩码单元所对应的变体模型，包括：1（B1）、2（B2）、8（B3）、16（B4）和本文的设定4相比较。结果显示模型的性能对这个因素十分敏感，这也验证了适当的粒度设定对于学习鲁棒的时尚相关表征的重要性。

表4 对使用ImageNet [47]预训练PVT权重的消融研究。

	TIR		ITR		M-CR		S-CR	
	$\mathcal{R}$@5	$\mathcal{R}$@10	$\mathcal{R}$@5	$\mathcal{R}$@10	$\mathcal{A}$	macro-$\mathcal{F}$	$\mathcal{A}$	macro-$\mathcal{F}$
w/o PVT	72.20%	86.60%	71.50%	85.90%	97.92%	0.879	92.90%	0.790
w/ PVT	**78.00%**	**89.50%**	**77.20%**	**91.10%**	**98.26%**	**0.896**	**93.57%**	**0.829**
Diff	+5.80%	+2.90%	+5.70%	+5.20%	+0.34%	+1.7%	+0.67%	+3.9%

掩码方式： 如图8所示，本文为MIR任务设计了四种不同的掩码策略，即网格掩码策略（C1）、涂鸦掩码策略（C2）、中心掩码策略（C3）和本文的随机网格（Final）掩码策略，定量的性能结果差异请参见表3（c）。可以看出，使用随机网格掩码（Final）的模型取得了最佳性能，而其他三种策略则表现较差。本文分析其原因是，与网格（C1）和中心（C3）方式相比，随机网格掩码（Final）能够帮助MVLT构建全面的表征。类似于本文采取的策略（Final），涂鸦掩码策略（C2）中也随机掩码了给定图像，然而或多或少在单独图像块（patch）中留下了未掩码的视觉线索。由于所提出的掩码策略很大程度保留了图像的语义，从而增强了模型对于可视区域知识的学习，这使得模型更容易地预测掩码区域。

预训练目标： 表3（d）给出了四个不同的变体模型来研究每个预训练目标的贡献，

表5 Fashion-Gen数据集上零样本检索的结果比较。

	TIR			ITR		
	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑
ViLBERT (Zero-shot)	7.18%	18.73%	29.84%	8.99%	15.34%	26.14%
CLIP (Zero-shot)	16.30%	40.60%	55.60%	13.60%	43.10%	57.60%
MVLT OUR)	**34.60%**	**78.00%**	**89.50%**	**33.10%**	**77.20%**	**91.10%**

表6 MS-COCO 2014数据集的检索结果。†表示使用了额外的特征提取器（例如：Faster RCNN）。

	TIR task (5K Test)			ITR task (5K Test)		
	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑	$\mathcal{R}$@1↑	$\mathcal{R}$@5↑	$\mathcal{R}$@10↑
Unicoder-VL†	48.40%	76.70%	85.90%	62.30%	87.10%	92.80%
UNITER-Base†	**50.30%**	78.50%	87.20%	64.40%	87.40%	93.10%
ViLT-Base/32	41.30%	72.00%	82.50%	61.80%	86.20%	92.60%
MVLT OUR)	49.66%	**79.88%**	**87.50%**	**65.38%**	**90.04%**	**93.60%**

即ITM（D1），ITM+MIR（D2），ITM+MLM（D3）以及本文的ITM+MIR+MLM（Final）。在TIR任务中，相比于D1变体模型和D2变体模型，本文可以看到D3变体模型在$\mathcal{R}$@5指标上具有更好的表现：74.10%（D1）< 76.00%（D2）< 76.20%（D3）本文认为，MLM任务可以帮助模型更为全面地学习语言知识，因此它提供了一个更精确的查询，使模型召回更匹配的图像。在ITR任务中，将D2变体模型与D1和D3变体模型的$\mathcal{R}$@5指标相比较时，本文发现类似的结论：70.80%（D1）< 75.50%（D2）< 76.30%（D3）。这表明，更好的视觉学习对于匹配更为准确的商品描述文本极为重要。

加载预训练权重： 如表4所示，本文通过对比实验来证明加载ImageNet [47]预训练后PVT权重的重要性。若不加载预训练PVT权重，MVLT的性能大幅度下降（ITR任务的$\mathcal{R}$@5指标: 77.20%→71.50%，S-CR任务的$\mathcal{A}$指标: 93.57%→92.90%）。这是由于在大规模通用数据集上的预训练模型已经学习了颜色、纹理、形状等信息，因此其对特定领域的应用也会带来性能提升。

4.4 更多讨论

MVLT模型在通用领域中表现如何？ 为了进一步探究模型在通用领域中的潜力，本文在这里讨论两个扩展的问题。*a*）通用模型可以直接迁移到时尚领域中吗？受到通用视觉-语言模型影响力的启发，如在表5中进一步探究了两个典型的通用模型（例如，ViLBERT模型[17]和CLIP模型[48]）的零样本性能。这再次表明了在特定领域设计预训练模型的必要性和优越性。*b*）*MVLT*模型在通用领域中也有用吗？本文进一步验证了我们的模型在一般领域的潜在能力。表6展示了在MS-COCO 2014数据集[49]上的性能，其中MVLT模型遵循与[35]相同的训练标准。结果表明，与最新模型（例如：Unicoder-VL模型[50]、UNITER模型[18]和ViLT模型[35]）相比，本文模型取得了可喜的结果，而没有使用额外的训练数据或者训练过程中使用特殊的检索损失函数，这表明MVLT模型在扩展到通用场景时也是一个十分有前景的解决方案。

为什么金字塔架构和MIR有效？ 如引言章节所述，时尚领域具有两个未充分研究的问题。为了解决可迁移性问题，金字塔架构[22]以原始模态数据作为输入，无需复杂的数据预处理流程，从根本上缓解了工业应用场景中的负担。此外，MIR任务不需要手工标签，例如：分类标签、边界框或像素级别分割标签。针对特征粒度问题[51]，金字塔架构[22]提供了语义丰富的多尺度特征。结合MIR任务的设计，本文的框架可以表征具有多粒度的时尚领域知识（例如：连衣裙、V领），这些特点在该领域很有帮助且是迫切需要的。

对于语义理解任务（例如：检索任务[52]、分类任务），性能良好的视觉-语言模型可以作为良好的基础，并通过使用额外的解码器而易于应用在下游任务中（例如：文本到图像的合成[53]、

图像描述）。本文因为专注于时尚领域的表征学习任务，而没有进行图像描述实验。

MVLT模型与MAE模型[32]比较： MAE模型通过允许模型探索像素到像素的关联来学习通用的视觉表征，所以本文所提出的MVLT模型和MAE模型在这方面是相似的。但是本文的MVLT模型首次将类似的视觉重建预训练目标引入多模态研究中（例如：时尚领域）。

5 总结与展望

本文提出了一个视觉-语言框架 MVLT，并做出了两个贡献：1）掩码图像重建（MIR）预训练目标；2）端到端的预训练方案。实验和消融分析证明了其在各种匹配和生成任务上的优越性。MVLT模型在检索和识别任务上以较大的优势超过了现有的最优方法Kaleido-BERT，进一步促进了时尚领域的发展。本文所提出的端到端方法简化了真实场景下的工作流（如：数据预处理和模型训练），使得大型电子商务网站的开发和业务效率提高近50%。

为缓解真实电商应用中的存储和计算限制，未来本文将在时尚领域中继续探索更为高效的方法，如使用哈希编码 [54]、网络剪枝和知识蒸馏等技术。

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186, DOI: 10.18653/v1/N19-1423.

[2] G.-P. Ji, M. Zhuge, D. Gao, D.-P. Fan, C. Sakaridis, and L. Van Gool, "Masked vision-language transformer in fashion," *Machine Intelligence Research*, 2023.

[3] A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16-16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*. [Online]: PMLR, 2021.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *International conference on computer vision*. Montreal, Canada: IEEE, 2021, pp. 9992–10 002, DOI: 10.1109/ICCV48922.2021.00986.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30. Long Beach Convention Center, Long Beach, US: Curran Associates, Inc., 2017.

[6] T.-X. Sun, X.-Y. Liu, X.-P. Qiu, and X.-J. Huang, "Paradigm shift in natural language processing," *Machine Intelligence Research*, vol. 19, no. 3, pp. 169–183, 2022, DOI: 10.1007/s11633-022-1331-6.

[7] S. Agarwal, G. Krueger, J. Clark, A. Radford, J. W. Kim, and M. Brundage, "Evaluating clip: towards characterization of

broader capabilities and downstream implications," *arXiv preprint arXiv:2108.02818*, 2021, Available: https://arxiv.org/abs/2108.02818.

[8] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[9] J. Lin, R. Men, A. Yang, C. Zhou, M. Ding, Y. Zhang, P. Wang, A. Wang, L. Jiang, X. Jia *et al.*, "M6: A chinese multimodal pretrainer," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2021, p. 3251–3261, DOI: 10.1145/3447548.3467206.

[10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. PMLR, 2021, pp. 8821–8831.

[11] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion iq: A new dataset towards retrieving images by natural language feedback," in *Conference on computer vision and pattern recognition*. Nashville, TN, USA: IEEE, 2021, pp. 11 307–11 317, DOI: 10.1109/CVPR46437.2021.01115.

[12] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang, "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 2251–2260, DOI: 10.1145/3397271.3401430.

[13] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleidobert: Vision-language pre-training on fashion domain," in *Conference on computer vision and pattern recognition*. Nashville, TN, USA: IEEE, 2021, pp. 12 642–12 652, DOI: 10.1109/CVPR46437.2021.01246.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on computer vision and pattern recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778, DOI: 10.1109/CVPR.2016.90.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, vol. 28. Montreal, Quebec, Canada: Curran Associates, Inc., 2015.

[16] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *arXiv preprint arXiv:2001.07966*, 2020, Available: https://arxiv.org/abs/2001.07966.

[17] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in neural information processing systems*, vol. 32. Vancouver, Canada: Curran Associates, Inc., 2019.

[18] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Glasgow, United Kingdom: Springer, 2020, pp. 104–120, DOI: 10.1007/978-3-030-58577-8_7.

[19] W.-L. Hsiao, I. Katsman, C.-Y. Wu, D. Parikh, and K. Grauman, "Fashion++: Minimal edits for outfit improvement," in *International conference on computer vision*. Montreal, Canada: IEEE, 2019, pp. 5046–5055, DOI: 10.1109/ICCV.2019.00515.

[20] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, "Learning type-aware embeddings for fashion compatibility," in *European conference on computer vision*. Munich, Germany: Springer, 2018, pp. 405–421, DOI: 10.1007/978-3-030-01270-0_24.

[21] D.-P. Fan, M. Zhuge, and L. Shao, "Domain specific pre-training of cross modality transformer model," 2022, uS Patent App. 17/186,745.

[22] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *International conference on computer vision*. Montreal, Canada: IEEE, 2021, pp. 548–558, DOI: 10.1109/ICCV48922.2021.00061.

[23] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in *European conference on computer vision*. Glasgow, United Kingdom: Springer, 2020, pp. 1–17, DOI: 10.1007/978-3-030-58601-0_1.

[24] Z. Al-Halah and K. Grauman, "From paris to berlin: Discovering fashion style influences around the world," in *Conference on computer vision and pattern recognition*. Seattle, WA, USA: IEEE, 2020, pp. 10 133–10 142, DOI: 10.1109/CVPR42600.2020.01015.

[25] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, p. 5100–5111.

[26] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*. [Online]: PMLR, 2020.

[27] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for

image-text matching," in *European conference on computer vision*. Munich, Germany: Springer, 2018, pp. 212–228, DOI: 10.1007/978-3-030-01225-0_13.

[28] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal lstm for dense visual-semantic embedding," in *International conference on computer vision*. Venice, Italy: IEEE, 2017, pp. 1899–1907, DOI: 10.1109/ICCV.2017.208.

[29] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European conference on computer vision*. Glasgow, United Kingdom: Springer, 2020, pp. 121–137, DOI: 10.1007/978-3-030-58577-8_8.

[30] T. Arici, M. S. Seyfioglu, T. Neiman, Y. Xu, S. Train, T. Chilimbi, B. Zeng, and I. Tutar, "Mlim: Vision-and-language model pre-training with masked language and image modeling," *arXiv preprint arXiv:2109.12178*, 2021, Available: https://arxiv.org/abs/2109.12178.

[31] H. Bao, L. Dong, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," in *ICLR*, 2022, Available: https://openreview.net/forum?id=p-BhZSz59o4.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Conference on computer vision and pattern recognition*. New Orleans, LA, USA: IEEE, 2022, pp. 15 979–15 988, DOI: 10.1109/CVPR52688.2022.01553.

[33] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020, Available: https://arxiv.org/abs/2004.00849.

[34] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "Vx2text: End-to-end learning of video-based text generation from multimodal inputs," in *Conference on computer vision and pattern recognition*. Nashville, TN, USA: IEEE, 2021, pp. 7001–7011, DOI: 10.1109/CVPR46437.2021.00693.

[35] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594.

[36] M. Yan, H. Xu, C. Li, B. Bi, J. Tian, M. Gui, and W. Wang, "Grid-vlp: Revisiting grid features for vision-language pre-training," *arXiv preprint arXiv:2108.09479*, 2021, Available: https://arxiv.org/abs/2108.09479.

[37] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *Conference on computer vision and pattern recognition*. Nashville, TN, USA: IEEE, 2021, pp. 12 971–12 980, DOI: 10.1109/CVPR46437.2021.01278.

[38] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert

for video-and-language learning via sparse sampling," in *Conference on computer vision and pattern recognition*. Nashville, TN, USA: IEEE, 2021, pp. 7327–7337, DOI: 10.1109/CVPR46437.2021.00725.

[39] H. Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang, "E2e-vlp: End-to-end vision-language pre-training enhanced by visual learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021, p. 503–513.

[40] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Advances in neural information processing systems*, vol. 34. Curran Associates, Inc., 2021, pp. 24 206–24 221.

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Munich, Germany: Springer, 2015, pp. 234–241, DOI: 10.1007/978-3-319-24574-4_28.

[42] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," in *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 2131–2140.

[43] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, "Fashion-gen: The generative fashion dataset and challenge," in *International conference on machine learning Workshops*, 2018.

[44] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014, Available: https://arxiv.org/abs/1411.2539.

[45] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *British Machine Vision Conference*. Newcastle, UK: BMVA Press, 2018.

[46] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, and X. Fan, "Position focused attention network for image-text matching," *arXiv preprint arXiv:1907.09748*, 2019, Available: https://arxiv.org/abs/1907.09748.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on computer vision and pattern recognition*. Miami, FL, USA: IEEE, 2009, pp. 248–255, DOI: 10.1109/CVPR.2009.5206848.

[48] 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.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Zurich, Switzerland: Springer, 2014, pp. 740–755, DOI: 10.1007/978-3-319-10602-1_48.

[50] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *AAAI Conference on Artificial Intelligence*. New York, NY, USA: AAAI Press, 2020, pp. 11 336–11 344.

[51] L. Y. Wu, D. Liu, X. Guo, R. Hong, L. Liu, and R. Zhang, "Multi-scale spatial representation learning via recursive hermite polynomial networks," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. Messe Wien, Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 1465–1473, DOI: 10.24963/ijcai.2022/204.

[52] D. Chen and et al., "Cross-modal retrieval with heterogenous graph embedding," in *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa, Portugal: Association for Computing Machinery, 2022, p. 3291–3300, DOI: 10.1145/3503161.3548195.

[53] D. Liu, L. Wu, F. Zheng, L. Liu, and M. Wang, "Verbal-person nets: Pose-guided multi-granularity language-to-person generation," *Transactions on Neural Networks and Learning Systems*, 2022, DOI: 10.1109/TNNLS.2022.3151631.

[54] Z. Zhang, H. Luo, L. Zhu, G. Lu, and H. T. Shen, "Modality-invariant asymmetric networks for cross-modal hashing," *Transactions on Knowledge and Data Engineering*, 2022, DOI: 10.1109/TKDE.2022.3144352.