arXiv:2304.08345v2 [cs.LG] 6 Jan 2025

VALOR：视觉-音频-语言全感知预训练模型与数据集

Jing Liu*、Sihan Chen*、Xingjian He、Longteng Guo、XinXin Zhu、
Weining Wang、Jinhui Tang，*IEEE*高级会员

摘要—本文提出了一种用于多模态理解与生成的**视**觉-**音**频-**语**言 **全**感知预**训**练模型（VALOR）。与广受研究的视觉-语言预训练模型不同，VALOR以端到端方式联合建模视觉、音频和语言之间的关系。它包含三个独立的编码器，用于学习单模态表征；还包含一个解码器，用于多模态条件文本生成。我们设计了两个前置任务来预训练VALOR模型，分别是多模态分组对齐（MGA）和多模态分组描述（MGC）。MGA将视觉、语言和音频映射到同一公共空间，同时实现视觉-语言、音频-语言和视听-语言对齐。MGC则学习如何在视觉、音频或两者同时作为条件时生成文本词元。为了推动视觉-音频-语言预训练研究，我们构建了一个名为VALOR-1M的大规模高质量三模态数据集，其中包含100万个带有人工标注视听描述的可听视频。大量实验表明，VALOR能够学习强大的多模态相关性，并泛化到各种下游任务（如检索、字幕生成和问答），且支持不同的输入模态（如视觉-语言、音频-语言和视听-语言）。VALOR在一系列公开跨模态基准测试上取得了新的最先进性能。代码和数据可在项目页面获取https://casia-iva-group.github.io/projects/VALOR。

索引词—视觉-音频-语言预训练，多模态理解，多模态预训练

1 引言

作为人类，我们通过多种媒介（如观看、阅读、聆听、触摸或嗅闻）从环境中感知信息，并进一步基于这些多模态线索理解或与世界互动。理想的智能系统也应当模仿这一点，具备跨模态理解与生成能力。人们已广泛研究各种跨模态应用，其中视觉-语言任务占据主要部分，包括文本到视觉检索 [1], [2], 视觉描述 [3], [4], [5]和视觉问答 [6], [7]。幸运的是，受自然语言处理领域自监督预训练方法取得的巨大成功启发， [8], [9],[10], 视觉-语言预训练得到了迅速发展，并在各种视觉-语言基准测试上取得了优于传统方法的主导性能。

然而，我们认为，仅对视觉与语言之间的关系进行建模，远不足以构建强大的多模态系统；此外，引入音频模态以构建三模态交互也是必要的。一方面，音频信号通常包含与视觉互补的语义信息，因此，利用三种模态可以帮助机器更好地理解周围全面且准确地。如图1 所示，我们只能通过观察视频帧了解房间内发生的事情，却可能无法察觉外面的警车，除非听到警笛声。另一方面，在统一的端到端框架中对三种模态进行建模，可以增强模型的泛化能力，并有利于各种视觉-语言、音频-语言、视听-语言以及视觉-音频下游任务。

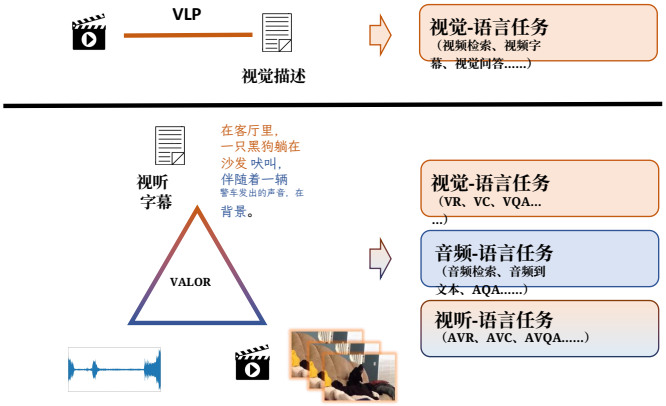

图 1：VALOR 使用相关联的视觉-音频-语言数据进行预训练，并可泛化至多种任务。AVR/VR/AR 分别表示文本到视听/视觉/音频检索，AVC/VC/AC 分别表示视听/视觉/音频字幕生成，AVQA/VQA/AQA 分别表示视听/视觉/音频问答。点击按钮播放音频。

• *Jing Liu*（通讯作者）和 *Sihan Chen* 对本文贡献相同，二人均就职于中国科学院大学人工智能学院、中国科学院自动化研究所模式识别国家重点实验室。电子邮箱： {*jliu, sihan.chen*}*@nlpr.ia.ac.cn*。*Xingjian He*、*Longteng Guo*、*XinXin Zhu* 和 *Weining Wang* 就职于中国科学院自动化研究所模式识别国家重点实验室。电子邮箱：*longtengguo1993@gmail.com*, {*xingjian.he, xinxin.zhu, weining.wang*}*@nlpr.ia.ac.cn*。*Jinhui Tang* 就职于南京理工大学计算机科学与工程学院。电子邮箱：*jinhuitang@njust.edu.cn*

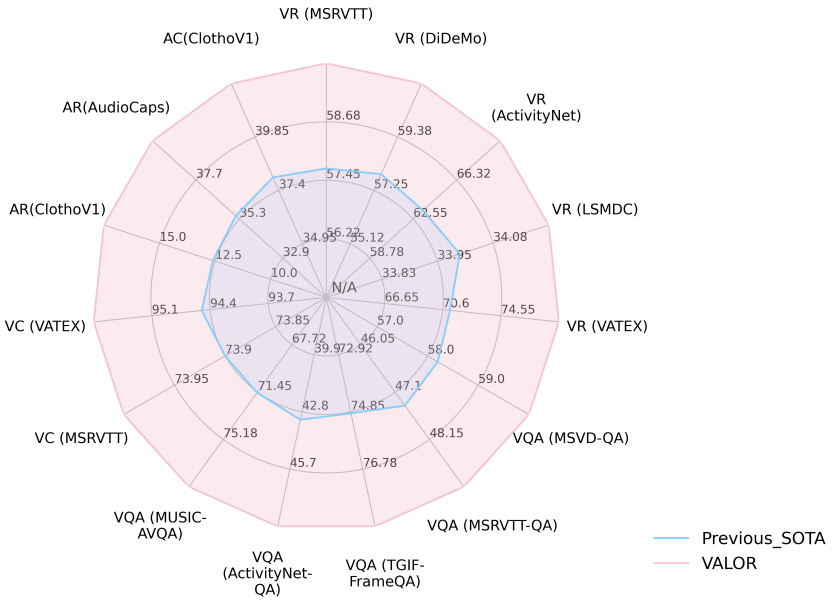

图2：与其他定制模型或基础模型相比，VALOR 在广泛的任务上取得了最先进的性能。VR、VC、VQA 分别代表文本到视频检索、视频描述生成和视频问答。

为此，如图 1所示，我们提出了一种**视**觉-**音**频-**语**言 **全**方位感知预**模**型（VALOR），以建立三种模态之间的通用联系，并实现三模态理解与生成。如图 5所示，VALOR 使用三个单模态编码器分别对视觉、音频和语言进行编码，并使用多模态解码器进行条件文本生成。我们设计了两个前置任务，即多模态分组对齐（MGA）和多模态分组描述（MGC），使 VALOR 具备处理判别式和生成式任务的能力。具体而言，MGA 将三种模态投射到同一个公共空间，并通过对比学习，在视觉-语言、音频-语言和视听-语言这三个模态组之间建立细粒度对齐。MGC 要求模型在视觉、音频或两者的交叉注意力层条件下，重建随机遮蔽的文本词元。得益于模态分组策略，VALOR 能够学习如何根据不同的模态组合对齐或生成文本，并可将这些能力迁移到各种跨模态下游任务中，包括视频、音频和视听检索、字幕生成或问答。

此外，我们认为，强相关的视觉-音频-语言三元组对于训练强大的三模态模型必不可少。当前的公开视觉-语言数据集无法用于三模态预训练，原因如下：i）所有图像-语言数据集，以及部分视频-语言数据集（如 WebVid-2.5M [13] ）都不包含音频信号。ii）即使部分视频-语言数据集（如 HowTo100M[11] 和 HD VILA 100M [12] ）包含音频模态，其中的音频也仅限于多样性较低的人类语音，文本则是自动语音识别转录文本，而非客观描述，且与语音内容重叠。为克服上述限制，我们构建了大规模、高质量的视觉-音频-语言数据集 VALOR-1M，以推动三模态预训练研究。该数据集包含一百万个开放领域可听视频，每个视频均由人工标注一条视听描述，同时描述音频和视觉内容。VALOR-1M 强大的视觉-语言和音频-语言相关性，以及其庞大的规模，使其成为三模态预训练的最佳选择。除 VALOR-1M 外，我们还建立了新的基准 VALOR-32K，用于评估视听能力。它包含两项新任务：视听检索（AVR）和视听描述（AVC）。

我们进行了大量消融研究，以展示所提出的 VALOR 模型和模态分组策略的有效性。定量和定性结果均证明，VALOR 能够有效利用视听线索完成 AVR 和 AVC 任务。我们在一系列公开的视频-语言、图像-语言和音频-语言基准测试上对 VALOR 进行了广泛验证，并取得了一系列新的最先进结果。具体而言，如图2所示，在包括 MSRVTT、DiDeMo、ActivityNet、LSMDC 和 VATEX 的文本到视频检索基准上，VALOR 的表现超过此前最先进的方法，R@1 分别提升 3.8%、6.2%、12.7%、0.6% 和 10.4%；在包括 MSRVTT-QA、MSVD-QA、TGIF-FrameQA 和 ActivityNet-QA 的开放式视频问答基准测试上，准确率分别提升 3.8%、3.4%、5.1% 和 12.5%；在包括 ClothoV1 和 AudioCaps 的文本到音频检索基准上，R@1 分别提升 38.9% 和 13.0%。此外，在 VATEX 字幕生成基准上，VALOR 的表现优于 GIT2 大模型，且仅使用 0.26% 的训练数据和 11.6% 的参数。 [14]

总而言之，本工作的贡献可概括如下：

I）我们提出了一种全感知预训练模型（VALOR），它建立了视觉、音频和语言之间的关联，以实现三模态理解与生成。

II）我们引入了采用模态分组策略的 MGA 和 MGC 预训练任务，以提升模型的泛化



图 3：视频-语言预训练数据集的可视化，包括 HowTo100M [11], HD VILA 100M [12], WebVid-2,5M [13] 和 VALOR-1M。点击按钮播放音频。

能力，使其能够处理不同的模态输入。

III）我们提出了 VALOR-1M 数据集，这是首个用于推动视觉-音频-语言研究的大规模人工标注三模态数据集；同时提出了 VALOR-32K 基准，用于评估视听-语言能力。

IV）在 VALOR-1M 和当前公开的视觉-语言数据集上进行预训练后，VALOR 在一系列跨模态基准测试中取得了新的最先进的性能，并实现了显著提升。

2 相关工作

本节首先介绍用于多模态预训练的常见跨模态数据集。随后，我们回顾视觉-语言预训练方法。最后，我们介绍利用视觉和文本之外更多模态进行视频-语言学习的典型方法。

2.1 多模态预训练的跨模态数据集

一般而言，理想的视觉-语言预训练数据集应满足两项基本要求：规模足够大，以及视觉-文本相关性强。考虑到句子级字幕标注比词级标签标注耗费的资源多得多，一些方法尝试收集包含人类语音的视频，并提取自动语音识别转录文本作为字幕。例如，Miech 等人收集了 HowTo100M [11],，该数据集包含来自 122 万个带旁白的教学类 YouTube 视频的 1.36 亿个视频片段，已成为早期视频-语言预训练方法所使用的主流数据集。Zellers 等人沿用这一方法，提出了 YT-Temporal- 1.8亿 [15] ，其中包含来自 600 万个 YouTube 视频的 1.8 亿个片段。Xue 等人收集了 HD VILA 100M [12] ，其中包含来自 330 万个 YouTube 视频的 1 亿个片段，具有更多样的数据和更高的图像分辨率。

然而，尽管这条途径能够轻松扩展，以获取大量视频-文本对，但字幕质量并不理想。除了可能存在语音识别错误外，自动语音识别转录文本通常传达的是说话者的主观想法和观点，而非对静态物体和正在发生的事件的客观描述。即使某些转录文本确实反映了视觉内容，也存在时间错位问题：它们可能对应于 {v1} 之前或之后的视频片段。

为了解决这一问题，同时兼顾数量与质量，Bain 等人沿用了图像-语言 Conceptual Captions 数据集（CC3M [17], CC12M [18]）的收集流程，构建了 WebVid [13], ，其中包含 250 万个与替代文本配对的视频。替代文本有时虽不够流畅、也不完整，但总体上与视频内容的相关性强于自动语音识别转录文本，因此已被最新的视频-语言预训练方法广泛采用。然而，上述数据集都不支持视觉-音频-语言预训练，因为它们缺少音频-语言关联；这促使我们收集 VALOR-1M 数据集，以推动三模态预训练的发展。

2.2 视觉-语言预训练

受 BERT 成功的影响，视觉-语言预训练迅速发展。我们将其主要研究方向总结如下。

I)跨模态预训练框架设计。根据不同的网络架构，视觉-语言模型主要可分为双编码器范式[19], [20] 和融合编码器范式 [21], [22]。前者通过简单的点积，在编码器输出端轻量地融合视觉与语言，可高效用于跨模态检索和零样本分类。后者则使用协同注意力 [22] 或合并注意力 [21] 深度融合两种模态，适用于字幕生成或视觉问答等更细粒度的任务。此外，人们提出了多种自监督前置任务，以更好地进行跨模态特征表示学习，包括掩码语言建模（MLM） [21], 、掩码视觉建模（MVM） [21], [23], 、视觉-文本匹配（VTM） [21], [24],、视觉-文本对比学习（VTC） [13], [25],等。在视觉表征方面，早期方法分别使用离线目标检测器（例如 Faster-RCNN[26]）提取目标级图像特征，或使用 3D 卷积神经网络（例如 S3D [19]）提取片段级视频特征。随着视觉 Transformer[27], [28], 的兴起，只需向模型输入图像或稀疏采样的视频帧，即可统一图像-语言和视频-语言任务。

II)统一多任务建模。 这一系列工作尝试通过统一框架对不同任务进行通用建模，并移除特定任务微调头，从而更高效地利用预训练数据。VL-T5 [29] 首先

采用序列到序列框架对视觉-语言任务进行建模，例如视觉问答和视觉定位。之后，目标检测和文本到图像生成等细粒度定位任务也分别通过 [30], [31], [32], 边界框坐标标记化 [33] 或图像标记化[34], 纳入其中。除序列到序列框架外，一些工作还通过对比学习 [35] 或掩码语言建模[36]统一多种视觉-语言任务。然而，即使上述方法统一了多种任务，它们仍局限于视觉-语言领域。相比之下，VALOR 能够推广至视觉-音频-语言领域，并适用于部分感知和全感知任务。

III)**视觉-语言基础模型**。 使用海量数据和参数训练的视觉-语言模型通常称为大型模型或基础模型，训练时通常采用对比学习[37], [38], [39], [40], 语言建模 [14], [41], [42], [43], 或两者兼用 [44]。基础模型在视觉-语言基准测试中取得了占主导地位的表现。例如，Flamingo [42] 将模型规模扩大到 80.2B 个参数，并在 VQAv2 数据集上取得 84.0 的准确率；而 GIT2 [14] 将数据规模扩大到 12.9B 个图像-文本对，并在 COCO 图像描述基准上取得 149.8 的 CIDEr 分数。然而，由于对计算资源和数据存储的需求很高，且分布式训练过程复杂，从参数和数据维度扩展视觉-语言预训练模型的效率有限。相比之下，我们认为，VALOR 可以通过引入音频并建立三模态连接，从模态维度进行扩展，这种方式有效且效率更高。

2.3 利用辅助模态增强视频-语言理解

鉴于视频天然是一种多模态媒介，且每种模态都蕴含丰富的语义信息，一些方法利用更多模态来增强视频-语言学习。MMT [45] 提出了一种多模态 Transformer，融合七种模态专家，用于文本到视频检索。SMPFF [46] 还引入客观描述和音频特征，以改进视频描述生成。在大规模预训练场景中，音频和字幕是最常用的辅助模态，用于增强视频表征。UniVL [47], VLM [48] 和 MV-GPT [19] 融合视频与字幕模态，并在 HowTo100M 数据集上进行预训练，以实现视频描述生成。VALUE [49]进一步利用字幕增强来应对更多任务，包括视频检索和问答。关于音频增强，AVLNet [50]和 MCN [51]利用音频来增强文本到视频检索。VATT [52] 提出了一种用于文本-视频和视频-音频对齐的分层对比损失，但其目标是学习单模态表征，而非提升跨模态能力。MERLOT Reserve [15] 和 i-Code [53]也将视觉、音频和语言作为预训练输入，但与 VALOR 存在本质区别：i) 这些方法存在严重的预训练-微调不一致。具体而言，预训练阶段的音频-语言关系建立在人类语音与自动语音识别转录文本之间，而微调阶段则建立在一般音频与客观描述之间。相比之下，VALOR 在强相关的三模态数据集上进行训练，并保持预训练-微调一致性，因此能够泛化到视频-语言、音频-语言和视听-语言任务。ii) 这些方法仅针对视频问答等判别任务，而 VALOR 得益于统一架构和精心设计的预训练任务，能够处理判别、对比和生成任务。

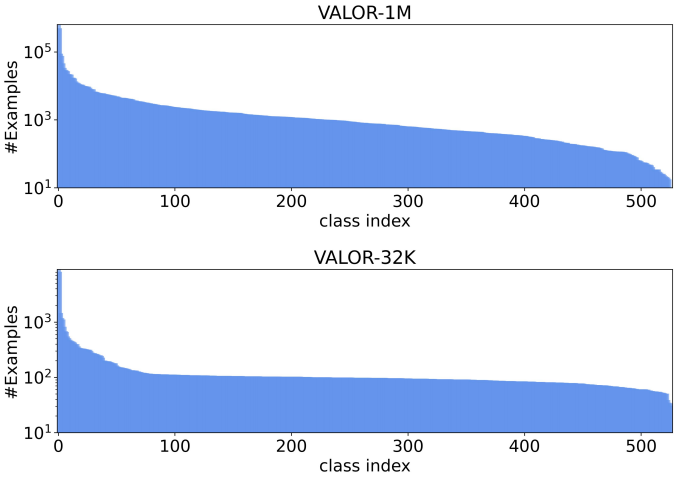

图4：VALOR-1M 和 VALOR-32K 中音频类别的分布。

3 用于视听-语言预训练的 VALOR 数据集

如第2.1节所述，字幕为自动语音识别转录文本或替代文本的视频-语言数据集，并非视觉-音频-语言预训练的最佳选择，因为文本句子与音频概念之间缺乏明确对应关系。为解决这一问题，我们通过标注公开视听数据，提出了一个视觉-音频-语言相关数据集 VALOR，用于三模态模型预训练和基准测试。在接下来的小节中，我们将详细介绍数据收集、标注和基准测试流程，并分析 VALOR 数据集的特征。

3.1 视听数据收集

理想情况下，视觉-音频-语言数据集中的视频应同时包含视觉和音频轨道，并具有较高的质量和多样性。为此，我们选用了来自 AudioSet [66],的大规模数据集的视频，该数据集是为音频事件识别而收集的。具体而言，AudioSet 包含从 YouTube 视频中截取的 200 多万个、时长为 10 秒的视频片段，每个视频都根据层级本体标注为 527 种音频类别之一。该数据集分为 200 万条非平衡训练集、2.2 万条平衡训练集和 2 万条评估集。在平衡训练集和评估集中，每种音频类别对应的视频数量大致相当，而非平衡训练集中的类别分布则不受限制。我们下载了 YouTube 链接仍然有效的 AudioSet 视频，过滤掉低质量和损坏的视频，最终获得约 100 万个视频。按照 [66], ，我们将数据集划分为 VALOR-1M（三模态预训练数据集）和 VALOR-32K（视听-语言下游基准数据集），

表 1：常见公开视频-语言预训练数据集和下游基准数据集的统计数据。音频：数据集是否包含音频。V-L：视觉-语言相关性。A-L：音频-语言相关性。#样本：视频/音频/图像的数量。#片段：视频片段或音频片段的数量。长度Cap：平均字幕长度。ACD：音频概念密度。

数据集	字幕	Task	领域	音频	V-L	A-L	#样本	#片段	长度Cap	音频概念密度 (%)
预训练数据集										
HowTo100M [11]	ASR	-	教学类	✓	✗	✗	1.22M	136M	4.0	3.4
HD_VILA_100M [12]	ASR	-	Open	✓	✗	✗	3.3M	103M	32.5	1.1
WebVid-2.5M [13]	替代文本	-	Open	✗	✓	✗	2.5M	2.5M	14.2	3.3
CC3M [17]	替代文本	-	Open	✗	✓	✗	3.3M	-	-	3.0
VALOR-1M	手动	-	Open	✓	✓	✓	118万	118万	16.4	9.7
下游基准测试										
MSVD [54]	手动	视频检索,视频字幕,视觉问答	Open	✗	✓	✗	2K	2K	7.0	6.7
MSRVTT [55]	手动	视频检索,视频字幕,视觉问答	Open	✓	✓	✗	7K	10K	9.3	4.7
VATEX [56]	手动	视频检索,视频字幕	Open	✓	✓	✗	41.3K	41.3K	14.3	4.1
YouCook2 [57]	手动	视频检索,视频字幕	烹饪	✓	✓	✗	2K	15.4K	8.8	4.3
DiDeMo [58]	手动	VR	Open	✓	✓	✗	10.5K	26.9K	8.0	6.3
ActivityNet [59]	手动	视频检索,视频字幕	动作	✓	✓	✗	20K	100k	13.5	3.4
LSMDC [60]	手动	VR	电影	✓	✓	✗	202	118K	9.1	2.5
Clotho视觉1 [61]	手动	音频检索、音频到文本	Open	✓	✗	✓	5.0K	5.0K	11.3	10.4
AudioCaps [62]	手动	音频检索、音频到文本	Open	✓	✗	✓	51.3K	51.3K	8.8	17.3
Pano-AVQA [63]	手动	AVQA	全景	✓	✓	✓	5.4K	5.4K	-	-
MUSIC-AVQA [64]	手动	AVQA	音乐	✓	✓	✓	9.3K	9.3K	-	-
AVQA [65]	手动	AVQA	Open	✓	✓	✓	57K	57K	-	-
VALOR-32K	手动	AVR、AVC	Open	✓	✓	✓	32K	32K	19.8	9.1

根据音频类别分布进行划分。具体而言，VALOR-1M 中的视频来自 AudioSet 的非平衡训练集，而 VALOR-32K 中的视频来自 AudioSet 的平衡训练集和评估集。如图所4示，与 VALOR-1M 相比，VALOR-32K 的音频类别分布更加均衡。

3.2 视听字幕标注

我们采用付费标注的方式，为 VALOR 数据集获取视听描述。考虑到这项标注任务是新颖的，而且比传统的视频描述标注更复杂，我们设计了一个三步交互式标注流程。

步骤 1，标注员培训。我们对 500 名标注员开展线上培训，强调描述应全面体现主体、活动、场景、物体和声音等重要组成部分。我们提供了一些视频-视听描述对，帮助标注员提前熟悉标注格式。我们还提供了一本将视频 ID 映射到其 AudioSet 标签的词典，并鼓励标注员在进行视听描述标注前，先查询这些标签作为先验参考。

步骤 2，第一阶段标注。在此阶段，我们向标注员提供 VALOR-32K 中的视频。我们会手动检查标注的描述，并反馈常见问题及对应的视频 ID。随后，我们要求标注员重新标注不符合要求的样例，以加深他们对标注要求的理解。

步骤 3，第二阶段标注。在这一阶段，标注员为 VALOR-1M 中的视频撰写视听描述。每条描述都会由三名标注员进一步审核，以确保标注质量；如果有超过一名标注员认为描述不合格，就需要重新标注。

整个标注和审核过程耗时约两个月。

3.3 VALOR-32K 基准

鉴于当前已建立的视听语言基准仅面向问答（AVQA）[63], [64], [65], ，我们建立了 VALOR-32K 基准，以扩展评估任务的范围。该基准包含两项任务：视听检索（AVR）和视听描述（AVC）。如图 8所示，AVC 要求模型为可听视频生成视听字幕；在 AVR 任务中，模型则需要根据给定的视听描述查询，检索最匹配的视频候选项。由于引入了音频模态，AVR 和 AVC 任务均比现有的文本到视频检索和视频描述生成任务更具挑战性。VALOR-32K 分为 25K/3.5K/3.5K 个视频，分别用于训练、验证和测试。AVR 和 AVC 任务的评估采用与视频检索和视频描述生成相同的评估指标。

3.4 VALOR 数据集的特点

VALOR 数据集是首个大规模视觉-音频-语言强相关数据集，其最大亮点在于丰富的音频概念和视听字幕。本小节将对 VALOR 数据集与公开的视频-语言数据集进行定量和定性比较。

定量比较。为了评估不同数据集的字幕中提及的音频概念丰富度，我们定义了一项名为音频概念密度（ACD）的指标。我们根据 [66] 提出的 632 类音频本体建立了一个音频概念集。具体来说，如果某个类别包含多个由逗号分隔的相似概念，我们就将其拆分，

将所有单词转换为小写并删除标点符号。最终，我们得到了759 个音频概念。给定一条字幕，我们先对其进行预处理，删除标点符号并转换为小写，然后检测每个音频概念是否存在。遍历整个数据集后，ACD 指标可按如下方式计算：

$$ACD = \frac{N_{AC}}{N_W} \tag{1}$$

其中 N_{AC} 等于检测到的音频概念总数，而N_W 为单词总数。如表1所示，VALOR 数据集的 ACD 指标远高于其他视频-语言数据集。此外，VALOR-1M 和 VALOR-32K 的平均字幕长度分别为 16.4 和 19.8，远长于其他数据集，例如 WebVid-2.5M（14.2）和 CC3M（10.3），这得益于额外的音频相关描述和较高的标注质量。

定性比较。我们将 VALOR-1M 与基于 ASR 转录字幕的数据集（如 HowTo100M 和 HD VILA 100M），以及基于替代文本字幕的数据集 WebVid-2.5M 进行比较。如图3 所示，HowTo100M 数据集中的字幕是不完整的句子，甚至难以理解，更不用说其中的视觉-语言关联了。HD VILA 100M 中的字幕较为完整，但视觉-语言关联仍然较弱。具体而言，字幕转录自两个人讨论度假推荐的一段对话，但蓝天、木制标牌和树木等重要视觉概念完全没有体现在字幕中。WebVid-2.5M 中的字幕具有更强的视觉关联，涵盖了更多视觉概念，但包含的音频概念较少，也较少直接描述音频信号。相比之下，VALOR 的标注同时关注视觉和音频线索，如样本中提及的黑狗和沙发等视觉概念，以及警笛等音频概念。

4 VALOR 模型

我们希望 VALOR 模型满足以下要求。I）模型可以进行完全端到端的训练，避免预先提取视觉或音频特征，从而让单模态编码器能够联合调优，学习有利于视觉-音频-语言交互的表征。II）应学习跨模态对齐、判别能力和生成能力，以提升 VALOR 在更广泛跨模态任务中的适应能力。III）考虑到不同模态应用于不同的下游领域，VALOR 应学习更具通用性的跨模态能力，而非局限于特定的模态组合。为此，我们对模型架构和预训练任务进行了专门设计，并将在以下小节中详细介绍。

4.1 模型架构

如图所示 5，VALOR 由文本编码器、视觉编码器、音频编码器和多模态解码器组成。该架构将单模态表征学习交由彼此独立的编码器负责，这些编码器的参数可继承自预训练模型，从而加快收敛并提升性能。

文本编码器。BERT [8] 模型用作文本编码器。首先使用 BERT 的分词器对原始句子进行分词，其词汇表大小为 30522。输入是词嵌入与位置嵌入之和。输出文本特征为 $F_t \in \mathbb{R}^{N_t \times C_t}$，其中 N_t 和C_t 分别是预先设定的最大词元长度和隐藏维度。

视觉编码器。 我们尝试了两种视觉编码器，包括 CLIP [37] 和 VideoSwin Transformer [67]。这两种模型都可以将图像或视频信号作为输入。对于视频输入，我们从视频片段中稀疏采样 N_v 帧，并使用图块嵌入层对图块进行编码。输出特征为 $F_v \in \mathbb{R}^{N_v \times S_v \times C_v}$，其中 S_v 是序列长度， C_v 是隐藏维度。每一帧都独立输入 CLIP 编码器，而在 VideoSwin Transformer 中则通过时间窗口注意力实现帧间交互。对于图像输入， N_v 等于 1。

音频编码器。 使用在 AudioSet 上预训练的音频频谱变换器（AST）[68], [69] 作为音频编码器。给定一个音频波形，我们将其切分为多个时长为 5 秒的音频片段，并随机采样 N_a 个片段作为输入。音频片段被转换为 64 维对数梅尔滤波器组特征，这些特征以 25 毫秒的汉明窗、每 10 毫秒计算一次。每个片段由此得到一张 64×512 频谱图。随后，将频谱图切分为图块，经过图块嵌入层处理后输入音频编码器。输出特征为 $F_a \in \mathbb{R}^{N_a \times S_a \times C_a}$，其中 S_a 是序列长度， C_a 是隐藏维度。

多模态解码器。 我们使用预训练的 BERT 作为多模态解码器。在每个 Transformer 块中，自注意力层和前馈层之间都加入了一个交叉注意力层，其参数采用随机初始化。在交叉注意力层中，文本特征关注条件特征，这些条件特征可以是视频输出特征、音频特征或二者的拼接。除交叉注意力层外，多模态解码器与文本编码器共享参数。

4.2 视觉-音频-语言跨模态学习

我们提出多模态分组对齐（MGA）和多模态分组描述（MGC）任务，以开展统一的视觉-音频-语言学习。它们分别基于模态分组策略，通过对比学习和因果掩码语言建模实现。我们主要考虑三种模态组：文本-视觉组（T-V）、文本-音频组（T-A）和文本-视听组（T-AV），分别对应三类主流下游任务（视觉-语言、音频-语言和视听-语言任务）。这一策略十分必要：可以想象，如果预训练期间只学习一种模态组（T-AV），那么视觉-语言和音频-语言任务的表现将受到限制，因为存在预训练-微调模态不一致的问题。

多模态分组对齐（MGA）。我们通过对比学习实现文本与模态 X 之间的细粒度对齐，其中 X 表示不同的模态，包括视觉（V）、音频（A）和视听（AV）。如果文本与模态 X 匹配，则将其视为正样本；如果不匹配，则视为负样本。我们按批次计算双向对比损失，以拉近

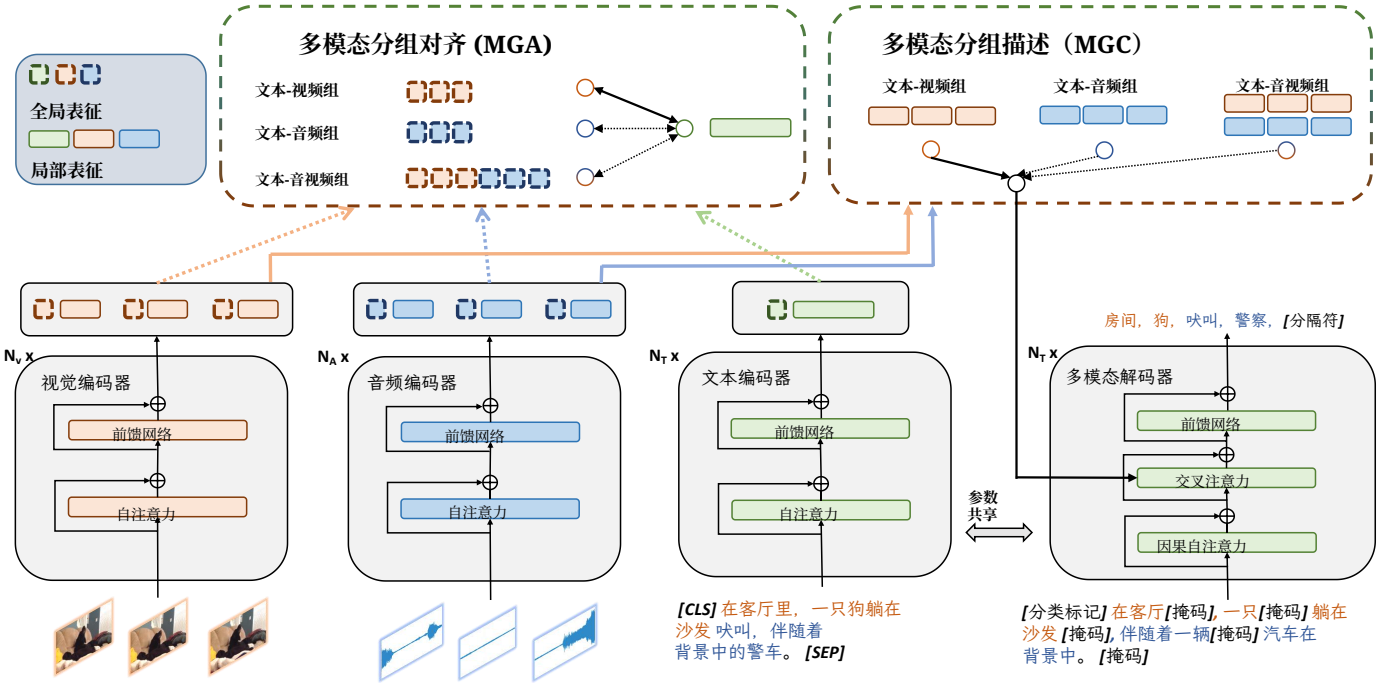

图5：VALOR整体预训练框架示意图。VALOR使用三个独立的编码器来获得单模态表征，并使用一个与文本编码器部分共享参数的多模态解码器进行文本生成。基于模态分组策略的MGA和MGC任务用于提升VALOR对不同类型任务和模态的泛化能力。

正样本之间的距离，并推远负样本，其公式如下：

$$L_{MGA(T-X)} = -\frac{1}{2}\sum_{i=1}^{B}\log\frac{\exp(s(T_i,X_i)/\tau)}{\sum_{j=1}^{B}\exp(s(T_i,X_j)/\tau)} - \frac{1}{2}\sum_{i=1}^{B}\log\frac{\exp(s(T_i,X_i)/\tau)}{\sum_{j=1}^{B}\exp(s(T_j,X_i)/\tau)} \tag{2}$$

其中， $s(\cdot,\cdot)$ 分别表示相似度函数和温度。

在计算相似度时 $s(\cdot,\cdot)$，我们不直接对齐文本与模态 X 的全局表征，而是建立每个文本词元与每个视频帧或音频片段之间的细粒度关联。具体而言，我们首先通过全局平均池化或使用 [CLS]词元特征，提取每个视频帧和音频片段的全局表征，然后通过三个线性投影层将三模态特征映射到同一个归一化语义空间。归一化后的特征分别表示为 $e_t \in \mathbb{R}^{N_t\times C}$、$e_v \in \mathbb{R}^{N_v\times C}$ 和 $e_a \in \mathbb{R}^{N_a\times C}$，其中 C 为共同的隐藏维度。视听特征 $e_{av} \in \mathbb{R}^{(N_v+N_a)\times C}$ 由 e_v 和 e_a 拼接而成。然后，通过计算 e_t 和 e_x 的点积得到细粒度相似度矩阵 $S_{TX} \in \mathbb{R}^{N_t\times N_x}$，其中 $e_x \in (e_v, e_a, e_{av})$；整体相似度则为双向分数之和。每个双向分数的计算方法是：沿矩阵的一个维度取 S_{TX} 的最大值，再沿另一个维度取平均值。考虑到不同文本词元、视觉帧或音频片段所包含的信息量并不相同，我们采用可学习加权平均，而非简单平均。权重通过将每个模态特征 e_t、e_v 和 e_a 输入相互独立的线性层，并通过 Softmax 函数进行归一化来获得。上述过程可表示为：

$$s(T,X) = \frac{1}{2}\sum_{i=1}^{N_t} f^i_{t,\theta}(e_t)\max_{j=1}^{N_x}(e^i_t)^{\mathrm{T}}e^j_x + \frac{1}{2}\sum_{j=1}^{N_x} f^j_{x,\theta}(e_x)\max_{i=1}^{N_t}(e^j_{av})^{\mathrm{T}}e^i_t \tag{3}$$

其中， f_θ 表示权重为 $W \in \mathbb{R}^{C\times 1}$的线性层。MGA总损失是三个分组对齐损失的平均值：

$$L_{MGA} = \frac{1}{3}(L_{MGA(T-AV)} + L_{MGA(T-V)} + L_{MGA(T-A)}) \tag{4}$$

多模态分组描述（MGC）。此任务采用因果掩码语言建模。具体而言，多模态解码器的输入文本词元会以 60% 的概率被随机替换为 [MASK] 词元，并将其输出特征输入多层感知机层，以重建原始词元。在多模态解码器的自注意力层中，使用因果注意力掩码来防止信息泄漏，并与自回归推理过程保持一致。文本、视觉和音频特征通过交叉注意力层进行融合。融合之前，我们首先沿时间维度展平，将 F_a 和F_v 重塑为二维，并通过线性层将它们转换为相同的隐藏维度，从而得到 $F_{a'} \in \mathbb{R}^{n_v\times C'}$ 和 $F_{v'} \in \mathbb{R}^{n_a\times C'}$，其中 $n_v = N_v \times S_v$、$n_a = N_a \times S_a$ 和 C' 等于多模态解码器的隐藏维度。融合后的视听

表2：VALOR_{B} 和VALOR_{L}的模型配置。#样本：所用视觉-文本对或视觉-音频-文本三元组的总数。Res：图像或视频帧的分辨率。

模型	三模态数据集	双模态数据集	#样本	视觉编码器	批量大小	迭代次数	参数	Res
VALOR_{B}	VALOR-1M	WebVid-2.5M+CC3M	6.5M	Video Swin $_{\text{B}}$	512	200K	342M	224
VALOR_{L}	VALOR-1M	WebVid-2.5M+CC14M+HD VILA 10M	33.5M	CLIP_{L}	1024	500K	593M	224

特征 $F_{av} \in \mathbb{R}^{(n_v+n_a)\times C'}$ 是将它们沿序列维度拼接而成的。以模态 X 为条件的 MGC损失可表示为：

$$L_{MGC(T-X)} = -\mathbb{E}_{(T,X)\in D}\log P(T_m|T_{<m}, F_x) \quad (5)$$

其中，D 、 T_m 和 $T_{<m}$ 分别表示训练批次、掩码词元以及当前掩码词元之前的词元，而 $F_x \in (F_{v'}, F_{a'}, F_{av})$。MGC 损失是三个分组描述生成损失的平均值：

$$L_{MGC} = \frac{1}{3}(L_{MGC(T-AV)} + L_{MGC(T-V)} + L_{MGC(T-A)}) \quad (6)$$

在每个训练步骤中，MGA和MGC会同时优化，并通过可调超参数 α 控制两个任务的比例，因此整体训练损失定义为：

$$L = \alpha L_{MGA} + L_{MGC} \quad (7)$$

4.3 适配下游任务

得益于上述介绍的MGA和MGC预训练任务，VALOR可以轻松适配不同类型的下游任务和模态。对于检索任务（AVR、VR、AR），我们分别使用$L_{MGA(T-AV)}$、$L_{MGA(T-V)}$、$L_{MGA(T-A)}$作为训练目标，且不使用多模态解码器。推理时，我们通过公式3计算每个查询与所有候选项之间的相似度分数，并对所有候选项进行排序。

对于字幕生成任务（AVC、VC、AC），我们分别使用 $L_{MGC(T-AV)}$、$L_{MGC(T-V)}$、 $L_{MGC(T-A)}$作为训练目标。推理时，文本词元以自回归方式生成。具体来说，输入“[CLS] [MASK]”以预测第一个词元 [TK1]，，输入“[CLS] [TK1] [MASK]”以预测下一个词元。重复此过程，直到生成[SEP] 词元。

对于问答任务（AVQA、VQA、AQA），我们将其建模为生成式问题，因此答案可以从整个词汇表中预测，而不是从预定义的高频答案候选集前k项中选取。训练期间，与字幕生成任务一样，使用MGC损失作为训练目标。具体来说，将问题词元和答案词元拼接后输入解码器，仅屏蔽答案词元，而问题词元则全部保持可见。多模态解码器中的自注意力掩码对问题词元采用双向机制，对答案词元采用因果机制。答案推理过程同样采用自回归方式。

5 实验

本节首先介绍基本实验设置，包括预训练数据集、下游基准测试和实现细节。之后，我们将 VALOR 与各种基准测试中的最先进方法进行比较。最后，我们呈现详细的消融研究，以证明所提出方法的有效性，并可视化 VALOR 的预测结果。

5.1 实验设置

5.1.1 预训练数据集

以下 4 个数据集用于 VALOR 的预训练。**VALOR-1M** 是提出的三模态数据集，包含一百万个开放领域可听视频，以及人工标注的视听字幕。

WebVid-2.5M [13] 是一个通过网络爬取构建的数据集，包含约 250 万个与替代文本配对的视频。近期还发布了其规模更大的版本 WebVid-10M，但本文未使用该版本。

CC14M由一系列图像语言数据集组成，包括 MSCOCO [90], Visual Genome [91],SBU[92],CC3M [17]和 CC12M [18]，，共包含 1400 万张图像或 2000 万个图像-文本对。由于下载时有太多图像无效，我们未使用 SBU 数据集。

HD_VILA 100M [12]是一个高分辨率开放领域视频文本数据集，包含 1 亿个带有自动语音识别转录文本的视频。受存储空间限制，我们仅使用其中随机抽取的 1000 万个视频子集（HD VILA 10M）。

– –

5.1.2 下游任务

对于检索任务，我们在 9 个公开数据集上评估 VALOR，包括视频检索（MSRVTT [55],DiDeMo [58], LSMDC [60],ActivityNet [59], VATEX [56] 和 MSCOCO [90]）、音频检索（ClothoV 1 [61] 和 AudioCaps [62]）以及 AVR（提出的 VALOR-32K）。对于 DiDeMo 和 ActivityNet 数据集，我们遵循其他研究的做法，将多条较短的时间描述拼接成较长的句子，并评估段落到视频的检索效果。我们使用前 K 名召回率（R@K，K=1、5、10）作为评估指标。

对于字幕生成任务，我们在 7 个公开数据集上评估 VALOR，包括视频字幕（MSVD [54], MSRVTT、VATEX 和 MSCOCO）、音频到文本（ClothoV1 和 AudioCaps）以及 AVC（提出的 VALOR-32K）。我们使用 BLEU4（B4） [93], METEOR（M） [94],ROUGE-L（R） [95], CIDEr（C） [96]和 SPICE（S） [97]作为评估指标。推理过程中使用束搜索，束宽为 3。

对于开放式问答任务，我们在 6 个公开数据集上进行评估，包括视觉问答（MSVD-QA [98],MSRVTT-QA [98], ActivityNet-QA [99], TGIF-Frame QA[100], VQAv2 [101]）和 AVQA（MUSIC-AVQA [64]）。我们使用准确率作为评估指标。推理时，我们采用贪心搜索，从整个词汇表中不受限制地生成答案。

5.1.3 实现细节

所有模型均基于 PyTorch 框架，使用 8 张 Tesla A100 显卡进行训练。预训练学习率为 1e-4。Warm

表3：与 VALOR-32K 文本到视听检索基准及5个文本到视频检索基准上的最先进方法的比较。报告 R@1/R@5/R@10。#样本表示所使用的视觉-文本对或视觉-音频-文本三元组的数量。Mod 表示使用的模态，V、A、S 分别是视觉、音频和字幕的缩写。+DSL 表示在评估过程中使用双重 softmax [70]后处理。VALOR-32K 上的结果由我们使用其公开发布的代码得出。

方法	#样本	Mod	VALOR-32K	MSRVTT	DiDeMo	ActivityNet	LSMDC	VATEX
A组：使用 *<**10M***个样例进行预训练								
ClipBert [71]	5.6M	V	-	22.0/46.8/59.9	20.4/48.0/60.8	-	-	-
冻结 [13]	6.1M	V	32.9/60.4/71.2	32.5/61.5/71.2	31.0/59.8/72.4	-	15.0/30.8/39.8	-
BridgeFormer [72]	5.5M	V	-	37.6/64.8/75.1	37.0/62.2/73.9	-	17.9/35.4/44.5	-
MILES [73]	5.5M	V	-	37.7/63.6/73.8	36.6/63.9/74.0	-	17.8/35.6/44.1	-
OA-Trans [74]	5.5M	V	-	35.8/63.4/76.5	34.8/64.4/75.1	-	18.2/34.3/43.7	-
Nagrani 等人 [75]	1.03M	V+A	-	35.8/65.1/76.9		-	-	-
LF-VILA [76]	8.5M	V	-	-	35.0/64.5/75.8	35.3/65.4/-	-	-
VALOR^-_B (我们)	5.5M	V	43.3/70.3/80.0	36.2/64.7/75.4	43.2/73.9/82.4	37.5/67.9/80.4	20.0/39.1/49.0	59.4/90.5/95.4
VALOR$_B$(我们)	6.5M	V+A	**67.9/89.7/94.4**	**43.0/72.2/82.1**	**52.2/80.8/86.8**	**50.5/79.6/89.1**	**25.1/45.8/55.2**	**67.5/94.1/97.4**
B组：使用 *>**10M***个样例进行预训练，或继承 ***CLIP*** 模型权重								
SINGULARITY [77]	17M	V	-	41.5/68.7/77.0	53.9/79.4/86.9	47.1/75.5/85.5	-	-
LAVENDER [36]	30M	V	-	40.7/66.9/77.6	53.4/78.6/85.3	-	26.1/46.4/57.3	-
MV-GPT [19]	53M	V+S	-	37.3/65.5/75.1	-	-	-	-
TACo [78]	136M	V	-	28.4/57.8/71.2	-	30.4/61.2/-	-	-
支持集 [79]	136M	V	-	30.1/58.5/69.3	-	29.2/61.6/-	-	44.9/82.1/89.7
MMT [45]	136M	V+A	-	26.6/57.1/69.6	-	28.7/61.4/-	12.9/29.9/40.1	-
AVLNet [50]	136M	V+A	21.6/47.2/59.8	22.5/50.5/64.1	-	-	11.4/26.0/34.6	-
Gabeur等人 [80]	136M	视觉+A+S	-	28.7/59.5/70.3	-	29.0/61.7/-	-	-
All-in-one [81]	138M	V	-	37.9/68.1/77.1	32.7/61.4/73.5	22.4/53.7/67.7	-	-
VIOLET [82]	186M	V	-	34.5/63.0/73.4	32.6/62.8/74.7	-	16.1/36.6/41.2	-
CLIP4Clip [83]	-	V	43.4/69.9/79.7	44.5/71.4/81.6	43.4/70.2/80.6	40.5/72.4/-	22.6/41.0/49.1	55.9/89.2/95.0
TS2-Net [84]	-	V	-	49.4/75.6/85.3	41.8/71.6/82.0	41.0/73.6/84.5	23.4/42.3/50.9	59.1/90.0/95.2
X-CLIP [85]	-	V	-	49.3/75.8/84.8	47.8/79.3/-	46.2/75.5/-	26.1/48.4/46.7	-
ECLIPSE [86]	-	V+A	-	-	44.2/-/-	45.3/75.7/86.2	-	-
DCR [87]	-	V	-	50.2/76.5/84.7	49.0/76.5/84.5	46.2/77.3/88.2	26.5/47.6/56.8	65.7/92.6/96.7
混元 TVR±DSL [88]	-	V	-	55.0/80.4/86.8	52.1/78.2/85.7	57.3/84.8/93.1	29.7/46.4/55.4	-
CLIP-VIP+DSL [20]	100M	V	-	57.7/80.5/88.2	55.3/82.0/89.3	61.4/85.7/92.6	30.7/51.4/60.6	-
InternVideo+DSL [89]	147.6M	V	-	55.2/-/-	57.9/-/-	62.2/-/-	34.0/-/-	71.1/-/-
VALOR$_L$(我们)	33.5M	V+A	73.2/91.6/95.4	54.4/79.8/87.6	57.6/83.3/88.8	63.4/87.8/94.1	31.8/52.8/62.4	76.9/96.7/98.6
VALOR$_L$+DSL（我们）	33.5M	V+A	**80.9/93.9/97.1**	**59.9/83.5/89.6**	**61.5/85.3/90.4**	**70.1/90.8/95.3**	**34.2/56.0/64.1**	**78.5/97.1/98.7**

up，并使用线性学习率衰减调度器。对于消融研究，除非另有说明，我们使用在 Kinetics-400 上预训练的 Video S win Transformer-small 作为视觉编码器。我们在 VALOR-1M 上预训练 4 个 epoch，批量大小为 512。

为了比较当前最先进的方法，我们训练了两个不同规模的模型，即 VALOR$_B$ 和 VALOR$_L$，具体配置见表2。与 VALOR$_B$相比，VALOR$_L$ 使用了更多训练数据、更大的批量大小、更多迭代次数，并采用了更强大的视觉编码器进行训练。除视觉编码器不同外，两个模型使用相同的文本/多模态编码器（BERT$_B$）和音频编码器（AST）。每次迭代时，我们根据预先设定的权重抽取一个数据集；如果抽取到双模态数据集，则不使用音频。预训练期间，每个视频抽取 1 个视频帧和 1 个音频片段。微调期间，我们使用特定于任务的学习率和采样数量。

5.2 与当前最先进方法的比较

5.2.1 视频-语言基准

文本到视频检索。如表3所示，在 VALOR-32K、MSRVTT、DiDeMo 和 LSMDC 数据集上，VALOR$_B$的表现明显优于 A 组中的所有模型。在 ActivityNet 和 VATEX 数据集上，VALOR$_B$ 甚至超越了 B 组中的所有模型，而预训练数据仅有 650 万条，这体现了 VALOR 的高效性和有效性。我们还仅使用 WebVid-2.5M 和 CC3M 训练了一个基础级模型，且在预训练和微调阶段均未使用音频，记为 VALOR^-_B。通过比较 VALOR$_B$ 与 VALOR^-_B，我们可以发现，VALOR-1M 数据集和音频模态对 VALOR 的高性能至关重要。此外，与 B 组中的模型相比，VALOR$_L$ 在 MSRVTT、DiDeMo、ActivityNet、LSMDC 和 VATEX 数据集上取得了新的最先进水平，分别比此前最先进水平（R@1）高出 3.8%、6.2%、12.7%、0.6% 和 10.4%。我们将 VALOR 的显著提升归因于：i）构建视觉-音频-语言对齐，而非双模态对齐；ii）在文本与视听信号之间构建细粒度对齐，而非粗粒度对齐。

值得注意的是，VALOR 还优于额外利用音频、字幕或两种模态的方法 [19],[45], [50], [75], [86],，这证明了 VALOR 中细粒度三模态对齐建模的有效性，也体现了利用强相关三模态预训练数据的重要性。与 MSRVTT 和 LSMDC 等短视频检索数据集（<30s）不同，DiDeMo 和 ActivityNet 等长视频数据集（>1min）更具挑战性，因为长视频与段落之间的时间关系更加复杂。VALOR 显著优于专门针对长视频检索的方法 [76], [86] ，且无需任何额外技巧。这表明 VALOR 具有强大的泛化能力，因为其预训练期间仅接触过短视频（约 10 秒）。此外，与使用视频-文本匹配（VTM）损失训练模型的方法 [36], [77] 相比，VALOR 兼具更高的推理效率和更好的性能。

视频描述生成。如表4所示，VALOR$_B$在4项基准测试中优于A组的所有模型。在

表4：VALOR-32K视听描述基准和3个视频字幕生成基准上与最先进方法的比较。鉴于大多数方法都使用强化学习 [102]方法来提升模型在VATEX数据集上的性能，为了公平比较，我们也采用了这些方法，并用*标记相应结果。VALOR-32K上的结果由我们使用其公开发布的代码得出。

方法	#样本	Mod	VALOR-32K				MSVD				MSRVTT				VATEX			
			B@4	M	R	C	B@4	M	R	C	B@4	M	R	C	B@4	M	R	C
***A*组：使用 <*10M*个样例进行预训练**																		
ORG-TRL [103]	-	V	-	-	-	-	54.3	36.4	73.9	95.2	43.6	28.8	62.1	50.9	32.1	22.2	48.9	49.7
OpenBook [104]	-	V	-	-	-	-	-	-	-	-	42.8	29.3	61.7	52.9	33.9	23.7	50.2	57.5
SwinBERT [105]	-	V	5.4	10.7	27.2	27.3	58.2	41.3	77.5	120.6	41.9	29.9	62.1	53.8	38.7	26.2	53.2	73.0
SMPFF [46]	-	V+A	7.5	12.6	28.6	37.1	-	-	-	-	48.4	30.6	64.9	58.5	39.7	26.0	53.6	70.5
VIOLETv2 [23]	5.5M	V	-	-	-	-	-	-	-	139.2	-	-	-	58.0	-	-	-	-
VALOR$_{\bar{B}}$（我们）	5.5M	V	8.0	13.5	29.4	44.3	74.3	47.1	83.8	156.1	48.1	30.4	64.3	61.5	40.7	26.1	53.8	71.6
VALOR$_B$(我们)	6.5M	V+A	**8.9**	**14.8**	**30.8**	**55.7**	**76.1**	**48.0**	**85.2**	**162.1**	**53.8**	**32.3**	**67.0**	**66.6**	**41.9**	**26.6**	**54.6**	**73.9**
***B*组：使用 >*10M*个样例进行预训练**																		
LAVENDER [36]	30M	V	-	-	-	-	-	-	-	150.7	-	-	-	60.1	-	-	-	-
支持集 [79]	136M	V	-	-	-	-	-	-	-	-	38.9	28.2	59.8	48.6	32.8	24.4	49.1	51.2
VALUE [49]	136M	V+S	-	-	-	-	-	-	-	-	-	-	-	-	32.9	24.0	50.0	58.1
MV-GPT [19]	136M	V+S	-	-	-	-	-	-	-	-	48.9	38.7	64.0	60.0	-	-	-	-
GIT$_L$ [14]	20M	V	-	-	-	-	75.8	48.7	85.5	162.9	48.7	30.9	64.9	64.1	41.6*	26.2*	54.3*	72.5*
GIT [14]	800M	V	-	-	-	-	79.5	**51.1**	87.3	**180.2**	53.8	32.9	67.7	73.9	41.6*	28.1*	55.4*	91.5*
GIT2 (5.1B) [14]	12.9B	V	-	-	-	-	82.2	52.3	88.7	185.4	54.8	33.1	68.2	75.9	42.7*	28.8*	56.5*	94.5*
VALOR$_L$(我们)	33.5M	V+A	**9.6**	**15.4**	**31.8**	**61.5**	80.7	51.0	87.9	178.5	54.4	32.9	68.0	74.0	**45.6***	**29.4***	**57.4***	**95.8***

表5：与 最先进的方法在5个开放式视频问答和视听问答基准测试上的比较。

方法	#样本	Mod	MSRVTT-QA	MSVD-QA	TGIF-FrameQA	ActivityNet-QA	MUSIC-AVQA
***A*组：使用 <*10M*个样例进行预训练**							
QueST [106]	-	V	34.6	34.6	59.7	-	-
MUSIC-AVQA [64]	-	V+A	-	-	-	-	71.5
ClipBERT [71]	5.6M	V	37.4	-	60.3	-	-
VIOLET [23]	5.5M	V	44.5	54.7	72.8	-	-
Clover [107]	5.5M	V	43.9	51.9	71.4	-	-
VALOR$_{\bar{B}}$（我们）	5.5M	V	44.5	54.9	73.0	43.7	74.8
VALOR$_B$（我们）	6.5M	V+A	**46.7**	**56.4**	**74.5**	**44.8**	**76.6**
***B*组：使用 >*10M* 个样例进行预训练**							
SINGULARITY [77]	17M	V	43.5	-	-	43.1	-
LAVENDER [36]	30M	V	45.0	56.6	73.5	-	-
JustAsk [108]	69M	V	41.5	46.3	-	38.9	-
MV-GPT [19]	53M	V+S	41.7	-	-	39.1	-
MERLOT [109]	180M	V	43.1	-	69.5	41.4	-
All-in-one [81]	228.5M	V	46.8	48.3	66.3	-	-
Flamingo（80B） [42]	2.3B	V	47.4	-	-	-	-
FrozenBiLM [110]	10M	V	47.0	54.8	68.6	43.2	-
InternVideo [89]	147.6M	V	47.1	55.5	72.2	-	-
VideoCoCa（2.1B） [111]	4.8B	V	46.0	56.9	-	-	-
GIT$_L$ [14]	20M	V	42.7	55.1	71.9	-	-
GIT [14]	800M	V	43.2	56.8	72.8	-	-
GIT2（5.1B） [14]	12.9B	V	45.6	58.2	74.9	-	-
VALOR$_L$（我们）	33.5M	V+A	**49.2**	**60.0**	**78.7**	**48.6**	**78.9**

B组中，我们主要将VALOR与GIT模型 [14],这一最近提出的大规模生成式预训练模型进行比较，该模型在许多视觉描述生成基准测试中取得了最先进水平的结果。具体而言，GIT根据不同的参数量和数据规模分为四种规模，分别命名为GIT$_B$、GIT$_L$、GIT和GIT2。需要注意的是，GIT$_L$ 使用了与VALOR$_L$ 相当规模的训练数据和相同的视觉编码器，即CLIP$_L$；而GIT使用了更大的视觉编码器（由Florence [39]预训练的CoSwin模型）和更大的数据规模。GIT2甚至使用了4.8B的DaViT [119] 作为视觉编码器，并将12.9B个视觉-文本对作为训练数据。通过比较结果可以发现，在全部三个基准测试的大多数指标上，VALOR$_L$ 以巨大优势优于GIT$_L$和GIT。此外，在VATEX基准测试中，VALOR$_L$ 甚至优于GIT2，而其参数量（11.6%）、数据规模（0.26%）和图像分辨率（224对384）都小得多。这些结果表明，与从模型参数量和数据规模维度进行扩展相比，学习视听条件文本生成（从模态维度扩大预训练模型）更加高效且有效。

开放式视频问答。 如表 5所示，VALOR$_B$在五个基准测试中超越了A组的所有模型。在B组中，FrozenBiLM使用了与VALOR相同的视觉编码器$_L$ （即CLIP$_L$），以及更强大的解码器（一个包含8.9亿参数的DeBERTa-V2-XLarge模型 [120]）。Flamingo拥有 135×参数和68.7× 比VALOR更多的训练数据$_L$。Video-CoCa继承了CoCa的权重，而CoCa拥有3.5× 参数和143.3× 比VALOR更多的训练数据$_L$。GIT2拥有8.6×参数和382.1× 比VALOR更多的训练数据$_L$。尽管参数量和训练数据都少得多，VALOR$_L$仍在MSRVTT-QA、MSVD-QA、TGIF-FrameQA和ActivityNet-QA基准测试中取得了新的最先进水平表现，并分别以3.8%、3.4%、5.1%和12.5%的优势超越了此前的最先进方法。在视听问答基准测试MUSIC-AVQA中，VALOR$_B$ 和VALOR$_L$ 分别将基线提升了7.1%和10.3%。

5.2.2 音频-语言基准测试

如表6所示，在文本到音频检索任务方面，VALOR在ClothoV1上取得了新的最先进水平，

表6：4个音频-语言基准上的与最先进方法的比较。

（a）文本到音频检索

方法	ClothoV1 AudioCaps					
	R@1	R@5	R@10	R@1	R@5	R@10
Oncescu 等人 [112]	9.6	-	40.1	25.1	-	73.2
Nagrani 等人 [75]	12.6	-	45.4	35.5	-	**84.5**
VALOR$_B$	**17.5**	**42.7**	**55.3**	**40.1**	**73.9**	83.1

(b) 音频字幕生成

方法	ClothoV1				AudioCaps			
	B@4	M	R	C	B@4	M	R	C
Kim等人 [62]	-	-	-	-	21.9	20.3	45.0	59.3
徐等人 [113]	15.6	16.2	36.8	33.8	-	-	-	-
陈等人 [114]	15.1	16.0	35.6	34.6	-	-	-	-
徐等人 [115]	15.9	16.9	36.8	37.7	23.1	22.9	46.7	66.0
Koh等人 [116]	16.8	16.5	37.3	38.0	-	-	-	-
ACT [117]	-	-	-	-	25.2	22.2	46.8	67.9
Liu 等人 [118]	-	-	-	-	25.1	23.2	48.0	66.7
VALOR$_B$	**16.2**	**17.4**	**38.2**	**42.3**	**27.0**	**23.1**	**49.4**	**74.1**

表7：在3个图像-语言基准上与最先进方法的比较。标有 * 的结果表示其通过强化学习获得。VALOR 在这三个基准测试中均采用 392 的图像分辨率。

方法	#样本	COCO检索			COCO图像描述		VQA v2	
		R@1	R@5	R@10	C	S	开发集	标准差
UNITER [21]	10M	52.9	79.9	88.0	-	-	73.82	74.02
Oscar [121]	10M	57.5	82.8	89.8	140.9*	25.2*	73.61	73.82
UFO [122]	10M	59.2	83.6	90.5	131.2	23.3	76.64	76.76
VinVL [123]	10M	58.8	83.5	90.3	140.0*	24.5*	76.52	76.60
ALBEF [25]	20M	60.7	84.3	90.5	-	-	75.84	76.04
METER [124]	10M	57.9	82.7	90.1	-	-	77.68	77.64
ALIGN [38]	1.8B	59.9	83.3	89.8	-	-	-	-
FILIP [125]	340M	61.2	84.3	90.6	-	-	-	-
Florence [39]	900M	63.2	85.7	-	-	-	80.16	80.36
BLIP [126]	135M	65.1	86.3	91.8	136.7	-	78.25	78.32
Flamingo（80B） [42]	2.3B	-	-	-	138.1	-	82.0	82.1
LEMON [127]	200M	-	-	-	145.5*	25.5*	-	-
SimVLM [41]	1.8B	-	-	-	143.3	25.4	80.03	80.34
CoCa [44]	4.8B	-	-	-	143.6	24.7	82.3	82.3
GIT$_L$ [14]	20M	-	-	-	144.6*	25.4*	75.5	-
GIT [14]	800M	-	-	-	151.1*	26.3*	78.6	78.8
GIT2（5.1B） [14]	12.9B	-	-	-	152.7*	26.4*	81.7	81.9
PALI (16.9B) [14]	1.6B	-	-	-	149.1	-	84.3	84.3
VALOR$_L$	33.5M	61.4	84.4	90.9	152.5*	25.7*	78.46	78.62

AudioCaps基准上分别以38.9%和13.0%的优势（R@1）超越了此前的最先进方法。Nagarani的方法在其提出的VideoCC3M数据集上预训练三模态模型 [75]，，该数据集通过从互联网收集与CC3M图像高度相似的视频构建，并直接将CC3M字幕用作视频字幕。相比之下，得益于VALOR-1M中的音频-语言相关性比VideoCC3M中的相关性更加明确且更强，VALOR在两个文本转音频基准上取得了明显更好的表现。此外，他们的方法需要分别为视频检索和音频检索预训练两个模型，而得益于所提出的模态分组策略，我们可以在单个预训练模型上针对视频、音频和视听检索任务进行微调。在音频描述基准方面，与直接在目标数据集上训练的模型相比，VALOR表现更佳，并在两个基准测试上取得了新的最先进结果。

5.2.3 图像-语言基准

我们在三个图像-语言基准测试上评估 VALOR$_L$，包括文本转图像检索、图像字幕生成和视觉问答。如表7所示，VALOR$_L$ 在三个基准测试中均取得了不错的表现。具体而言，VALOR$_L$ 在 COCO检索基准测试上的表现与 FILIP 相当 [125]。在 COCO 字幕基准测试中，VALOR$_L$ 优于 GIT，且表现与 GIT2模型相当，同时所需参数和数据大幅减少。在 VQAv2基准测试中，VALOR$_L$ 大幅优于规模相近的 GIT$_L$ ，并取得了与规模更大的 GIT模型相当的表现。

5.3 消融研究

5.3.1 视觉-音频-语言跨模态学习

我们首先开展实验，以展示视觉-音频-语言跨模态学习的必要性。具体而言，我们在两种设置下（使用和不使用 VALOR 预训练），针对具有不同输入模态的 6 项任务中的 9 个基准测试训练模型。如图6 所示，与仅使用视觉或音频模态相比，同时利用两种模态在两种预训练设置下均能使 9 个基准测试取得稳定提升。这些结果证明，结合两种模态确实有助于模型普遍地理解视频。此外，视觉-音频-语言预训练还能进一步增强模型的三模态推理能力。与视觉-语言任务相比，音频模态对于视听-语言任务的作用更大。例如，引入音频模态可使音视频检索（VALOR）和视频检索（MSRVTT）的性能分别提升 26.0% 和 14.9%。这是因为视听-语言任务与音频直接相关，而视觉-语言任务则将音频用作辅助模态。

5.3.2 模态分组策略

如第4.2节所述，在预训练期间，我们在MGA和MGC任务中均采用模态分组策略，旨在增强模型对输入模态不同的任务的泛化能力。为验证其有效性，我们使用不同的模态组对模型进行预训练，并在多个基准测试上进行评估。以MGA为例，从表8 所示结果中，我们可以得出以下结论。I）分别使用T-V（M1）或T-A组（M2）进行MGA预训练，在相应的T-V或T-A基准测试上能取得较好结果，但在T-AV基准测试上的性能较低。相比之下，使用T-AV组（M3）进行预训练，在T-AV基准测试上能取得更好的性能，但由于三模态预训练与双模态适配之间存在不一致，其在T-V和T-A基准测试上的性能分别弱于M1和M2。II）在M3的基础上，额外引入T-V组或T-A组有助于提升相应基准测试上的性能，但会降低另一项的性能。例如，M4在T-V基准测试上的性能优于M3，且与M1相当，但其在T-A基准测试上的性能相比M3下降更多。III）使用T-AV+T-V+T-A组（M6）训练的模型，在零样本和微调设置下，均能在T-AV、T-V和T-A基准测试上取得不错的性能。IV）在M6的基础上，进一步引入V-A、A-TV和V-TA组可以提升相应基准测试上的性能，但也会导致主流T-AV、T-V和T-A基准测试上的性能明显下降。为

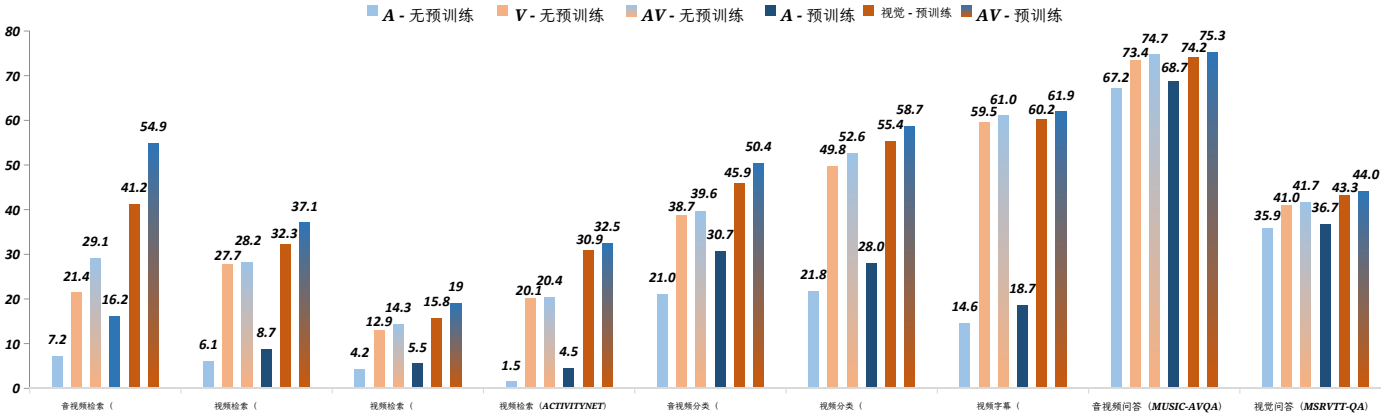

图6：在6项任务的9个基准测试中，使用音频（A）、视觉（V）或两种模态（AV）的实验结果。是否预训练表示模型是否在 VALOR-1M 上进行过预训练。检索、字幕生成和问答任务分别报告 R@1、CIDEr 和准确率指标。

表8：使用不同模态组在VALOR-1M上预训练的模型在MGA任务中的下游性能。仅使用 L_{MGA}。报告零样本R@1 / 微调R@1。

Name	预训练模态组						T-V		T-A		T-AV	V-A	A-TV	V-TA
	T-V	T-A	T-AV	V-A	A-TV	V-TA	VALOR-32K	MSVD	VALOR-32K	ClothoV1	VALOR-32K			
M1	✓						**38.3/42.7**	**30.7**/36.0	0.0/7.2	0.2/9.0	38.3/46.9	-	-	-
M2		✓					0.0/18.1	0.2/14.7	17.1/16.7	8.4/16.5	14.5/30.8	-	-	-
M3			✓				34.9/40.1	28.7/34.8	10.2/12.3	8.1/14.7	50.5/55.4	-	-	-
M4	✓		✓				36.9/42.5	30.4/**36.2**	8.8/11.1	7.2/14.4	50.5/55.5	-	-	-
M5		✓	✓				33.9/39.1	27.6/33.6	**17.1/17.9**	**9.6/16.8**	52.0/54.8	-	-	-
M6	✓	✓	✓				37.1/41.8	29.8/35.4	16.9/17.0	9.2/16.6	**50.7/55.6**	5.6/10.2	18.5/20.5	39.8/45.6
M7	✓	✓	✓	✓	✓	✓	34.5/40.7	28.5/34.3	16.9/17.6	8.2/15.9	48.2/54.0	**11.7/18.5**	**24.6/27.9**	**41.2/52.9**

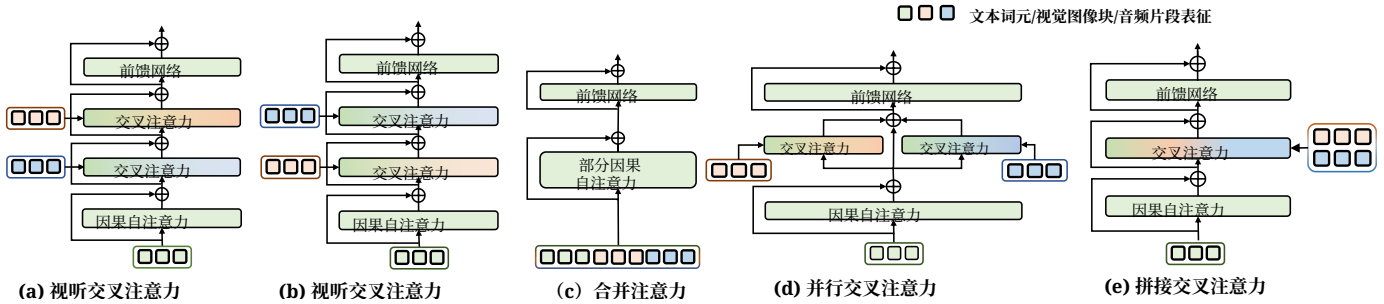

图7：多模态解码器中用于MGC任务的不同注意力机制变体示例

用于 MGC 任务的多模态解码器中采用不同注意力机制的 f 个变体 .

表9：在MGC任务中，使用不同模态组在VALOR-1M上预训练的模型的下游性能。仅使用 L_{MGC}。报告零样本CIDEr / 微调CIDEr。

模态组			AVC（VALOR-32K）	VC（MSVD）	AC（Clotho）
T-V	T-A	T-AV			
✓			31.5/47.5	15.4/**122.6**	1.6/28.2
	✓		20.7/39.2	0.8/95.0	7.6/39.2
		✓	**43.1/50.0**	12.5/119.9	10.5/38.1
✓		✓	41.6/49.9	14.7/121.7	6.6/36.2
	✓	✓	42.7/49.6	11.5/116.7	**8.1**/40.1
✓	✓	✓	41.4/49.8	**15.6**/122.2	8.0/**40.4**

为此，我们选择 M6 作为默认设置。除了 MGA，模态分组策略也适用于 MGC 任务，并且可以从表 9.中观察到类似结论。

表 10：针对 MGA 任务，采用不同视听融合方法在 VALOR-1M 上预训练的模型的下游性能。模型仅使用 $L_{MGA(T-AV)}$ 在 VALOR-1M 上进行预训练。MSR 是 MSRVTT 数据集的简称。粗体行表示默认设置。

方法	AVR (VALOR-32K)		视频检索 (MSR)	
	R@1	R@5	R@1	R@5
粗粒度 + 分数融合	53.4	81.3	35.0	66.0
粗粒度 + 特征融合	54.9	81.5	34.9	66.0
细粒度 +分数融合	54.5	82.3	35.7	**67.0**
细粒度 +特征融合	54.9	81.7	36.7	65.9
细粒度 +特征融合 +加权平均	**55.4**	**82.7**	**36.8**	66.5

5.3.3 视听融合

MGA 中的视听融合（T-AV）。MGA任务中的T-AV组旨在实现

表 11：针对 MGC 任务，采用不同 E 和不同视听融合方法在 VALOR-1M 上预训练的模型的下游性能。模型仅使用 $L_{MGC(T-AV)}$ 在 VALOR-1M 上进行预训练。图像描述任务报告 CIDEr指标。粗体行表示默认设置。

方法	AVC（VALOR-32K）	视频字幕（MSR）	视觉问答（MSR）
合并注意力	48.2	56.0	43.4
视听交叉注意力	49.6	57.0	43.7
视听交叉注意力	49.6	57.4	43.6
并行交叉注意力	49.5	57.4	43.8
拼接交叉注意力	**50.0**	**59.0**	**44.1**

表12：在不同 MGA 和 MGC 任务组合设置下，基于 VALOR-1M 预训练的模型的下游性能。检索和字幕生成分别报告 R@1 和 CIDEr。

共享权重	α	AVR (VALOR-32K)	AVC (VALOR-32K)	VQA (MSR)
✗	1	53.8	50.3	44.1
✓	1	54.7	50.0	44.1
✓	0.5	53.0	**50.8**	**44.1**
✓	1.5	54.9	50.4	44.0
✓	3.0	**55.4**	49.9	43.8

语言与音频和视觉的融合之间的细粒度对齐。我们将其与粗粒度对齐方式进行比较，后者将整句的全局表征与整个视频和音频的全局表征融合结果进行对齐。此外，我们还比较了两种视听融合方法，包括特征融合和分数融合。特征融合会先融合音频和视觉特征（粗粒度对齐时沿隐藏维度拼接，细粒度对齐时沿序列维度拼接），再计算其与文本的相似度；而分数融合则分别计算文本-视频和文本-音频的相似度分数，再将二者相加作为总分。从表 10所示结果可以看出，在四种组合中，细粒度对齐结合特征融合在 VALOR-32K 和 MSRVTT 数据集上均取得了最佳结果。采用第4.2节介绍的加权平均方法，还能在两个基准测试上持续提升性能。

MGC（T-AV）中的视听融合。带有 T-AV 组的 MGC 任务要求模型在视觉和音频特征的共同条件下预测掩码词元。我们对多模态解码器中采用不同注意力机制的五种变体进行了比较，如图7所示。具体而言，视听交叉注意力（a）和视觉-音频交叉注意力（b）变体引入了两个交叉注意力层，并依次关注两种模态。合并注意力（c）直接拼接三模态特征，并在自注意力层中使用部分因果掩码，使视觉和音频能够充分关注彼此，同时防止文本信息泄漏。拼接式交叉注意力（e）引入一个交叉注意力层，并关注两种模态的拼接结果。并行交叉注意力（d）为两种模态使用独立参数，而非像（e）那样共享权重。从表11所示结果可知，拼接式交叉注意力机制在 AVC、VC 和 VQA 任务上始终取得最佳效果。

MGA 和 MGC 的组合。当模型与 MGA 和 MGC 任务一同训练时，我们共享文本编码器和解码器的公共参数，并使用一个超参数 α 来平衡两种损失。如表 12所示，与表 8 和表 9中分别训练的结果相比，同时训练这两项任务会导致 AVR 性能下降（53.8 对 55.6），但 AVC 任务性能有所提升（50.3 对 49.6）。参数共享可使 AVR 任务的性能提高 0.9 个点，而 AVC 任务的性能仅略有下降。使用更大的 α 会使模型更侧重于 MGA 任务，从而提升 AVR 性能，但 AVC 性能会降低。视频问答对不同设置的敏感度相对较低。最终，我们选择参数共享并设置$\alpha = 1.5$，因为其在多个任务上表现良好。

5.3.4 VALOR-1M 数据集的影响

在表14中，我们将 VALOR-1M 与其他公开视频-语言预训练数据集进行比较，包括 WebVid-2.5M、CC3M 和 HD VILA 10M。在 VALOR-1M 上预训练的模型使用视觉和音频模态，而在其他数据集上预训练的模型仅使用视觉模态。所有模型均在包含视觉和音频模态的 MSRVTT 数据集上进行微调。从结果可以看出，在 VALOR-1M 上预训练的模型在全部三个基准测试中都以明显优势超越了其他模型，这得益于 VALOR-1M 中高质量的视听字幕。

5.3.5 模型架构选择

我们在表13中比较了不同的模型架构。结果显示，当 BERT_B 用作文本编码器时，采用更强大的视觉编码器能够在全部四个检索基准上带来更大的提升。此外，我们还比较了以 CLIP_L 作为视觉编码器时，不同文本编码器的表现。如果使用 BERT_B 作为文本编码器，那么在 MGA 学习开始时，三个单模态编码器尚未对齐。相比之下，如果使用 CLIP_L 作为文本编码器，那么在 MGA 学习之前，视觉与文本就已完成对齐。结果显示，当预训练数据仅限于 VALOR-1M 时，CLIP_L 在 3 个视频检索基准上明显优于 BERT_B，这得益于大规模 CLIP 对比预训练。然而，CLIP_L 在 AVR 基准上的表现不如 BERT_B。我们推测，预先对齐视觉与语言会使模型倾向于忽略视听-语言相关性的学习，因为在预训练开始时，仅利用视觉信息所造成的损失较小。当使用更多训练数据（33.5M）时，BERT_B 在视频检索基准上的劣势大幅减轻，同时其在 AVR 基准上的优势仍然存在。此外，我们还尝试将文本编码器从 BERT_B 扩展到 BERT_L，发现除 MSRVTT 外，其余两个基准上的表现均有所提升。最终，考虑到效率和效果，我们选择 CLIP_L 和 BERT_B作为 VALOR_L的默认设计。

5.4 可视化

在图8中，我们对 VALOR-32K 基准测试进行了定性比较，比较对象包括 VALOR_B 以及 AVLNet [50] 和 SMPFF [46], 等任务专用方法，这些方法都以视觉和音频为输入。从可视化结果中

表13：采用不同架构的模型的下游性能。预训练和微调均使用视觉和音频。所有模型均使用 AST_B 作为音频编码器。

视觉编码器	文本编码器	#样本	AVR（VALOR-32K）	视频检索（MSRTT）	视频检索 (DiDeMo)	视频检索 (LSMDC)
Video $Swin_B$	$BERT_B$	1M	56.8/83.1/89.9	39.3/69.0/80.4	41.1/73.1/81.8	19.8/41.3/52.5
$CLIP_B$	$BERT_B$	1M	61.8/86.0/92.3	45.0/74.1/84.4	46.3/75.6/84.2	22.7/45.4/56.1
$CLIP_L$	$BERT_B$	1M	64.3/87.3/92.8	47.7/76.1/84.6	47.3/77.9/85.1	27.8/51.4/60.9
$CLIP_L$	$CLIP_L$	1M	59.3/84.4/91.0	50.6/79.0/87.8	53.4/80.2/88.3	29.2/51.6/58.8
$CLIP_L$	$BERT_L$	1M	66.5/88.1/93.4	46.1/76.7/85.4	49.0/77.4/84.9	29.2/50.7/60.5
$CLIP_L$	$BERT_B$	33.5M	**73.2**/**91.6**/**95.4**	54.4/79.8/87.6	**57.6**/**83.3**/**88.8**	31.8/**52.8**/62.4
$CLIP_L$	$CLIP_L$	33.5M	67.8/89.5/94.0	**55.3**/**80.5**/**88.1**	57.1/82.9/88.6	**32.6**/52.6/**62.7**



图8：不同模型在 VALOR-32K 基准上的预测结果可视化。点击按钮播放音频。

表14：在不同数据集上预训练的模型的下游性能。为公平比较，在每个数据集上使用相同的迭代次数训练模型。

数据集	视频检索 (MSRVTT)	视频字幕 (MSRVTT)	视觉问答 (MSRVTT)
HD VILA 10M	30.4	55.4	43.2
WebVid-2.5M	35.7	58.8	43.8
CC3M	36.3	58.8	44.0
VALOR-1M	**39.5**	**60.7**	**44.5**

我们可以发现，与 AVLNet 相比，给定视听文本查询时，VALOR 能够准确地为视频候选项排序，并成功检索到真实标注对应的视频。与排名第一的视频相比，排名第二的视频也展示了一名身穿白衣的男子在水上，并伴有风声，但缺少鸟鸣声。无论从视觉还是音频角度来看，排名第三的视频都与查询更加不符。对于 AVC 任务，SMPFF 要么错误地识别声音（样本 1），要么描述视觉识别错误（如样例 2），或完全忽略音频信息（如样例 3）。相比之下，VALOR 能全面识别视觉和音频概念，并为全部三个样例生成准确的描述。

6 结论

本文提出了一种统一的视觉-音频-语言跨模态预训练模型 VALOR，通过两项专门设计的预训练任务——多模态分组对齐和多模态分组描述，实现三模态理解与生成。大量实验表明，VALOR 具有良好的通用性和可扩展性。本文还提出了首个强相关的视觉-音频-语言数据集 VALOR-1M，以推动三模态预训练研究，并提出了用于视听-语言检索和字幕生成基准测试的 VALOR-32K。在 VALOR-1M 和其他公开视觉-语言数据集上训练后，VALOR 在

下游视觉、音频和视听检索、字幕生成及问答任务中取得了一系列新的最先进性能。未来，我们计划通过生成并筛选伪视听描述等无监督方法，扩大 VALOR-1M 数据集的规模。此外，我们还计划在当前的 VALOR 框架中加入视觉和音频生成建模。

参考文献

[1] J.-H. Kim、J. Jun 和 B.-T. Zhang，“双线性注意力网络”，神经信息处理系统进展，第31卷，2018年。[2] J. Dong、X. Li、C. Xu、X. Yang、G. Yang、X. Wang 和 M. Wang，“通过文本进行视频检索的双重编码”，*IEEE* 模式分析与机器智能汇刊，2021年。[3] F. Liu、X. Wu、C. You、S. Ge、Y. Zou 和 X. Sun，“对齐源视觉域与目标语言域以实现无配对视频字幕生成”，*IEEE* 模式分析与机器智能汇刊，2021年。[4] Y. Yan、N. Zhuang、B. Ni、J. Zhang、M. Xu、Q. Zhang、Z. Zhang、S. Cheng、Q. Tian、Y. Xu 等，“通过基于图的多粒度交互学习实现细粒度视频字幕生成”，*IEEE* 模式分析与机器智能汇刊，第44卷第2期，第666–683页，2019年。[5] B. Wang、L. Ma、W. Zhang、W. Jiang、J. Wang 和 W. Liu，“基于门控融合网络和词性序列引导的可控视频字幕生成”，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第2641–2650页。[6] P. Anderson、X. He、C. Buehler、D. Teney、M. Johnson、S. Gould 和 L. Zhang，“用于图像字幕生成和视觉问答的自底向上与自顶向下注意力”，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第6077–6086页。[7] L. Peng、Y. Yang、Z. Wang、Z. Huang 和 H. T. Shen，“MRA-Net：通过多模态关系注意力网络改进视觉问答”，*IEEE* 模式分析与机器智能汇刊，第44卷第1期，第318–329页，2020年。[8] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，*arXiv*预印本 *arXiv:1810.04805*，2018年。[9] A. Radford、K. Narasimhan、T. Salimans、I. Sutskever 等，“通过生成式预训练提升语言理解能力”，2018年。[10] T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal、A. Neelakantan、P. Shyam、G. Sastry、A. Askell 等，“语言模型是少样本学习器”，神经信息处理系统进展，第33卷，第1877–1901页，2020年。[11] A. Miech、D. Zhukov、J.-B. Alayrac、M. Tapaswi、I. Laptev 和 J. Sivic，“HowTo100M：通过观看一亿个配音视频片段学习文本-视频嵌入”，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第2630–2640页。[12] H. Xue、T. Hang、Y. Zeng、Y. Sun、B. Liu、H. Yang、J. Fu 和 B. Guo，“利用大规模视频转录文本提升高分辨率视频-语言表征”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第5036–5045页。[13] M. Bain、A. Nagrani、G. Varol 和 A. Zisserman，“Frozen in Time：用于端到端检索的视频与图像联合编码器”，载于*IEEE/CVF*国际计算机视觉会议论文集，2021年，第1728–1738页。[14] J. Wang、Z. Yang、X. Hu、L. Li、K. Lin、Z. Gan、Z. Liu、C. Liu 和 L. Wang，“GIT：用于视觉与语言的生成式图像到文本变换器”，*arXiv*预印本 *arXiv:2205.14100*，2022年。[15] R. Zellers、J. Lu、X. Lu、Y. Yu、Y. Zhao、M. Salehi、A. Kusupati、J. Hessel、A. Farhadi 和 Y. Choi，“MERLOT Reserve：通过视觉、语言和声音获取神经脚本知识”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第16 375–16 387页。[16] A. Miech、J.-B. Alayrac、L. Smaira、I. Laptev、J. Sivic 和 A. Zisserman，“从未经筛选的教学视频中端到端学习视觉表征”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第9879–9889页。

[17] P. Sharma、N. Ding、S. Goodman 和 R. Soricut，《Conceptual captions：一个经过清洗并添加上位词的图像替代文本数据集，用于自动图像描述》，载于《计算语言学协会第56届年会论文集》（第1卷：长篇论文），2018年，第2556–2565页。[18] S. Changpinyo、P. Sharma、N. Ding 和 R. Soricut，《Conceptual 12M：推动网络规模图文预训练，以识别长尾视觉概念》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2021年，第3558–3568页。[19] P. H. Seo、A. Nagrani、A. Arnab 和 C. Schmid，《面向多模态视频描述的端到端生成式预训练》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第17 959–17 968页。[20] H. Xue、Y. Sun、B. Liu、J. Fu、R. Song、H. Li 和 J. Luo，《Clip-vip：将预训练图文模型适配到视频语言表征对齐》，arXiv预印本 arXiv:2209.06430，2022年。[21] 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《欧洲计算机视觉会议论文集》。Springer，2020年，第104–120页。[22] J. Lu、D. Batra、D. Parikh 和 S. Lee，《VilBERT：为视觉与语言任务预训练任务无关的视觉语言表征》，《神经信息处理系统进展》，第32卷，2019年。[23] T.-J. Fu、L. Li、Z. Gan、K. Lin、W. Y. Wang、L. Wang 和 Z. Liu，《采用掩码视觉建模的端到端视频语言Transformer实证研究》，arXiv预印本 arXiv:2209 01540，2022年。。[24] L. Li、Y.-C. Chen、Y. Cheng、Z. Gan、L. Yu 和 J. Liu，《HERO：用于视频+语言全方位表征预训练的层次化编码器》，载于《2020年自然语言处理实证方法会议论文集》（EMNLP），2020年，第2046–2065页。[25] J. Li、R. Selvaraju、A. Gotmare、S. Joty、C. Xiong 和 S. C. H. Hoi，《先对齐后融合：采用动量蒸馏进行视觉与语言表征学习》，《神经信息处理系统进展》，第34卷，第9694–9705页，2021年。[26] S. Ren、K. He、R. Girshick 和 J. Sun，《Faster R-CNN：借助区域提议网络实现实时目标检测》，《神经信息处理系统进展》，第28卷，2015年。[27] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等，《一张图像胜过16×16个词：用于大规模图像识别的Transformer》，arXiv预印本 arXiv:2010.11929，2020年。[28] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo，《Swin Transformer：采用移位窗口的层次化视觉Transformer》，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第10 012–10 022页。[29] J. Cho、J. Lei、H. Tan 和 M. Bansal，《通过文本生成统一视觉与语言任务》，载于《国际机器学习会议论文集》。PMLR，2021年，第1931–1942页。[30] Z. Yang、Z. Gan、J. Wang、X. Hu、F. Ahmed、Z. Liu、Y. Lu 和 L. Wang，《UniTAB：统一基于视觉定位的视觉语言建模中的文本和框输出》，载于《欧洲计算机视觉会议论文集》。Springer，2022年，第521–539页。[31] P. Wang、A. Yang、R. Men、J. Lin、S. Bai、Z. Li、J. Ma、C. Zhou、J. Zhou 和 H. Yang，《OFA：通过简单的序列到序列学习框架统一架构、任务和模态》，载于《国际机器学习会议论文集》。PMLR，2022年，第23 318–23 340页。[32] J. Lu、C. Clark、R. Zellers、R. Mottaghi 和 A. Kembhavi，《Unified-IO：用于视觉、语言和多模态任务的统一模型》，arXiv预印本 arXiv:2206.08916，2022年。[33] T. Chen、S. Saxena、L. Li、D. J. Fleet 和 G. rey E. Hinton，《Pix2Seq：用于目标检测的语言建模框架》，ArXiv，第abs/2109.10852卷，2021年。[34] A. Van Den Oord、O. Vinyals 等，《神经离散表征学习》，《神经信息处理系统进展》，第30卷，2017年。[35] X. Zhu、J. Zhu、H. Li、X. Wu、H. Li、X. Wang 和 J. Dai，《Uni-perceiver：面向通用感知的统一架构预训练，用于零样本和少样本任务》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第16 804–16 815页。[36] L. Li、Z. Gan、K. Lin、C.-C. Lin、Z. Liu、C. Liu 和 L. Wang,

《LAVENDER：将视频-语言理解统一为掩码语言建模》，arXiv预印本 arXiv:2206.07160，2022 年。[37] 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视觉模型》，载于《国际机器学习会议论文集》。PMLR，2021年，第8748–8763页。[38] C. Jia、Y. Yang、Y. Xia、Y.-T. Chen、Z. Parekh、H. Pham、Q. Le、Y.-H. Sung、Z. Li 和 T. Duerig，《通过噪声文本监督扩展视觉与视觉-语言表征学习》，载于《国际机器学习会议论文集》。PMLR，2021年，第4904–4916页。[39] L. Yuan、D. Chen、Y.-L. Chen、N. Codella、X. Dai、J. Gao、H. Hu、X. Huang、B. Li、C. Li 等，《Florence：一种面向计算机视觉的新型基础模型》，arXiv预印本 arXiv:2111.11432，2021年。[40] H. Pham、Z. Dai、G. Ghiasi、K. Kawaguchi、H. Liu、A. W. Yu、J. Yu、Y.-T. Chen、M.-T. Luong、Y. Wu 等，《开放词汇图像分类的联合扩展》，arXiv预印本 arXiv: 2111.10050，2021 年。[41] Z. Wang 、 J. Yu 、 A. W. Yu 、 Z. Dai、Y. Tsvetkov 和 Y. Cao，《SimVLM：采用弱监督的简单视觉语言模型预训练》，arXiv预印本 arXiv:2108.10904，2021年。[42] J.-B. Alayrac、J. Donahue、P. Luc、A. Miech、I. Barr、Y. Hasson、K. Lenc、A. Mensch、K. Millican、M. Reynolds 等，《Flamingo：用于少样本学习的视觉语言模型》，arXiv预印本 arXiv:2204.14198，2022年。[43] X. Chen、X. Wang、S. Changpinyo、A. Piergiovanni、P. Padlewski、D. Salz、S. Goodman、A. Grycner、B. Mustafa、L. Beyer 等，《PALI：联合扩展的多语言图像-语言模型》，arXiv预印本 arXiv:2209.06794，2022年。[44] J. Yu、Z. Wang、V. Vasudevan、L. Yeung、M. Seyedhosseini 和 Y. Wu，《CoCa：对比式图像描述模型是图像-文本基础模型》，arXiv预印本 arXiv:2205.01917，2022年。[45] V. Gabeur、C. Sun、K. Alahari 和 C. Schmid，《用于视频检索的多模态Transformer》，载于《欧洲计算机视觉会议论文集》。Springer，2020年，第214–229页。[46] S. Chen、X. Zhu、D. Hao、W. Liu、J. Liu、Z. Zhao、L. Guo 和 J. Liu，《MM21视频理解挑战赛预训练：采用预训练技术进行视频描述生成》，载于《第29届 ACM国际多媒体会议论文集》，2021年，第4853–4857页。[47] H. Luo、L. Ji、B. Shi、H. Huang、N. Duan、T. Li、J. Li、T. Bharti 和 M. Zhou，《UniVL：用于多模态理解与生成的统一视频与语言预训练模型》，arXiv 预印本 arXiv:2002.06353，2020年。[48] H. Xu、G. Ghosh、P.-Y. Huang、P. Arora、M. Aminzadeh、C. Feichtenhofer、F. Metze 和 L. Zettlemoyer，《VLM：面向视频理解的任务无关视频-语言模型预训练》，arXiv预印本 arXiv:2105.09996，2021年。[49] L. Li、J. Lei、Z. Gan、L. Yu、Y.-C. Chen、R. Pillai、Y. Cheng、L. Zhou、X. E. Wang、W. Y. Wang 等，《VALUE：用于视频与语言理解评估的多任务基准》，arXiv预印本 arXiv:2106.04632，2021年。[50] A. Rouditchenko、A. Boggust、D. Harwath、B. Chen、D. Joshi、S. Thomas、K. Audhkhasi、H. Kuehne、R. Panda、R. Feris 等，《AVLNet：从教学视频中学习音频-视觉语言表征》，arXiv预印本 arXiv:2006.09199，2020年。[51] B. Chen、A. Rouditchenko、K. Duarte、H. Kuehne、S. Thomas、A. Boggust、R. Panda、B. Kingsbury、R. Feris、D. Harwath 等，《用于从未标注视频进行自监督学习的多模态聚类网络》，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第8012–8021页。[52] H. Akbari、L. Yuan、R. Qian、W.-H. Chuang、S.-F. Chang、Y. Cui 和 B. Gong，《VATT：从原始视频、音频和文本进行多模态自监督学习的Transformer》，《神经信息处理系统进展》，第34卷，第24 206–24 221页，2021年。[53] Z. Yang、Y. Fang、C. Zhu、R. Pryzant、D. Chen、Y. Shi、Y. Xu、Y. Qian、M. Gao、Y.-L. Chen 等，《i-Code：集成式可组合多模态学习框架》，arXiv预印本 arXiv:2205.01818，2022年。[54] D. Chen 和 W. B. Dolan，《为释义评估收集高度平行的数据》，载于《计算语言学协会第49届年会：人类语言技术论文集》，2011年，第190–200页。[55] J. Xu、T. Mei、T. Yao 和 Y. Rui，《MSR-VTT：用于衔接视频与语言的大型视频描述数据集》，载于《IEEE 计算机视觉与模式识别会议论文集》，2016年，第5288–5296页。

[56] X. Wang、J. Wu、J. Chen、L. Li、Y.-F. Wang 和 W. Y. Wang，“Vatex：一个面向视频与语言研究的大规模高质量多语言数据集”，载于《IEEE/CVF国际计算机视觉会议论文集》，2019年，第 4581–4591 页。[57] L. Zhou、C. Xu 和 J. J. Corso，“从网络教学视频中自动学习操作流程”，载于第三十二届 AAAI人工智能会议，2018年。[58] L. Anne Hendricks、O. Wang、E. Shechtman、J. Sivic、T. Darrell 和 B. Russell，“使用自然语言定位视频中的时刻”，载于《IEEE国际计算机视觉会议论文集》，2017年，第 5803–5812 页。[59] R. Krishna、K. Hata、F. Ren、L. Fei-Fei 和 J. Carlos Niebles，“视频中的事件密集字幕生成”，载于《IEEE国际计算机视觉会议论文集》，2017年，第 706–715 页。[60] A. Rohrbach、A. Torabi、M. Rohrbach、N. Tandon、C. Pal、H. Larochelle、A. Courville 和 B. Schiele，“电影描述”，《国际计算机视觉期刊》，第 123 卷，120，2017年 1，94第。–。[61] K. Drossos、S. Lipping 和 T. Virtanen，“Clotho：一个音频描述数据集”，载于 ICASSP 2020-2020 IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2020年，第 736–740 页。[62] C. D. Kim、B. Kim、H. Lee 和 G. Kim，“AudioCaps：为真实场景中的音频生成字幕”，载于《2019年计算语言学协会北美分会会议：人类语言技术论文集》，第 1 卷（长短论文），2019年，第 119–132 页。[63] H. Yun、Y. Yu、W. Yang、K. Lee 和 G. Kim，“Pano-AVQA：360度视频中的有依据视听问答”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第 2031–2041 页。[64] G. Li、Y. Wei、Y. Tian、C. Xu、J.-R. Wen 和 D. Hu，“在动态视听场景中学习回答问题”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第 19 108–19 118 页。[65] P. Yang、X. Wang、X. Duan、H. Chen、R. Hou、C. Jin 和 W. Zhu，“AVQA：一个用于视频视听问答的数据集”，载于《第30届 ACM国际多媒体会议论文集》，2022年，第 3480–3491 页。[66] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，“AudioSet：一个用于音频事件的本体与人工标注数据集”，载于《2017年 IEEE国际声学、语音与信号处理会议（ICASSP）》。IEEE，2017年，第 776–780 页。[67] Z. Liu、J. Ning、Y. Cao、Y. Wei、Z. Zhang、S. Lin 和 H. Hu，“视频Swin Transformer”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第 3202–3211 页。[68] Y. Gong、Y.-A. Chung 和 J. Glass，“AST：音频频谱变换器”，arXiv 预印本 arXiv:2104.01778，2021年。[69] Y. Gong、C.-I. Lai、Y.-A. Chung 和 J. Glass，“SSAST：自监督音频频谱变换器”，载于《AAAI人工智能会议论文集》，第 36 卷，第 10 期，2022年，第 10 699–10 709 页。[70] X. Cheng、H. Lin、X. Wu、F. Yang 和 D. Shen，“通过多流语料对齐和双重Softmax损失改进视频文本检索”，arXiv 预印本 arXiv:2109.04290，2021年。[71] J. Lei、L. Li、L. Zhou、Z. Gan、T. L. Berg、M. Bansal 和 J. Liu，“少即是多：通过稀疏采样实现视频与语言学习的 ClipBERT”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2021年，第 7331–7341 页。[72] Y. Ge、Y. Ge、X. Liu、D. Li、Y. Shan、X. Qie 和 P. Luo，“通过选择题连接视频文本检索”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第 16 167–16 176 页。[73] Y. Ge、Y. Ge、X. Liu、A. J. Wang、J. Wu、Y. Shan、X. Qie 和 P. Luo，“MILES：注入语言语义的视觉BERT预训练，用于视频文本检索”，arXiv 预印本 arXiv:2204.12408，2022年。[74] J. Wang、Y. Ge、G. Cai、R. Yan、X. Lin、Y. Shan、X. Qie 和 M. Z. Shou，“面向检索的对象感知视频语言预训练”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2022年，第 3313–3322 页。[75] A. Nagrani、P. H. Seo、B. Seybold、A. Hauth、S. Manen、C. Sun 和 C. Schmid，“从图像描述中学习音视频模态”，arXiv 预印本 arXiv:2204.00679，2022年。[76] Y. Sun、H. Xue、R. Song、B. Liu、H. Yang 和 J. Fu，“通过多模态时间对比学习进行长视频语言预训练”，arXiv 预印本 arXiv:2210.06031，2022年。

[77] J. Lei、T. L. Berg 和 M. Bansal，“揭示视频与语言学习中的单帧偏差”，arXiv 预印本 arXiv:2206 03428，2022年。[78] J. Yang、Y. Bisk 和 J. Gao，“TACo：面向视频文本对齐的令牌感知级联对比学习”，载于《IEEE/CVF国际计算机视觉会议论文集》，2021年，第 11 562–11 572 页。[79] M. Patrick、P.-Y. Huang、Y. Asano、F. Metze、A. Hauptmann、J. Henriques 和 A. Vedaldi，“视频文本表示学习的支持集瓶颈”，arXiv 预印本 arXiv:2010 02824，2020年。[80] V. Gabeur、A. Nagrani、C. Sun、K. Alahari 和 C. Schmid，“为跨模态视频检索屏蔽模态”，载于《IEEE/CVF 冬季计算机视觉应用会议论文集》，2022年，第 1766–1775 页。[81] A. J. Wang、Y. Ge、R. Yan、Y. Ge、X. Lin、G. Cai、J. Wu、Y. Shan、X. Qie 和 M. Z. Shou，“All in One：探索统一的视频语言预训练”，arXiv 预印本 arXiv:2203.07303，2022年。[82] T.-J. Fu、L. Li、Z. Gan、K. Lin、W. Y. Wang、L. Wang 和 Z. Liu，“VIOLET：采用掩码视觉令牌建模的端到端视频语言变换器”，arXiv 预印本 arXiv:2111.12681，2021年。[83] H. Luo、L. Ji、M. Zhong、Y. Chen、W. Lei、N. Duan 和 T. Li，“CLIP4Clip：针对端到端视频片段检索与字幕生成的 CLIP 实证研究”，《神经计算》，第 508 卷，304，2022年 293页。–。[84] Y. Liu、P. Xiong、L. Xu、S. Cao 和 Q. Jin，“TS2-Net：用于文本视频检索的令牌移位与选择变换器”，载于《欧洲计算机视觉会议》。Springer，2022年，第 319–335 页。[85] Y. Ma、G. Xu、X. Sun、M. Yan、J. Zhang 和 R. Ji，“X-clip：用于视频文本检索的端到端多粒度对比学习”，载于《第30届 ACM国际多媒体会议论文集》，2022年，第 638–647 页。[86] Y.-B. Lin、J. Lei、M. Bansal 和 G. Bertasius，“ECLIPSE：利用视觉与声音进行高效长程视频检索”，arXiv 预印本 arXiv:2204.02874，2022年。[87] Q. Wang、Y. Zhang、Y. Zheng、P. Pan 和 X.-S. Hua，“用于文本视频检索的解耦表示学习”，arXiv 预印本 arXiv:2203.07111，2022年。[88] S. Min、W. Kong、R.-C. Tu、D. Gong、C. Cai、W. Zhao、C. Liu、S. Zheng、H. Wang、Z. Li 等，“用于文本视频检索的混元 TVR”，arXiv 预印本 arXiv:2204.03382，2022年。[89] Y. Wang、K. Li、Y. Li、Y. He、B. Huang、Z. Zhao、H. Zhang、J. Xu、Y. Liu、Z. Wang 等，“InternVideo：通过生成式与判别式学习构建通用视频基础模型”，arXiv 预印本 arXiv:2212.03191，2022年。[90] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar 和 C. L. Zitnick，“微软 COCO：情境中的常见物体”，载于《欧洲计算机视觉会议》。Springer，2014年，第 740–755 页。[91] R. Krishna、Y. Zhu、O. Groth、J. Johnson、K. Hata、J. Kravitz、S. Chen、Y. Kalantidis、L.-J. Li、D. A. Shamma 等，“Visual Genome：利用众包密集图像标注连接语言与视觉”，《国际计算机视觉期刊》，第 123 卷，第 1 期，第 32–73 页，2017年。[92] V. Ordonez、G. Kulkarni 和 T. Berg，“Im2text：使用一百万张带说明文字的照片描述图像”，《神经信息处理系统进展》，第 24 卷，2011年。[93] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，“BLEU：一种机器翻译自动评估方法”，载于《计算语言学协会第40届年会论文集》，2002年，第 311–318 页。[94] S. Banerjee 和 A. Lavie，“METEOR：一种改进与人工判断相关性的机器翻译评估自动指标”，载于 ACL 研讨会论文集：机器翻译和/或摘要生成的内在与外在评估指标，2005年，第 65–72 页。[95] S. Banerjee、A. Lavie 等，“一种改进与人工判断相关性的机器翻译评估自动指标”，载于 ACL-2005 研讨会论文集：机器翻译和 / 或摘要生成的内在与外在评估指标，2005年，第 65–72 页。[96] R. Vedantam、C. Lawrence Zitnick 和 D. Parikh，“CIDEr：基于共识的图像描述评估”，载于《IEEE计算机视觉与模式识别会议论文集》，2015年，第 4566–4575 页。[97] P. Anderson、B. Fernando、M. Johnson 和 S. Gould，“SPICE：语义命题式图像字幕评估”，载于《欧洲计算机视觉会议》。Springer，2016年，第 382–398 页。

[98] D. Xu、Z. Zhao、J. Xiao、F. Wu、H. Zhang、X. He 和 Y. Zhuang，《通过逐步精炼对外观与运动的注意力实现视频问答》，载于第*25*届*ACM*多媒体国际会议论文集，2017年，第1645–1653页。[99] Z. Yu、D. Xu、J. Yu、T. Yu、Z. Zhao、Y. Zhuang 和 D. Tao，《ActivityNet-QA：通过问答理解复杂网络视频的数据集》，载于*AAAI*人工智能会议论文集，第33卷，第01期，2019年，第9127–9134页。[100] Y. Jang、Y. Song、Y. Yu、Y. Kim 和 G. Kim，《Tgif-qa：迈向视觉问答中的时空推理》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第2758–2766页。[101] Y. Goyal、T. Khot、D. Summers-Stay、D. Batra 和 D. Parikh，《让VQA中的V发挥作用：提升图像理解在视觉问答中的作用》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第6904–6913页。[102] S. J. Rennie、E. Marcheret、Y. Mroueh、J. Ross 和 V. Goel，《用于图像字幕生成的自批判序列训练》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第7008–7024页。[103] Z. Zhang、Y. Shi、C. Yuan、B. Li、P. Wang、W. Hu 和 Z.-J. Zha，《通过教师推荐学习实现视频描述生成的对象关系图》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第13 278–13 288页。[104] Z. Zhang、Z. Qi、C. Yuan、Y. Shan、B. Li、Y. Deng 和 W. Hu，《采用检索-复制-生成网络的开放式视频描述生成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第9837–9846页。[105] K. Lin、L. Li、C.-C. Lin、F. Ahmed、Z. Gan、Z. Liu、Y. Lu 和 L. Wang，《Swinbert：采用稀疏注意力进行视频描述生成的端到端Transformer》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第17 949–17 958页。[106] J. Jiang、Z. Chen、H. Lin、X. Zhao 和 Y. Gao，《分而治之：用于视频问答的问题引导时空上下文注意力》，载于*AAAI*人工智能会议论文集，第34卷，第07期，2020年，第11 101–11 108页。[107] J. Huang、Y. Li、J. Feng、X. Sun 和 R. Ji，《Clover：迈向统一的视频-语言对齐与融合模型》，*arXiv*预印本 *arXiv:2207.07885*，2022年。[108] A. Yang、A. Miech、J. Sivic、I. Laptev 和 C. Schmid，《尽管发问：从数百万个带旁白的视频中学习回答问题》，载于*IEEE/CVF*国际计算机视觉会议论文集，2021年，第1686–1697页。[109] R. Zellers、X. Lu、J. Hessel、Y. Yu、J. S. Park、J. Cao、A. Farhadi 和 Y. Choi，《Merlot：多模态神经脚本知识模型》，神经信息处理系统进展，第34卷，第23 634–23 651页，2021年。[110] A. Yang、A. Miech、J. Sivic、I. Laptev 和 C. Schmid，《通过冻结的双向语言模型实现零样本视频问答》，*arXiv*预印本 *arXiv:2206.08155*，2022年。[111] S. Yan、T. Zhu、Z. Wang、Y. Cao、M. Zhang、S. Ghosh、Y. Wu 和 J. Yu，《利用对比式图像描述模型的零样本迁移进行视频文本建模》，*arXiv*预印本 *arXiv:2212.04979*，2022年。[112] A.-M. Oncescu、A. Koepke、J. F. Henriques、Z. Akata 和 S. Albanie，《通过自然语言查询进行音频检索》，*arXiv*预印本 *arXiv:2105.02192*，2021年。[113] X. Xu、H. Dinkel、M. Wu 和 K. Yu，《一种基于CRNN-GRU的音频描述强化学习方法》，载于*DCASE*，2020年，第225–229页。[114] K. Chen、Y. Wu、Z. Wang、X. Zhang、F. Nian、S. Li 和 X. Shao，《基于Transformer和预训练卷积神经网络的音频描述生成》，载于*DCASE*，2020年，第21–25页。[115] X. Xu、H. Dinkel、M. Wu、Z. Xie 和 K. Yu，《通过迁移学习研究自动音频描述生成中的局部与全局信息》，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第905–909页。[116] A. Koh、X. Fuzhao 和 C. E. Siong，《利用迁移学习和重构潜在空间相似性正则化实现自动音频描述生成》，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第7722–7726页。[117] X. Mei、X. Liu、Q. Huang、M. D. Plumbley 和 W. Wang，《音频描述生成Transformer》，*arXiv*预印本 *arXiv:2107.09817*，2021年。

[118] X. Liu、X. Mei、Q. Huang、J. Sun、J. Zhao、H. Liu、M. D. Plumbley、V. Kilic 和 W. Wang，《利用预训练BERT进行音频描述生成》，载于*2022*年第*30*届欧洲信号处理会议（*EUSIPCO*）。IEEE，2022年，第1145–1149页。[119] M. Ding、B. Xiao、N. Codella、P. Luo、J. Wang 和 L. Yuan，《DaViT：双注意力视觉变换器》，*arXiv*预印本 *arXiv:2204.03645*，2022年。[120] P. He、X. Liu、J. Gao 和 W. Chen，《DeBERTa：采用解耦注意力的解码增强型BERT》，*arXiv*预印本 *arXiv:2006 03654*，2020。。[121] 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-语言任务的对象语义对齐预训练》，载于欧洲计算机视觉会议。Springer，2020年，第121–137页。[122] J. Wang、X. Hu、Z. Gan、Z. Yang、X. Dai、Z. Liu、Y. Lu 和 L. Wang，《UFO：用于视觉-语言表征学习的统一变换器》，*arXiv*预印本 *arXiv:2111.10023*，2021年。[123] P. Zhang、X. Li、X. Hu、J. Yang、L. Zhang、L. Wang、Y. Choi 和 J. Gao，《VinVL：重新审视视觉-语言模型中的视觉表征》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第5579–5588页。[124] Z.-Y. Dou、Y. Xu、Z. Gan、J. Wang、S. Wang、L. Wang、C. Zhu、P. Zhang、L. Yuan、N. Peng 等，《端到端视觉语言变换器训练的实证研究》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第18 166–18 176页。[125] L. Yao、R. Huang、L. Hou、G. Lu、M. Niu、H. Xu、X. Liang、Z. Li、X. Jiang 和 C. Xu，《FILIP：细粒度交互式语言-图像预训练》，*arXiv*预印本 *arXiv:2111.07783*，2021年。[126] J. Li、D. Li、C. Xiong 和 S. Hoi，《BLIP：通过自举语言-图像预训练实现统一的视觉语言理解与生成》，*arXiv*预印本 *arXiv:2201.12086*，2022年。[127] X. Hu、Z. Gan、J. Wang、Z. Yang、Z. Liu、Y. Lu 和 L. Wang，《扩展视觉语言预训练以用于图像字幕生成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第17 980–17 989页。