arXiv:2603.01096v1 [cs.CV] 1 Mar 2026

通过概念空间对齐实现统一的视觉-语言建模

Yifu Qiu[1, 2, *], **Paul-Ambroise Duquenne**[2], **Holger Schwenk**[2]

[1]爱丁堡大学, [2]Meta FAIR
*本研究是在 Yifu Qiu 于 Meta FAIR 实习期间完成的

我们推出了 v-SONAR，这是一个从纯文本嵌入空间扩展而来的视觉-语言嵌入空间SONAR (Omnilingual Embeddings Team 等，2026），支持 1500 种文本语言和 177 种语音语言。为构建 v-SONAR，我们提出了一种事后对齐流程，将现有视觉编码器的表征映射到 SONAR 空间。我们对 v-SONAR进行了全面评估，结果表明，其嵌入在文本到视频检索任务上表现出色。搭配 OMNISONAR 文本解码器后，v-SONAR 在视频描述任务上的表现进一步超越了最先进的视觉-语言模型，包括 DREAM-1K(BLEU 23.9 对比 19.6）以及PE-VIDEO (BLEU39.0 对比 30.0）。

借助v-SONAR，我们首先证明，在LCM；LCM 团队等人，2024年）所述框架中运行、且仅使用英语文本训练的大型概念模型（SONAR 能够以零样本方式完成单视觉概念与多视觉概念理解。最后，我们提出v-LCM，它通过视觉-语言指令微调扩展了LCM 。v-LCM 通过v-SONAR 和SONAR将视觉与语言输入编码为统一的潜在嵌入序列，并采用与LCM纯文本预训练相同的潜在扩散目标进行下一嵌入预测训练。在大规模多语言、多模态指令微调数据混合集上的实验凸显了v-LCM的潜力：v-LCM 在涵盖图像/视频描述生成和问答的任务上与最先进的视觉语言模型表现相当，同时在全部62种测试语言中的61种从高资源到低资源的语言上显著优于这些模型。

日期：2026年3月3日
通讯作者：Holger Schwenk，邮箱：schwenk@meta.com

Meta

1 引言

语言和模态无关的嵌入空间已成为多语言、多模态表征学习的一种强大范式（Artetxe and Schwenk, 2019; Feng et al., 2020; Ni et al., 2021; Duquenne 等人, 2023; Wang 等人, 2024b; Chen et al., 2024a）。这类空间在广泛的应用中取得了最先进的性能，例如双语文本挖掘（Schwenk et al., 2019; NLLB 团队 et al., 2022; Ramesh et al., 2022）和语音-文本挖掘（Duquenne 等人, 2021b）。除此之外，采用编码器-解码器架构的嵌入空间（如 SONAR（Duquenne 等人, 2023））进一步支持了直接在潜在嵌入空间中进行生成式建模。大型概念模型（LCM；LCM 团队 et al. 2024）沿着这一方向更进一步，表明基于扩散的语言建模可以直接在语言无关的潜在空间中运行，即在连续嵌入而非离散词元上运行。尽管取得了这些进展，现有嵌入空间仍局限于文本和语音，限制了其在视觉-语言任务中的潜力。

在本研究中，我们提出了v-SONAR，将OMNISONAR (Omnilingual Embeddings Team 等人，2026)扩展至图像和视频模态。据我们所知，这使得OMNISONAR 成为覆盖四种模态的最通用嵌入空间[1]，支持多达 1500 种语言。我们采用教师-学生训练（Reimers和Gurevych，2020年；Duquenne 等人，2021a；Heffernan 等人，2022），以事后方式将最先进的视觉编码器感知编码器(Bolya 等人，2025)的表征与SONAR的语义空间

[1]OmniSONAR 支持 1,500 种语言的文本、177 种语言的语音，以及新增的图像和视频模态 .

中对齐。对齐过程遵循由粗到细的课程，分三个阶段使用视觉描述数据：（1）大规模图像-描述对（1200万），用于粗粒度对齐；（2）合成视频-描述对（200万），用于时间适配；（3）高质量人工标注的视频描述（20万），用于细粒度对齐。我们对v-SONAR 进行了广泛评估。在零样本视频检索任务中，它在PE-VIDEO上取得了 73.03 的 Recall@1，大幅超过 SigLIP2-g-opt（63.91）。在零样本视频描述任务中，它优于最先进的视觉-语言模型，在BLEU 上分别提升了 +18， +4.3，测试集为PE-VIDEO和DREAM-1K，相较于感知语言模型（Cho 等人，2025）。

通过对齐 v-SONAR 与 SONAR，我们表明，在 SONAR 中运行的潜在扩散语言模型 LCM(LCM 团队等，2024年) 经英文文本语料库训练后，能够以零样本方式处理由 v-SONAR 编码的视觉嵌入。在单概念理解任务（即视频描述）中，LCM在 PE-VIDEO、DREAM-1K 和 VATEX 上的表现仅略逊于现有视觉语言模型。类似地，LCM 在多概念推理任务（即长视频摘要生成）上仍具竞争力，评测数据集为 VIDEOXUM。

从视觉语言建模的角度来看，LCM 提出了一种新范式，将视觉和语言模态统一到由 SONAR 和 v-SONAR 共享的模态无关潜在空间中，并通过潜在扩散目标直接预测下一个嵌入。因此，我们进一步引入经过视觉语言指令微调的 LCM，以探索如何最大限度地发挥其在各种下游视觉语言任务中的效用，并将其命名为 v-LCM。 v-LCM 使用 v-SONAR 和 SONAR 对多模态数据（图像、视频和文本）进行编码，并采用相同的潜扩散策略进行训练，遵循原始双塔 LCM 框架（LCM 团队等，2024年）进行文本预训练。

我们在v-LCM 上进行了评估，使用的是多语言多模态指令微调数据集 M3IT（Li 等人，2023年），该数据集涵盖 8 类任务，支持图像和视频两种模态，并覆盖 80 种语言。v-LCM 在图像/视频描述生成、视觉问答及其他生成任务上的表现，可与 InternVL（Chenetal.，2024b）、Qwen-VL（Wang 等人，2024c；Bai 等人，2025）和感知语言模型等其他视觉语言模型不相上下。值得注意的是，在 M3IT 的多语言评估中，v-LCM 在测试的 62 种语言中，有 61 种语言的表现优于其他视觉语言模型，涵盖从高资源到低资源的语言。本文的贡献如下：

- 我们通过事后由粗到细的对齐策略，将语言与模态无关的嵌入空间（Sonar）首次扩展至图像和视频，提出了 v-Sonar。
- 我们证明，v-Sonar 在视频检索和图像描述方面实现了最先进的零样本性能，并能稳健地泛化至多语言环境。
- 我们表明，最初仅使用纯文本数据训练的 LCM，能够有效处理 v-Sonar嵌入，执行零样本单概念和多概念视觉理解任务。
- 我们通过在 v-Sonar 和 Sonar 的共享潜在空间中统一视觉与语言，将 LCM 扩展为潜在扩散视觉语言模型（v-LCM）。在 M3IT 上，v-LCM 的图像描述和问答表现与最先进的视觉语言模型不相上下，同时在 61 种非英语语言上的表现优于它们。

2 v-Sonar

我们首先介绍v-SONAR，这是一个视觉-语言嵌入空间，通过事后将最先进的视觉编码器 感知编码器与多语言文本嵌入空间SONAR对齐而构建。我们选择感知编码器作为基础编码器，主要有两个原因：（1）它在图像和视频两种模态上均实现了最先进的性能（Bolya 等人，2025年）；（2）它与轻量级文本编码器一同经过预训练，因此更容易与SONAR进行事后对齐。这一设计选择使感知编码器区别于其他视觉编码器，例如 v-JEPA（Bardes 等人，2023年；Assran 等人，2025）和 DINO（Oquab 等人，2023年；Siméoni 等人，2025，后者主要侧重于视觉特征学习，而未明确考虑文本对齐。

架构 v-Sonar 的架构 v-SONAR 如图1左侧面板所示。给定输入图像或视频，感知编码器 (PE) 首先会分别对每一帧进行编码。随后，我们在 PE 之上叠加一个轻量级投影器，将编码器的表征适配到 SONAR 空间。投影器首先

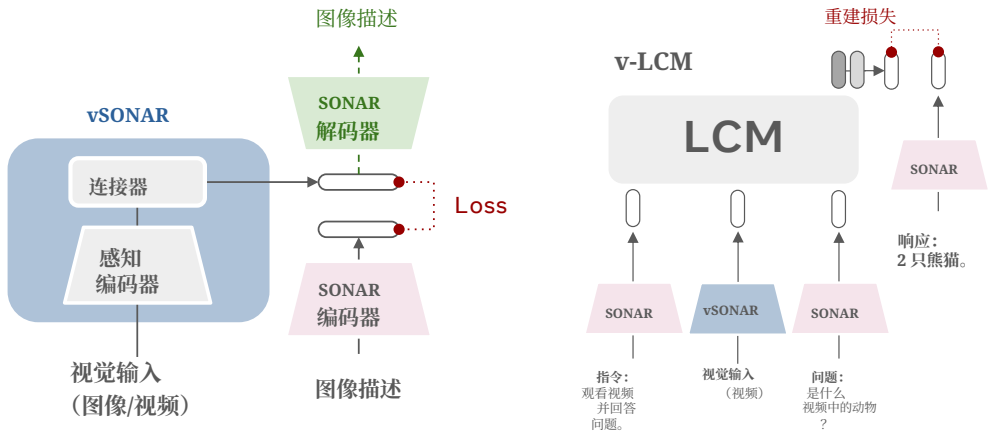


图1 左：v-Sonar 示意图。右：使用视觉-语言指令微调训练 v-LCM。

向所有帧的嵌入注入位置嵌入，从而编码时间顺序信息；随后通过单个时间注意力层实现帧级交互。最后，一个注意力层将帧嵌入聚合为单个视频级表示，作为下游任务的最终嵌入。实现细节请参见附录C 。

从视觉到语言的对齐。 我们使用图像描述数据来训练感知编码器，使其与 Sonar 对齐。我们假设视觉输入和图像描述具有相同的语义含义，因此它们的高层表示在潜在模态无关空间中应尽可能接近。

因此，给定一组 N 配对的视觉输入和图像描述 $\mathcal{D} = \{(V_i, T_i)\}_{i=1}^{N}$，其中 V_i 是图像或视频， T_i 是其对应的图像描述，我们旨在学习一种映射，使视觉嵌入 $\mathbf{z}_v = f_\theta(V_i)$ 和文本嵌入 $\mathbf{z}_t = g(T_i)$ 共享同一语义空间，其中 f_θ 表示可训练的视觉编码器， g 表示冻结的 Sonar文本编码器。为实现语义对齐，我们使用均方误差（MSE）损失，最小化 Sonar 空间中视觉嵌入与文本嵌入之间的差异：

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \| f_\theta(V_i) - g(T_i) \|_2^2. \tag{1}$$

按照教师-学生训练（Reimers和Gurevych，2020年；Duquenne 等人，2023年），SONAR 保持冻结状态，我们仅更新轻量级投影器和视觉编码器中的参数。我们还尝试加入额外的对比损失（Oord等人，2018年；Radford等人，2021），但未发现显著收益；详情和结果见附录A。

我们设计了一种由粗到细的课程，逐步使视觉编码器适应更复杂的语义。对齐分三个阶段进行。在阶段 1，我们使用来自 PLM 数据流水线的 1200 万组大规模图像-描述对初始化对齐（Cho 等人，2025），该流水线由 Segment- Anything（Kirillov 等人，2023年）和 OpenImages（Kuznetsova 等人，2020年）组成。此阶段建立了视觉嵌入与文本嵌入之间的基本映射。随后，我们引入 200 万组配对

模型	xsim ↓	xsim++ ↓
Sonar1	1.37	15.27
OmniSONAR	0.65	6.14

表1 Flores 中 200 种语言的相似度搜索。

来自 YouTube1B 语料库的 PLM 合成视频字幕数据（Cho 等人，2025）。此步骤使视觉编码器适应时间动态，同时保持与SONAR的语义一致性。最后，我们使用来自PE-VIDEO （Bolya 等人，2025）的 20 万组经人工核验的高质量视频-描述对来优化对齐。

我们使用两个版本的SONAR编码器：SONAR1 是已发布并开源的版本（Duquenneet al.，2023年）。这是LCM所支持的版本。我们提前获得了一个改进版本，名为OMNISONAR，该版本使用更多数据进行训练，并增加了三个阶段的对比训练和

	方法	R@1↑	R@5↑	R@10↑	平均倒数排名↑	AC↑	V.迹↑	V.对数行列式↑	T.迹↑	T.对数行列式↑
PE-Vid	SigLIP2-G-OPT	47.55	71.47	79.41	58.47	0.396	0.393	-1.7×10^4	0.582	-1.7×10^4
	PECoreG	63.91	85.98	91.61	73.77	0.476	0.479	-1.4×10^4	0.686	-1.4×10^4
	v-Sonar	**73.03**	**89.75**	**93.81**	**80.50**	**0.519**	**0.700**	-1.2×10^4	**2.216**	-8.0×10^3
DREAM	SigLIP2-G-OPT	61.50	83.50	89.10	71.50	0.263	0.401	-1.8×10^4	0.662	-1.8×10^4
	PECoreG	**72.10**	**89.80**	**93.60**	**79.90**	0.307	0.495	-1.4×10^4	0.639	-1.4×10^4
	v-Sonar	63.30	84.10	89.00	72.46	**0.410**	**0.559**	-1.2×10^4	**2.523**	-8.5×10^3
Vatex	SigLIP2-G-OPT	27.52	57.70	70.06	41.27	0.289	0.352	-1.7×10^4	0.660	-1.7×10^4
	PECoreG	18.90	42.42	54.72	30.42	0.379	0.480	-1.4×10^4	0.508	-1.4×10^4
	v-Sonar	**40.75**	**68.63**	**78.88**	**53.59**	**0.427**	**0.558**	-1.2×10^4	**1.660**	-8.3×10^3

表 2 在 PE-Video、Dream-1k 和 Vatex 上的零样本检索性能。我们报告了召回率@1/5/10 和 MRR分数。我们还报告了嵌入空间的解析指标，包括：1）迹反映总体方差；2）对数行列式（logdet）近似表示空间体积。每列的最佳值以**粗体**标出。

自蒸馏（Omnilingual Embeddings Team et al.，2026）。如表1所示，OmniSONAR在多语言相似性搜索这一代理指标上的表现显著优于Sonar1。指标xsim++包含困难负例（Chen et al.，2023年）。我们在Sonar两个版本的视觉描述生成任务上的消融实验见附录 D。

2.1 v-LCM

大型概念模型（LCM；LCM 团队等，2024年）是一种直接在 Sonar 嵌入空间中运行的潜在扩散语言模型。它遵循自回归范式，根据前面的干净嵌入预测下一个句子嵌入。对于文本模态，所有嵌入均由固定的 Sonar 编码器和解码器进行编码和解码。为对下一个嵌入的条件分布进行建模，LCM 采用扩散目标：给定一个干净嵌入 $x^0 \in \mathbb{R}^d$，前向过程会在方差保持调度下逐步向其添加高斯噪声（Karras et al.，2022）：

$$q(x_t \mid x^0) = \mathcal{N}\big(x_t; \alpha_t x^0, \sigma_t^2 \mathbf{I}\big), \quad x_t = \alpha_t x^0 + \sigma_t \epsilon, \ \ \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

其中，(α_t, σ_t) 由单调递减的对数信噪比调度 $\lambda_t = \log(\alpha_t^2/\sigma_t^2)$ 决定。反向过程由去噪器 $\mu_\theta(x_t, t, c)$ 参数化，并以上下文嵌入 c 为条件，其转移为高斯转移：

$$p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t, c), \sigma_t^2 \mathbf{I}\big). \tag{3}$$

训练通过最小化原始干净嵌入上的重建损失来进行：

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x^0, \epsilon} \left\| x^0 - \mu_\theta(\alpha_t x^0 + \sigma_t \epsilon, t, c) \right\|_2. \tag{4}$$

我们采用 LCM 的双塔变体，将上下文编码器（对前序嵌入进行编码）与去噪器（迭代重建下一个嵌入）分离。

从视觉语言建模的角度来看，LCM代表了一种新范式：它在模态无关的潜在空间先验中融合视觉和文本模态的信息，而非使用离散的视觉和文本词元（Chameleon Team，2024年）。这使得自回归生成能够完全在潜在空间中进行。在此基础上，我们进一步提出v-LCM，这是LCM 的扩展版本，通过视觉语言指令微调进行训练，以增强其在广泛下游视觉语言任务中的实用性。在v-LCM中，视觉输入（图像和视频）使用v-Sonar编码到Sonar 潜在空间中，而文本指令和提示词则由Sonar编码。随后将得到的视觉和文本嵌入拼接成一个序列，并采用与LCM原始纯文本训练相同的潜在扩散框架进行处理，以预测序列中的下一个嵌入。

3 实验

我们首先通过使用与OmniSONAR文本编码器对齐的v-SonarOmni评估文本-视频检索和图像描述，验证将视觉编码器与Sonar对齐的有效性，并进行若干消融实验。

模型	Bleu	R-1	R-2	R-L	BS-P	BS-R	BS-F
PE-Video							
InternVL2.5-1B	19.4	32.1	9.0	23.4	31.2	27.3	29.3
InternVL2-1B	24.1	35.8	10.7	25.5	30.8	32.1	31.5
Qwen2-VL-2B-Instruct	29.9	41.7	18.8	31.2	34.8	40.0	37.3
Qwen2.5-VL-3B-Instruct	30.0	41.3	16.1	28.9	30.2	38.6	34.4
PLM-1B	21.5	37.6	11.9	26.6	35.8	26.2	31.0
PLM-3B	21.1	37.5	11.7	26.4	36.6	26.1	31.3
v-Sonar 搭配 OmniSONAR 解码器	**39.0**	**50.1**	**23.3**	**38.0**	**44.4**	**41.6**	**43.0**
Dream-1k							
InternVL2.5-1B	10.2	21.5	3.7	15.6	**26.1**	11.2	18.6
InternVL2-1B	14.6	25.0	4.3	17.2	23.8	15.2	19.5
Qwen2-VL-2B-Instruct	19.7	27.1	5.2	18.5	12.9	14.8	13.9
Qwen2.5-VL-3B-Instruct	16.1	23.9	4.4	15.9	1.6	15.6	8.6
PLM-1B	18.5	27.0	6.4	19.3	14.5	16.8	15.5
PLM-3B	19.6	28.6	6.7	20.4	19.9	18.1	19.0
v-Sonar（搭配 OmniSONAR 解码器）	**23.9**	**32.7**	**8.4**	**22.7**	19.7	**21.6**	**20.7**
Vatex							
InternVL2.5-1B	41.5	23.3	4.4	19.2	37.6	45.2	40.2
InternVL2-1B	**47.8**	**27.3**	**6.4**	**22.4**	**36.9**	**50.4**	**42.4**
Qwen2-VL-2B-Instruct	32.1	19.8	6.0	16.4	18.7	46.3	30.8
Qwen2.5-VL-3B-Instruct	29.4	18.3	5.1	15.0	12.1	47.1	27.6
PLM-1B	33.4	21.8	5.7	19.1	15.1	48.1	29.6
PLM-3B	34.0	22.1	5.9	19.3	16.9	48.6	30.8
v-Sonar（配合 OmniSONAR 解码器）	26.7	17.2	5.0	14.8	13.6	42.8	26.5
Vatex-zh							
InternVL2-1B-Instruct	22.3	14.1	2.9	11.8	8.7	18.2	12.6
InternVL2.5-1B-Instruct	**33.2**	22.5	4.4	18.8	19.1	31.5	24.2
v-Sonar（配合 OmniSONAR 解码器）	30.6	**32.1**	**8.53**	**26.9**	**23.2**	**48.5**	**33.7**

表3 在 PE-Video、Dream-1k 和 Vatex（英语和中文）上的视频描述性能。指标包括 BLEU、ROUGE（R-1、R-2、R-L）以及 BERTScore（BERTScore精确率、BERTScore召回率、BS-F）。

随后，我们转向LCM的零样本评估，以及v-LCM 在M3IT（Li 等人，2023年）上的评估；这需要使用 v-SONAR1，因为LCM 是在SONAR1上训练的。

3.1 使用 v-SonarOmni 进行概念空间对齐

文本-视频检索 我们将 v-Sonar 视为成对的视觉-文本编码器，首先按照 Bolya 等人（2025）的设置，评估其在文本到视频检索任务中的零样本性能。我们将 v-Sonar 与两个强基线进行比较：最先进的 SigLIP2 视觉编码器（Tschannen 等人，2025）和 v-Sonar 的衍生模型感知编码器（Bolya 等人，2025）。按照 Cho 等人（2025）的评估方案，我们在三个广泛使用的视频描述基准上进行评估：PE-Video（1.5 万对图像描述数据）（Bolya 等人，2025）、Vatex（5000 对图像描述数据）（Wang 等人，2019）和 Dream-1k（1000 对图像描述数据）（Wang 等人，2024a）。除 Recall@1/5/10 等标准检索指标外，我们还引入三项互补指标来分析嵌入空间的属性：（1）对齐一致性（AC）：视觉与文本相似度分数之间的秩相关，反映跨模态对齐质量。（2）迹：视觉与文本嵌入协方差矩阵的迹，表示表征的分布范围。（3）对数行列式（logdet）：协方差矩阵行列式的对数，可解释为嵌入椭球体的体积。

表 2 总结了实验结果和嵌入空间统计数据。在三个数据集上，PE-VIDEO、DREAM-1K和VATEX上，V-SONAR 的表现显著优于 SigLIP2，在 Recall@1 上分别提升了 9.12、1.8 和 13.23 个百分点

，进一步证明了我们的方法在检索任务上的有效性。与原始感知编码器相比，v-Sonar 在 PE-Video 和 Vatex 上取得了显著提升（Recall@1 分别提高了 9.12 和 21.85 分），但在 Dream-1k 上的 Recall@1 降低了 8.8 分，这表明我们的课程对齐策略保留了强大的检索能力。这些结果证实，视觉编码器可以通过事后方式与纯文本训练的嵌入空间（Sonar）成功对齐。最后，我们的嵌入空间分析显示，v-Sonar 保持了更分散的分布。此外，通过冻结原始 Sonar 空间，v-Sonar 实现了最大的文本嵌入离散度，其迹和 logdet 值均为所有对比模型中最高。

文本-视频描述生成 与 SigLIP 2 或 Perception Encoder 等传统视觉编码器不同，将 v-Sonar 对齐到 Sonar嵌入空间后，我们可以利用 Sonar 解码器直接将 v-Sonar 的编码向量言语化。因此，我们对 v-Sonar 进行视频描述的零样本评估，并将其与若干最先进的视觉语言模型（视觉语言模型）进行比较，包括 InternVL-2/2.5（Chen等，2024b）、Qwen-VL 2/2.5（Wang等，2024c；Bai等，2025）以及 Perception语言模型（Cho等，2025）。为进行公平比较，我们选取规模介于 1B 至 3B 的视觉语言模型，因为 Sonar 解码器为 1.5B，v-Sonar 为 1.9B。我们使用包括 BLEU 和 Rouge 分数在内的词汇指标，以及包括 BERTScore-精确率/召回率/F1 在内的语义指标对模型进行评估，遵循（Zhang等，2025）。

结果如表3所示。对于PE-Video 和Dream-1k等细致的图像描述基准，我们发现v-Sonar 与Sonar解码器配合使用时，能够达到最先进的性能。特别是，v-Sonar 在BLEU上比排名第二的模型 Qwen2.5-VL-3B-Instruct 高出 9 分。唯一的例外是Vatex ，其字幕相对较短，通常只有一个句子；v-Sonar 的表现不及 InternVL2。不过，这在预期之中，因为我们主要使用细致的图像描述数据来对齐v-Sonar 和Sonar 。此外，我们发现v-Sonar 的表现仍可与 PLM 和 Qwen-VL系列相媲美。我们使用Vatex-中文验证集进行多语言评估，并主要将v-Sonar 与 InternVL 进行比较，因为据报道 QwenVL 在训练过程中使用了Vatex 中文划分（

Wang 等人，2024c；Bai 等人，2025），以及 PLM-1/3B 均无法支持流畅的中文生成。我们发现，在 Vatex 中文划分中，v-Sonar 仍然优于 InternVL 系列，表明其在多语言评估中的优势。

v-Sonar 1 与 *v-SonarOmni*。我们展示了 Sonar （Duquenne 等人，2023年）与OmniSONAR 在 图 2中的对比。我们同时报告了 Sonar 空间的预言机性能（我们使用Sonar 编码器对参考描述进行编码，并用 Sonar 解码器进行解码）。同时，我们报告了使用 v-Sonar的零样本性能（我们使用 v-Sonar 对视频进行编码，并用 Sonar 解码器进行解码）。 Sonar 预言机性能可作为 v-Sonar 在利用 Sonar 解码器时所能达到的性能上界的估计。

我们观察到，这两个 Sonar 版本都具有很强的预言机性能，这表明 Sonar从文本空间到其表示空间的编码和解码几乎没有损失。具体而言，在 PE-Video、Vatex和Dream-1k上，OmniSONAR 的 BLEU 得分分别可达 81、96 和 70。比较 Sonar 和 OmniSONAR的零样本性能可见，Sonar1 的表现明显较差。我们推测，Sonar更难对齐，因为据报道其空间发生了坍塌。我们对 Sonar和 OmniSONAR 的分析也支持这一观察：在 PVD 中，Sonar 和OmniSONAR 的嵌入范数分别为 0.264 和 1.69，协方差迹分别为 0.049 和 1.83。检索任务的比较见附录 D。

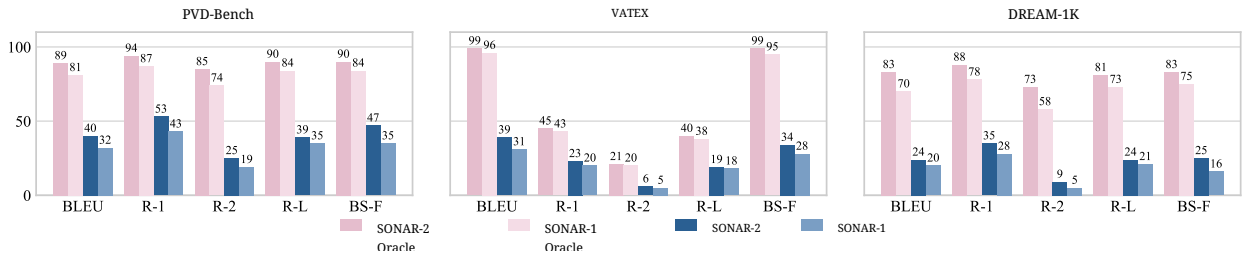


图 2 在三个图像描述数据集上，对使用 SONAR 版本 1 和版本 2（OmniSONAR）嵌入空间训练的 v-Sonar 的比较。

		MSE	余弦相似度	BLEU	R-1	R-2	R-L	BERTScore精确率	BERTScore召回率	BERTScore
Architecture	线性投影	1.45×10^{-3}	0.694	38.0	49.7	21.6	36.7	47.2	40.1	43.7
	完整位置编码	1.54×10^{-3}	0.672	37.1	48.5	21.3	36.5	46.9	38.8	42.9
	+异步学习率	1.43×10^{-3}	0.700	39.7	51.3	23.3	37.7	48.1	42.1	45.1
	+归一化初始化	1.39×10^{-3}	0.708	39.8	51.8	24.0	38.5	49.4	42.2	45.8
	+注意力池化	1.39×10^{-3}	0.708	39.8	51.9	24.0	38.5	49.7	42.4	46.0
	+ 时间注意力	1.39×10^{-3}	0.708	39.8	51.9	24.0	38.5	49.7	42.4	46.1
Pipeline	完整流程	1.36×10^{-3}	0.716	40.1	52.6	24.9	39.2	50.8	43.2	47.0
	w/o 合成视频字幕	1.39×10^{-3}	0.710	39.6	51.9	24.1	38.6	50.0	42.4	46.2
	w/o 图像字幕 & 合成视频字幕	1.39×10^{-3}	0.708	39.8	51.9	24.0	38.5	49.7	42.4	46.1

表4 模型架构与三阶段训练流程的消融研究。SV：我们在第二阶段使用合成视频字幕数据进行课程训练。IC：我们在第一阶段使用图像描述数据进行课程训练。

3.2 消融研究

我们在PE-Video测试集上（表 4）对模型架构设计和我们提出的训练流程进行了消融研究。

模型架构 我们对投影器网络的架构选择进行了消融研究。作为基线，我们评估了线性投影（Linear Proj.），即冻结感知编码器并仅训练一个线性层；以及全模型微调（Full PE），即联合更新编码器。线性投影的表现更好，这表明编码器的对比预训练已实现了良好的语义对齐，而全模型微调则受到随机初始化的投影器产生的不稳定梯度影响。为缓解这一问题，我们采用了一系列逐步提升下游性能的策略，包括为投影器和编码器设置异步学习率、使用初始化技巧、采用基于注意力的聚合策略处理视频帧特征，以及添加时间注意力层。

数据混合 接下来，我们对流程中的第二阶段合成视频字幕训练课程和第一阶段图像描述训练课程进行了消融研究。我们观察到，这两个阶段都能提升 PE-Video 在图像描述任务上的下游性能。移除 1200 万组图像字幕配对数据和 200 万组视频字幕配对数据后，BLEU 分别下降了 0.3 和 0.2。

3.3 使用大型概念模型对 v-Sonar 嵌入进行零样本处理

由于 LCM（LCM 团队等，2024年）直接运行于 Sonar1 之上，因此它应能无缝迁移其能力，并理解 v-Sonar 中与 Sonar1 对齐的视觉概念。我们从单一视觉概念理解任务逐步扩展到多重视觉概念理解任务，考察 LCM 如何处理 v-Sonar：LCM 接收由 Sonar 编码的指令和来自 v-Sonar 的视觉嵌入，并预测目标嵌入。请注意，在这两项实验中，我们都没有对 LCM 进行微调，也没有使用任何视频数据或图像描述数据。因此，LCM 仅在英文文本语料库上训练，包括预训练和指令微调，具体设置与 LCM 团队等（2024年）相同。我们将 LCM 的表现与 InternVL 系列（Chen等，2024b）、Qwen-VL（Bai等，2025；Team，2024）和 PLM（Cho等，2025）的 7/8B 规模视觉语言模型进行比较。

单一视觉概念理解：视频描述 我们在LCM 视频描述任务上的结果见表 5。与最强基线相比，零样本LCM在BLEU得分上分别落后PE-Video、Dream-1k 和Vatex1.15/4.44/4.76 分。在各模型中，PLM-8B 的整体表现最出色。LCM 与具有竞争力的视觉语言模型之间性能差距相对较小，这表明LCM 能够理解单一视觉概念，尽管它从未接受过视频数据训练。

多重视觉概念理解：长视频摘要生成 接下来，我们评估 LCM 在需要理解多个视觉嵌入的场景中的表现。长视频被均匀切分为多个片段

		视频描述 / 摘要生成								M3IT 图像				M3IT 视频		
		PE-Video		Dream-1k		Vatex		VideoXum		COCO	VIQUAE	VisualMRC	ScienceQA	ActivNetQA	MSRVTT-QA	IVQA
		R-L	BS	R-L	BS	R-L	BS	R-L	BS	R-L	R-L	R-L	准确率	R-L	R-L	R-L
InterVL2	1B	25.5	31.5	17.2	19.5	22.4	42.4	15.3	17.7	12.6	24.0	30.6	53.9	40.6	27.6	39.5
	4B	15.0	18.4	12.2	14.7	19.2	42.4	15.6	17.4	17.5	20.0	35.3	89.6	27.5	24.5	31.8
	8B	18.6	23.4	16.4	19.4	16.2	42.4	29.1	26.1	21.0	21.6	42.9	87.2	29.7	27.1	38.4
InternVL-2.5	1B	23.4	29.3	15.6	18.6	19.2	40.3	17.1	23.2	13.2	10.8	27.3	69.0	16.6	11.7	19.3
	4B	14.6	18.0	13.3	14.8	17.3	36.2	18.1	23.0	15.1	23.1	45.2	86.4	26.8	21.8	24.9
	8B	21.4	26.0	17.0	17.0	20.6	42.4	24.9	20.5	16.8	17.3	42.8	**93.1**	20.9	16.9	22.7
Qwen2-VL	2B	31.2	**37.3**	18.5	13.9	16.4	30.8	23.6	29.8	24.9	**50.2**	56.1	54.5	53.7	39.6	49.4
	7B	26.9	32.6	19.8	18.1	28.5	**51.6**	26.0	32.4	23.7	49.7	**57.4**	70.4	41.9	22.7	39.1
Qwen2.5-VL	3B	**28.9**	34.4	15.9	8.6	15.0	27.6	26.0	32.9	25.1	48.3	55.7	55.0	52.1	41.6	48.5
	7B	22.2	25.9	15.7	10.5	27.5	50.8	24.1	28.9	18.5	34.5	45.0	61.6	46.0	41.4	54.2
感知语言模型	1B	26.6	31.0	19.3	15.5	19.1	29.6	21.8	33.2	27.5	30.8	45.5	73.6	27.8	14.5	39.1
	3B	26.4	31.3	20.4	19.0	19.3	30.8	**27.0**	**36.4**	34.3	23.7	51.1	89.8	28.0	19.4	26.1
	8B	27.4	31.9	**20.8**	**19.7**	19.0	30.8	26.2	33.7	36.3	31.0	50.0	87.7	40.5	25.3	41.4
LCM	LCM	25.5	27.9	18.5	16.6	23.8	30.8	21.5	22.1	18.0	34.3	33.5	44.7	51.7	36.0	48.9
	v-LCM	27.4	30.0	19.8	19.2	**28.8**	48.7	20.6	25.3	**38.8**	39.4	34.1	76.2	**63.6**	**48.7**	**63.9**

表 5 M3IT 及此前视频基准（PE-Video、Dream-1k、Vatex、VideoXum）上的视觉语言任务主要结果。

，每个片段包含 8 帧，并由 v-Sonar分别编码为独立的视频嵌入。鉴于LCM 在文档摘要生成方面表现出色（LCM 团队等，2024年），且能处理多个 Sonar嵌入；我们推测，它应该能够对来自 v-Sonar的视频嵌入集合进行零样本摘要生成。本次评估中，我们使用 VideoXum (Lin 等，2023年），该数据集包含时长为一到五分钟的视频，并将其均匀切分为每段 8 帧的片段。

我们报告了VideoXum的结果，见表 5。同样，在相同规模的有竞争力视觉语言模型中，PLM 的表现最为出色。v-Sonar + LCM 的 BertScore-F1 得分为 22.1，低于表现最佳的 PLM-8B（33.7），但略高于 InternVL-2.5-8B 的 20.5。这些发现表明，即使训练期间从未接触过任何视频数据，
LCM 展现出对长视频的多个 v-Sonar嵌入的非平凡理解能力。

在 *v-Sonar* 空间中推理随后，我们研究 LCM 是否真正利用了 v-Sonar 中的潜在表示进行多模态推理。为此，我们比较了两种设置：（1）将视频片段直接编码为 v-Sonar嵌入，再将其输入 LCM 进行零样本摘要生成；（2）使用 Sonar 解码器将视频嵌入解码为图像描述，再用 Sonar 重新编码，并将这些 Sonar 嵌入输入 LCM。我们的假设是，与 Sonar 中的文本等效内容相比，v-Sonar嵌入保留了更丰富的视觉特征，因此，如果 LCM 依赖视觉表示，就应能取得更好的表现。

我们将视频分为短视频（<90秒）、中等长度视频（90–150秒）和长视频（>150秒）三类，并在图3中报告ROUGE-L 分数。在所有类别中，LCM 结合 v-Sonar的表现始终优于结合Sonar的表现。值得注意的是，随着视频长度增加，Sonar的性能会下降，而v-Sonar

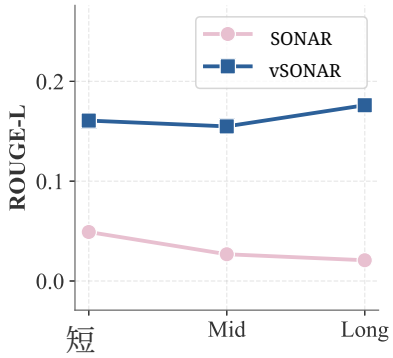


图 3 在 v-Sonar 空间中运行时，LCM 的表现优于仅接收文本 Sonar 输入的情况。我们在 VideoXum 中对 LCM-7B-IFT 进行评估，并比较其在短、中、长视频输入类别上的 Rouge-L 分数。

的性能保持稳定，凸显了其鲁棒性。这些结果支
持我们的假设：LCM 直接在 v-Sonar 提供的视觉嵌入空间中进行推理，该空间包含比 Sonar 文本输入表征更丰富的视觉信息。

v-LCM 结合视觉细节进行推理。 为评估我们的对齐是否感知编码器到 SONAR 保留了对下游推理至关重要的视觉细节，我们进一步评估了 M3IT 上的视觉常识推理（VCR）表现（表 6）。在此基准测试中，模型必须在解读相关边界框的同时，运用常识知识对物体进行推理。与单纯的语义级对齐相比，VCR 对视觉落地和布局感知的要求要高得多（见图21）。我们报告了相对于真实推理依据的词元级 F1 分数，以及与参考解释的语义相似度。v-LCM 表现出色，这表明尽管仅使用语义级图注进行训练，v-LCM 仍能有效利用布局定位以及 v-Sonar 在推理任务中保留的空间关系。

VCR	F1	Sim.
LCM	0.385	0.258
vLCM	**0.671**	**0.529**
PLM-8B	0.441	0.432
Qwen-2.5-7B	0.275	0.402
Qwen-2-7B	0.502	0.513
InterVL2.5-8B	0.155	0.158
InterVL2-8B	0.325	0.340

表 6 VCR 结果：词元级 F1 与语义相似度的比较。

3.4 v-LCM

任务内表现 接下来，我们评估 v-LCM，它经过 M3IT（Li 等人，2023年）上的监督微调，以更好地捕捉视觉概念并进行推理。我们主要采用 M3IT（Li 等人，2023年），因为它支持多种任务，覆盖多达 80 种语言，涵盖高资源语言和低资源语言。评估涉及 7 个数据集，涵盖 M3IT 定义的 5 项任务：（1）图像描述（COCO）；（2）视觉问答（VIQUAE）；（3）文档图像问答（VisualMRC）；（4）视频描述（MSRVTT）；（5）问答（IVQA、MSRVTT-QA、ActivityNetQA）。此外，我们还报告v-LCM在上一节介绍的视频描述和长视频摘要基准测试上的表现。按照 Li 等人 (2023年) 的评估方案，我们使用 ROUGE-L 评估生成任务（例如图像描述和开放式问答），并使用准确率评估选择题问答。

表 5 对比了 LCM与表现强劲的开源视觉语言模型。我们发现，v-LCM在大多数基准测试中都显著优于零样本 LCM。例如，v-LCM 在 IVQA 上取得了 63.9 的 R-L 分数，在 ActivityNetQA 上取得了 63.6 的 R-L 分数，超过了 LCM的 48.9 和 51.7，充分体现了使用视觉指令微调数据训练带来的显著提升。尽管 v-LCM 在 VisualMRC、VIQUAE 和 ScienceQA 等部分基准测试中不及表现最佳的模型，但它在视频问答任务上取得了最先进的结果，包括 IVQA、ActivityNetQA 和 MSRVTT-QA。同时，在我们此前的视频描述和摘要生成数据集上，其表现仍具竞争力： v-LCM 在 PE-Video 上取得了 27.4 的 R-L 分数，在 DREAM-1K 上取得了 19.8 的 R-L 分数，仅比最佳模型低 1.5 和 1 个 ROUGE 分，凸显了 LCM 对训练期间未见数据集的泛化能力。

多语言性 我们进一步对v-LCM在涵盖 62 种语言的 M3IT基准测试上进行多语言评估，[2] 利用了v-LCM完全在SONAR和v-SONAR的潜在空间中运行这一特点，因此能够将输出解码为SONAR支持的任意语言。评估涵盖五种不同任务，包括图像分类（ImageNet）、图像问答（VQA-V2、OKVQA）、视频问答（MSRVTT-QA）、视频描述（MSRVTT）和叙事生成（VIST），覆盖从高资源语言（如中文）、中资源语言（如日语）到低资源语言（如爪哇语）的多种语言。我们使用ROUGE-L 的实现，该实现来自（Shohan 等人，2024年），用于多语言评估。

如图 4所示，v-LCM 在 62 种语言中的 61 种语言上始终优于 Qwen2.5-VL-7B 和 PLM-8B，唯一的例外是荷兰语。在一些高资源语言中，提升幅度较小（例如

[2]Sonar、M3IT 和多语言 ROUGE 支持的所有语言的交集。

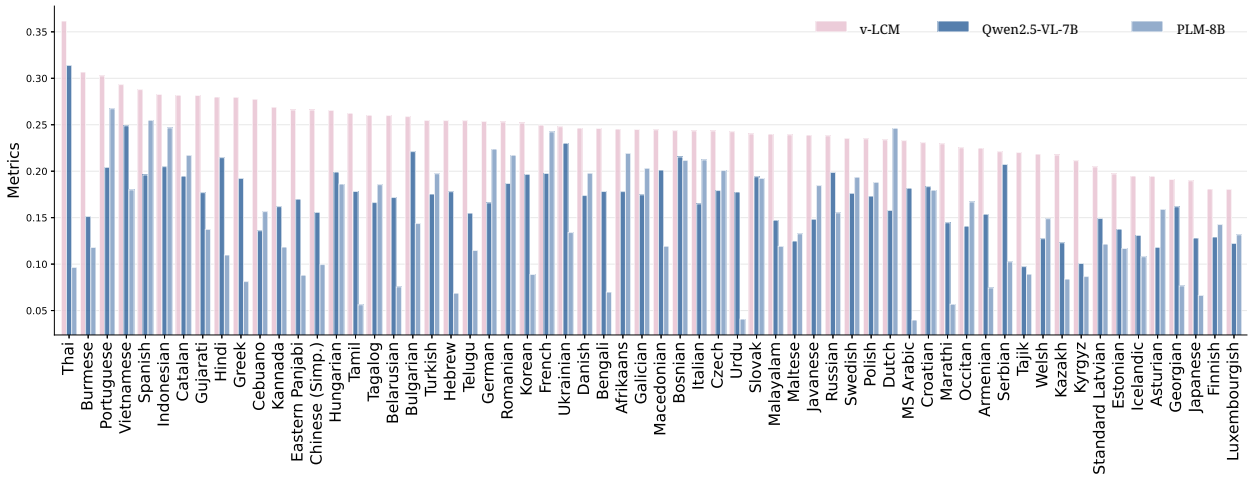


图 4 v-LCM 在 62 种语言上的表现、Qwen2.5-VL-7B 和 PLM-8B 在 M3IT 测试集上的表现，涵盖 MSRVTT、MSRVTT-QA、ImageNet、VQA-V2、VIST 和 OKVQA。我们报告所有数据集的平均ROUGE-L 分数。各数据集的详细结果见附录 E。

法语），而在中低资源语言中，提升则十分显著，包括缅甸语、塔吉克语和泰卢固语。值得注意的是，对于乌尔都语、现代阿拉伯语和泰米尔语等语言，基于 LLaMA-3.2 的 PLM-8B 不支持这些语言（Touvron 等人，2023年；Dubey 等人，2024年），v-LCM 仍能成功生成有意义的输出，而竞争模型则完全无法生成。

4 相关工作

多模态学习的一种核心范式是将视觉与语言表示对齐到共享嵌入空间中。早期方法如 CLIP（Radford等人，2021）和 ALIGN（Jia 等人，2021）通过配对图像和字幕进行大规模对比学习，从而实现向下游任务的零样本迁移。后续研究将这一思路扩展到视频-语言预训练（Lei 等人，2021；Xu 等人，2021；Wang 等人，2022）。近期工作着重于将预训练编码器对齐到统一空间：感知编码器（Bolya 等人，2025）将多种感知模态映射到共享潜在空间；而在 Florence（Yuan 等人，2021）和 SigLIP2（Tschannen 等人，2025）等模型中扩展数据和架构，则进一步提升了对齐质量。近期研究还表明，将大语言模型用作文本编码器可以增强视觉-语言对齐（Stone 等人，2025）；此外，研究者提出了事后对齐策略，作为联合训练的轻量级替代方案（Brokowski 等人，2025；Yang 等人，2025）。

多语言文本嵌入模型也取得了并行进展，例如 LASER（Artetxe and Schwenk，2019；Heffernan 等人，2022）、LaBSE（Feng 等人，2020）和 Sonar（Duquenne 等人，2023；Omnilingual Embeddings Team 等人，2026），这些模型证明了语言无关嵌入空间在数百种语言中的有效性。模块化方法进一步探索了语言专用组件，以减少通用嵌入空间中的干扰（Huang 等人，2024）。这些通用文本嵌入为对齐视觉编码器提供了理想目标，因为它们无需在每种语言中使用多模态数据，就能继承跨语言泛化能力。先前的研究已在语音到文本对齐中探索过类似策略（Chung 等人，2018；Duquenne 等人，2021a；Laperrière 等人，2024；Du 等人，2024），但将视觉嵌入大规模对齐到此类通用文本空间，仍缺乏充分研究。

5 结论

我们通过将图像和视频模态纳入 Sonar 嵌入空间，推出了 v-Sonar。据我们所知，这使 Sonar 成为覆盖四种模态（文本、语音、图像和视频）且支持多达 200 种语言的最通用嵌入空间。我们提出了一种三阶段训练方法，将基于感知编码器的池化表示映射到 Sonar 语义表示。我们

在文本到视频检索和视频描述任务中取得了极具竞争力的结果。大型概念模型（LCM；LCM 团队等，2024年）是一种近期提出的方法，旨在更高的语义概念层面（即 SONAR）进行推理。通过 V-SONAR进行编码后，我们展示了 LCM能够以零样本方式处理图像或视频嵌入，而无需这些模态的训练数据。我们还推出了 V-LCM ，并采用多模态指令微调，使其达到最先进视觉语言模型的水平，同时在涵盖 61 种从高资源到低资源的语言的任务中显著超越这些模型。

参考文献

Mikel Artetxe 和 Holger Schwenk。用于零样本跨语言分类及其他任务的大规模多语言句子嵌入。*TACL*，第 597–610 页，2019 年。

Mido Assran、Adrien Bardes、David Fan、Quentin Garrido、Russell Howes、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zholus 等。V-JEPA 2：自监督视频模型助力理解、预测与规划。*arXiv* 预印本 *arXiv:2506.09985*，2025。

Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025。

Adrien Bardes、Quentin Garrido、Jean Ponce、Xinlei Chen、Michael Rabbat、Yann LeCun、Mido Assran 和 Nicolas Ballas。V-JEPA：用于视觉表征学习的潜在视频预测。已提交至 *ICLR'25*，2023 年。

Daniel Bolya、Po-Yao Huang、Peize Sun、Jang Hyun Cho、Andrea Madotto、Chen Wei、Tengyu Ma、Jiale Zhi、Jathushan Rajasegaran、Hanoona Rasheed 等。感知编码器：最佳视觉嵌入并非位于网络的输出端。*arXiv* 预印本 *arXiv:2504.13181*，2025。

Trevor Brokowski、Alexandre Sallinen 和 Mary-Anne Hartley。SSCA：SigLIP-2 Sonar概念对齐。载于第二届视觉概念研讨会，2025。

Chameleon Team。Chameleon：混合模态早期融合基础模型。*arXiv* 预印本 *arXiv:2405.09818*，2024年。

Jianlv Chen、Shitao Xiao、Peitian Zhang、KunLuo、Defu Lian 和 Zheng Liu。M3-Embedding：通过自知识蒸馏实现多语言性、多功能性和多粒度文本嵌入，2024a。

Mingda Chen、Kevin Heffernan、Onur Çelebi、Alexandre Mourachko 和 Holger Schwenk。xSIM++：一种用于评估低资源语言双语文本挖掘性能的改进代理指标。载于 Anna Rogers、Jordan Boyd-Graber 和 Naoaki Okazaki（编），*ACL*，第 101–109 页，2023年。 https://aclanthology.org/2023.acl-short.10/。

Zhe Chen、Jiannan Wu、Wenhai Wang、Weijie Su、Guo Chen、Sen Xing、Muyan Zhong、Qinglong Zhang、Xizhou Zhu、Lewei Lu 等。Internvl：扩展视觉基础模型并使其适用于通用视觉语言任务。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 24185–24198 页，2024b。

Jang Hyun Cho、Andrea Madotto、Effrosyni Mavroudi、Triantafyllos Afouras、Tushar Nagarajan、Muhammad Maaz、Yale Song、Tengyu Ma、Shuming Hu、Suyog Jain 等。Perceptionlm：用于细致视觉理解的开放获取数据和模型。*arXiv*预印本 *arXiv:2504.13180*，2025。

Yu-An Chung、Wei-Ning Hsu、Hao Tang 和 James Glass。语音和文本嵌入空间的无监督跨模态对齐。*神经信息处理系统进展*，31，2018年。

Zhihao Du、Qian Chen、Shiliang Zhang、Kai Hu、Heng Lu、Yexin Yang、Hangrui Hu、Siqi Zheng、Yue Gu、Ziyang Ma 等。CosyVoice：一种基于监督式语义标记的可扩展多语言零样本文本转语音合成器。*arXiv* 预印本 *arXiv:2407.05407*，2024年。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan 等。Llama 3 2407，2024年的模型群。*arXiv* 电子预印本，页码 arXiv–。

Paul-Ambroise Duquenne、Hongyu Gong 和 Holger Schwenk。用于大规模语音挖掘的多模态和多语言嵌入。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，第 34 卷，第 15748–15761 页。Curran Associates, Inc.，2021a。 https://proceedings.neurips.cc/paper_files/paper/2021/file/8466f9ace6a9acbe71f75762ffc890f1-Paper.pdf。

Paul-Ambroise Duquenne、Hongyu Gong 和 Holger Schwenk。面向大规模语音挖掘的多模态和多语言嵌入。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan 编，

神经信息处理系统进展，第34卷，第15748–15761页。Curran Associates，Inc.，2021年b。https://proceedings.neurips.cc/paper_files/paper/2021/file/8466f9ace6a9acbe71f75762ffc890f1-Paper.pdf。

Paul-Ambroise Duquenne、Holger Schwenk 和 Benoît Sagot。Sonar：句子级多模态和语言无关表示。*arXiv*预印本 *arXiv:2308.11466*，2023 年。Fangxiaoyu Feng、Yinfei Yang、Daniel Cer、Naveen Arivazhagan 和 Wei Wang。语言无关的BERT句子嵌入。*arXiv*预印本 *arXiv:2007.01852*，2020年。Kevin Heffernan、Onur Çelebi 和 Holger Schwenk。利用蒸馏句子表示为低资源语言进行双语文本挖掘。*arXiv*预印本 *arXiv:2205.12654*，2022年。Yongxin Huang、Kexin Wang、Goran Glavaš 和 Iryna Gurevych。模块化句子编码器：将语言专门化与跨语言对齐分离。*arXiv*预印本 *arXiv:2407.14878*，2024年。Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。通过噪声文本监督扩展视觉和视觉语言表示学习。载于 国际机器学习会议，第4904–4916页。PMLR，2021。Tero Karras、Miika Aittala、Timo Aila 和 Samuli Laine。阐释基于扩散的生成模型的设计空间。神经信息处理系统进展，35：26565–26577，2022年。Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo 等。任意分割。载于*IEEE/CVF*国际计算机视觉会议论文集，第4015–4026页，2023年。Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov 等。《Open Images数据集v4：大规模统一图像分类、目标检测和视觉关系检测》。载于国际计算机视觉期刊，128 1981，2020 7 1956( ): – 。Gaëlle Laperrière、Sahar Ghannay、Bassam Jabaian 和 Yannick Estève。采用双任务学习方法，对多语言语义语音编码器进行微调，以实现口语理解。*arXiv*预印本 *arXiv:2406.12141*，2024年。

LCM 团队、Loïc The Barrault、Paul-Ambroise Duquenne、Maha Elbayad、Artyom Kozhevnikov、Belen Alastruey、Pierre Andrews、Mariano Coria、Guillaume Couairon、Marta R Costa-jussà、David Dale 等。大概念模型：在句子表示空间中的语言建模。*arXiv* 预印本 *arXiv:2412.08821*，2024年。

Jie Lei、Tamara L. Berg 和 Mohit Bansal。少即是多：通过稀疏采样实现视频与语言学习的 ClipBERT。发表于 *CVPR*，2021。

Lei Li、Yuwei Yin、Shicheng Li、Liang Chen、Peiyi Wang、Shuhuai Ren、Mukai Li、Yazheng Yang、Jingjing Xu、Xu Sun 等。M3IT：面向多模态多语言指令微调的大规模数据集。*arXiv* 预印本 *arXiv:2306.04387*，2023年。Jingyang Lin、Hang Hua、Ming Chen、Yikang Li、Jenhao Hsiao、Chiuman Ho 和 Jiebo Luo。VideoXum：视频的跨模态视觉与文本摘要。*IEEE*多媒体汇刊，26:5548–5560，2023年。Jianmo Ni、Gustavo Hernández Ábrego、Noah Constant、Ji Ma、Keith B. Hall、Daniel Cer 和 Yinfei Yang。Sentence-T5：基于预训练文本到文本模型的可扩展句子编码器，2021年。https://arxiv.org/abs/2108.08877。

NLLB 团队、Marta R. Costa-jussà、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi、Janice Lam、Daniel Licht、Jean Maillard、Anna Sun、Skyler Wang、Guillaume Wenzek、Al Youngblood、Bapi Akula、Loic Barrault、Gabriel Mejia-Gonzalez、Prangthip Hansanti、John Hoffman、Semarley Jarrett、Kaushik Ram Sadagopan、Dirk Rowe、Shannon Spruit、Chau Tran、Pierre Andrews、Necip Fazil Ayan、Shruti Bhosale、Sergey Edunov、Angela Fan、Cynthia Gao、Vedanuj Goswami、Francisco Guzmán、Philipp Koehn、Alexandre Mourachko、Christophe Ropers、Safiyyah Saleem、Holger Schwenk 和 Jeff Wang。不落下任何语言：以人为中心的机器翻译扩展，2022。https://arxiv.org/abs/2207.04672。

Omnilingual Embeddings Team、Pere-Lluís Huguet Cabot、João Maria Janeiro、Yen Meng、Ioannis Tsiamas†、Vivek Iyer、Jaehyeong Jo、Guillem Ramírez、Belen Alastruey、Loic Barrault、David Dale、Kevin Heffernan、Artyom Kozhevnikov、Alex Mourachko、Christophe Ropers、Holger Schwenk、Yu-An Chung、Marta R. Costa-Jussa 和 Paul-Ambroise Duquenne。SONAR2：面向文本和语音的全语言句子嵌入，2026 年。即将出版。

Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。通过对比预测编码进行表示学习。*arXiv* 预印本 *arXiv:1807.03748*，2018年。

Maxime Oquab、Timothée Darcet、Théo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无需监督地学习鲁棒视觉特征。arXiv 预印本 arXiv:2304.07193，2023年。Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于《国际机器学习会议》，第 8748–8763 页。PMLR，2021。Gowtham Ramesh、Sumanth Doddapaneni、Aravinth Bheemaraj、Mayank Jobanputra、Raghavan AK、Ajitesh Sharma、Sujit Sahoo、Harshita Diddee、Mahalakshmi J、Divyanshu Kakwani、Navneet Kumar、Aswin Pradeep、Srihari Nagaraj、Kumar Deepak、Vivek Raghavan、Anoop Kunchukuttan、Pratyush Kumar 和 Mitesh Shantadevi Khapra。Samanantar：规模最大的公开可用平行语料库集合，涵盖 11 种印度语系语言。TACL，：–。https://aclanthology.org/.tacl-./。Nils Reimers 和 Iryna Gurevych。利用知识蒸馏将单语句子嵌入转化为多语言句子嵌入。载于 EMNLP，第 4512–4525 页，2020年。Holger Schwenk、Guillaume Wenzek、Sergey Edunov、Edouard Grave、Armand Joulin 和 Angela Fan。CCMatrix：从网络中挖掘数十亿条高质量平行句。载于 https://arxiv.org/ abs/ 1911.04944，2019年。Faisal Tareque Shohan、Mir Tafseer Nayeem、Samsul Islam、Abu Ubaida Akash 和 Shafiq Joty。XL-HeadTags：利用多模态检索增强，实现新闻标题和标签的多语言生成。载于 ACL（Findings），2024。Oriane Siméoni、Huy V Vo、Maximilian Seitzer、Federico Baldassarre、Maxime Oquab、Cijo Jose、Vasil Khalidov、Marc Szafraniec、Seungeun Yi、Michaël Ramamonjisoa 等。DINOv3。arXiv 预印本 arXiv:2508.10104，2025。Austin Stone、Hagen Soltau、Robert Geirhos、Xi Yi、Ye Xia、Bingyi Cao、Kaifeng Chen、Abhijit Ogale 和 Jonathon Shlens。通过改进语义引导学习视觉组合。载于《计算机视觉与模式识别会议论文集》，第 3740–3750 页，2025。通义千问团队。Qwen2 技术报告。arXiv 预印本 arXiv:2407.10671，2024。Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。Llama：开放且高效的基础语言模型。arXiv 预印本 arXiv:2302.13971，2023。Michael Tschannen、Alexey Gritsenko、Xiao Wang、Muhammad Ferjad Naeem、Ibrahim Alabdulmohsin、Nikhil Parthasarathy、Talfan Evans、Lucas Beyer、Ye Xia、Basil Mustafa 等。SigLIP 2：具备更强语义理解、定位能力和密集特征的多语言视觉-语言编码器。arXiv 预印本 arXiv:2502 14786，2025。Jiawei Wang、Liping Yuan、Yuchen Zhang 和 Haomiao Sun。Tarsier：大型视频描述模型的训练与评估方法。arXiv 预印本 arXiv:2407.00634，2024a。Liang Wang、Nan Yang、Xiaolong Huang、Linjun Yang、Rangan Majumder 和 Furu Wei。多语言E5文本嵌入：技术报告，2024b。https://arxiv.org/abs/2402.05672。Minghang Wang、Zeqiu Ma、Yuanjun Xie、Xuewen Li 等。All-in-one：探索统一的视频-语言预训练。载于 CVPR，2022。Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-VL：提升视觉-语言模型对任意分辨率下世界的感知能力。arXiv 预印本 arXiv:2409.12191，2024c。Xin Wang、Jiawei Wu、Junkun Chen、Lei Li、Yuan-Fang Wang 和 William Yang Wang。Vatex：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》，第 4581–4591 页，2019年。Hu Xu、Gargi Ghosh、Po-Yao Huang、Dahuang Zhang、Joshua Ainslie 等。VideoCLIP：用于零样本视频-文本理解的对比预训练。载于 EMNLP，2021。Jingfeng Yang、Ziyang Wu、Yue Zhao 和 Yi Ma。使用固定文本编码器进行语言-图像对齐。arXiv 预印本 arXiv:2506.04209，2025。Lu Yuan、Dongdong Chen、Yi-Ling Chen、Noel Codella、Xiyang Dai 等。Florence：面向计算机视觉的新型基础模型。载于 CVPR，2021。

Chunhui Zhang、Yiren Jian、Zhongyu Ouyang 和 Soroush Vosoughi。《预训练图像-文本模型其实是视频描述生成器》。载于*2025* 年计算语言学协会美洲分会会议论文集：人类语言技术（第 *2* 卷：短论文），第 292–305 页，2025 年。

附录

用于对齐感知编码器与 Sonar 的对比损失

我们还探索了在均方误差损失之外，使用对比损失来对齐 Perception 编码器与 Sonar。具体而言，给定一个由 B 配对样本 $\{(V_i,\ T_i)\}_{i=1}^{B}$组成的小批次，我们不仅要最小化匹配对之间的距离（$f_\theta(V_i)$ $g(T_i)$），还要拉开不匹配对之间的距离。我们将对比损失定义为：

$$\mathcal{L}_{\text{con}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\text{sim}(f_\theta(V_i), g(T_i))/\tau\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}(f_\theta(V_i), g(T_j))/\tau\right)}, \tag{5}$$

其中，$\text{sim}(\cdot,\cdot)$ 表示余弦相似度，τ 是温度参数。最终损失是均方误差对齐损失与对比损失的加权组合：

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda\mathcal{L}_{\text{con}}, \tag{6}$$

其中，λ 控制对比项的强度。然而，在我们的初步实验中，加入对比项并未使图像描述性能显著优于仅均方误差目标（表 7），但提升了检索性能。不过，由于我们将 v-Sonar 用于 v-LCM 的下游用途更接近生成任务，因此选择仅均方误差损失作为最终损失。

	图像描述			检索	
	Bleu	R-L	BS-F1	R@1	MRR
仅均方误差	**38.9**	**37.8**	**44.9**	49.0	60.3
均方误差 + 对比	38.6	37.5	44.5	**52.4**	**63.7**

表 7 仅使用均方误差损失与加入对比损失的消融研究。报告了在 PE-Video 基准上使用单个多层感知机层作为 v-Sonar 连接器进行图像描述和检索的结果。

为进一步解释表 7中检索性能与图像描述性能之间的权衡，我们假设对比训练可能会使视觉向量偏离精确的 SONAR 流形。它只通过余弦间隔约束相对排序，并生成范数或局部协方差与 SONAR 解码器训练时所用数据不同的嵌入。即使检索性能有所提升，这种轻微的流形偏移也可能降低生成式重建效果。

我们分析了v-Sonar 嵌入在表 8中的统计特性。显而易见，v-Sonar 采用对比损失训练后，嵌入的分离度更高，体现为其范数、迹和体积更大。这表明，与仅均方误差相比，对比损失会将 vSONAR嵌入推向分布更分散的状态（范数、协方差迹和体积的值更高）。然而，无论是均方误差还是余弦相似度，其对齐一致性都较差；我们通过计算每个样本的视觉与文本相似度排序之间的相关性来衡量这一点。这表明，对比损失破坏了局部协方差结构，并可能破坏与 SONAR 流形的对齐。

	视觉范数	V.迹	视觉体积	余弦对齐一致性	均方误差对齐一致性
MSE	1.22	0.48	-10007.16	**0.41**	**0.32**
均方误差+对比	**1.30**	**1.74**	**-8107.94**	0.31	0.13

表 8 v-Sonar 表征在 PE-Video 上的其他统计数据。我们报告 v-Sonar 预测的视频嵌入的范数（V.Norm）、迹（V.Trace）、体积（V.Volume），以及以余弦相似度（AC Cosine）和均方误差（AC MSE）衡量的对齐一致性。

B 数据集统计

表9 总结了我们用于对齐的三阶段训练流程中所使用的数据集。PLM-Image 数据集（SA1B 和 OpenImages）提供了大规模的图像-描述对，对于改善落地和语言丰富度尤为重要。PLM-Video-Auto-YT1B 数据集提供了平均时长为 22.75 秒的视频-文本对，使模型能够捕捉多模态内容中的时间动态。最后，PE-Video 数据集提供了经过精心整理的人工标注的视频-描述对，长度适中，可在后续阶段提供质量更高的监督信号。这些数据集共同平衡了规模与质量，确保不同模态之间既有广泛覆盖，又能精准对齐。

数据集	#样本	时长 (s)	描述长度（句）	描述长度（词）
PLM-Image-Auto-SA1B	7.99M	–	10.7	181.8
PLM-Image-Auto-OpenImages	1.37M	–	7.9	132.1
PLM-Video-Auto-YT1B	2.14M	22.8	2.3	95.5
PE-Video	118K	16.7	4.4	51.4

表9 用于三阶段训练对齐的数据集统计信息。我们报告样本数量、平均视频时长（如适用），以及以句子和单词计的平均图像描述长度。

C 实现

v-SONAR 架构 我们的模型构建于 Perception Encoder Vision Transformer 主干网络 PE-Core-G14-448[3] 之上。编码器以 448×448 像素的分辨率处理 RGB 图像，将其划分为 14×14 个图像块，并为每个画面生成 1024 个图像块。视觉塔由 50 个 Transformer 层组成，每层的隐藏宽度为 1024，包含 16 个注意力头和一个 4096 维的前馈网络，参数量约为 19 亿。对于视频输入，我们均匀采样 8 帧，并从编码器中提取 1536 维的帧级嵌入，随后将其投影到 1024 维的 SONAR 嵌入空间。为了连接感知特征与目标空间，我们附加了一个轻量级连接器，其权重可选择性地初始化为高斯分布（$\mu = 0,\ \sigma = 1e-5$），并使用零偏置以确保稳定性。为捕捉时间动态，连接器使用正弦位置编码增强编码器输出，并在帧之间应用时间多头自注意力模块（8 个注意力头，dropout 为 0.1），同时结合残差连接。随后，使用基于注意力的池化对所得序列进行聚合，其中可学习的 CLS 标记通过一个 8 头注意力模块关注帧嵌入；此外，我们也评估了平均池化和最大池化变体。最终的池化表示（1536 维）随后通过线性多层感知机层映射到 1024 维的 SONAR 空间。

v-SONAR 训练详情 为稳定训练，投影器使用均值为零、方差较小（$1e$–5）的高斯分布进行初始化，从而缓解将高维感知编码器特征映射到目标嵌入空间时出现的梯度爆炸。我们采用两阶段训练方案：前 2,000 步冻结感知编码器，仅优化投影器，使其能够调整而不影响预训练编码器。随后，联合优化投影器和感知编码器，并采用异步学习率：投影器使用较高的学习率（$1e$–4），以便快速调整；感知编码器使用较低的学习率（$1e$–5），以保留预训练知识。

我们采用三阶段课程训练模型。阶段 1（图像描述）训练 15 个训练轮次，批次大小为 512，基础学习率为 1×10^{-5}，连接器学习率为 1×10^{-4}，并为连接器设置 4000 步预热。阶段 2（合成视频字幕数据）训练 10 个训练轮次，有效批次大小为 128，学习率为 1×10^{-5}，连接器学习率为 1×10^{-4}，并设置 2000 步预热。阶段 3（人工验证的视频描述数据）采用与阶段 2 相同的设置。所有阶段均使用 AdamW、余弦学习率衰减和 500 步线性预热计划进行优化。我们将随机种子固定为 42，并在每个阶段使用 2000 个验证样本进行评估。训练通过以下方式分布式进行：

[3]https://huggingface.co/Facebook/PE-Core-G14-448

在 64 个英伟达 A100-80G GPU 上采用全分片数据并行（FSDP），使用 bfloat16 精度，并通过梯度累积提高内存效率；采用提前停止策略，耐心值为 3 个训练轮次，并保存表现最佳的验证模型检查点。

v-LCM 训练详情 为训练 v-LCM，我们采用 LCM 双塔架构（LCM 团队等，2024年），并使用基于扩散的下一句微调目标。优化器为 AdamW，$\epsilon = 10^{-6}$，权重衰减设为 0.01，梯度裁剪阈值设为 25.0，学习率为 3×10^{-5}，采用余弦衰减调度，在前 300 步进行预热，并逐步降至最终学习率 10^{-6}。训练最多进行 10,000 步，批量大小根据需要动态调整，最多使用 7168 个潜在嵌入，梯度累积设为 1。每 1000 步保存一次检查点，并根据验证集表现选取最佳模型。该准则采用 LCM 中使用的 0.15 条件引导概率，损失通过求和损失函数进行归约。训练采用全分片数据并行（FSDP）和 bf16 精度，以提高效率。数据加载设置为从 M3IT 的整个训练集中均匀采样，采用按长度排序的批处理，不进行打包。实验在 1 个节点上使用 8 个 A100 GPU（80GB）进行。

D Sonar 与 OmniSONAR

我们还报告了表 10中的检索性能对比。尽管OmniSONAR总体上占据优势，但 SONAR1 仍极具竞争力（例如，在 PVD-Bench 上的 R@1 为 64.9，与 PECoreG 相比）。

		R@1	R@5	R@10	平均倒数排名
PVD-Bench	SONAR1	0.649	0.843	0.895	0.737
	OmniSONAR	0.730	0.898	0.938	0.805
DREAM-1K	SONAR1	0.536	0.759	0.838	0.638
	OmniSONAR	0.633	0.841	0.890	0.725
VATEX	SONAR1	0.119	0.263	0.350	0.195
	OmniSONAR	0.408	0.686	0.789	0.536

表 10 在三个文本到视频检索数据集上，对使用 SONAR 版本 1 和 2 嵌入空间训练的 v-Sonar 进行比较。

E 详细多语言评估

本节展示了我们在所有受支持数据集上进行多语言评估的结果。具体而言，我们测试了 Sonar 和 M3IT 涵盖的所有语言，并报告 ROUGE 分数作为主要指标。 图10展示了 ImageNet 上图像描述任务的结果，而 图5和 图6 分别报告了 MSR-VTT 上视频描述和视频问答的结果。 图7 展示了 OKVQA 上的结果，图8展示了 VIST 上故事生成的表现，图9展示了 VQA-v2 上图像问答的结果。除 ImageNet 外，我们的模型在所有测试语言上始终优于基线，这很可能是因为 ImageNet 被广泛使用，且现有视觉语言模型对其覆盖充分；唯一的例外是 VQA-v2 中的泰语。

F v-Sonar 潜在空间的可视化

为了定性评估对齐的效果，我们使用 t-SNE（感知编码器 到 Sonar 的潜在空间，在每个对齐阶段前后进行可视化，展示视频和 SONAR 嵌入图 11）。每个对齐阶段结束后，我们都观察到更好的聚类结构：视频嵌入与对应的 Sonar 嵌入彼此距离更近，这表明对齐成功缩小了潜在空间中的模态差距，并促进了跨模态共享语义表示，从而验证了该对齐策略。

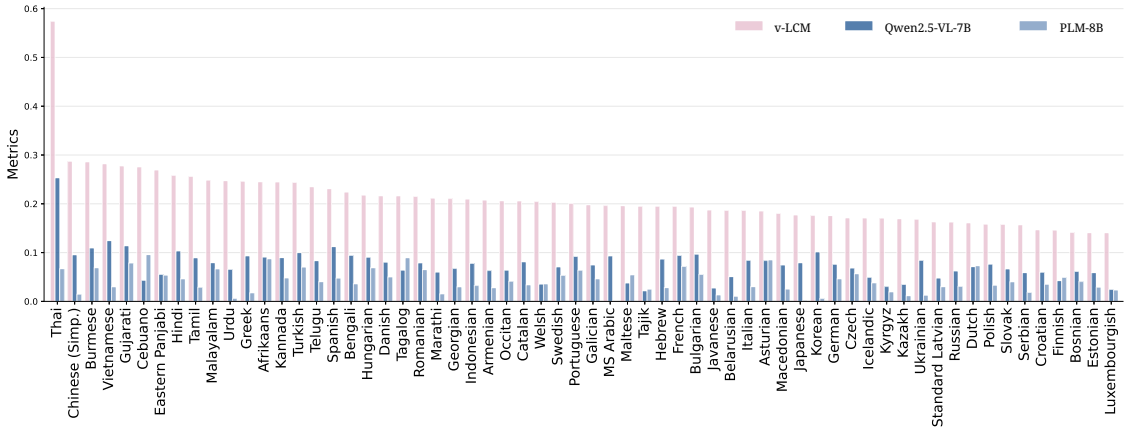


图5 针对MSRVTT的M3IT 61种语言评估。

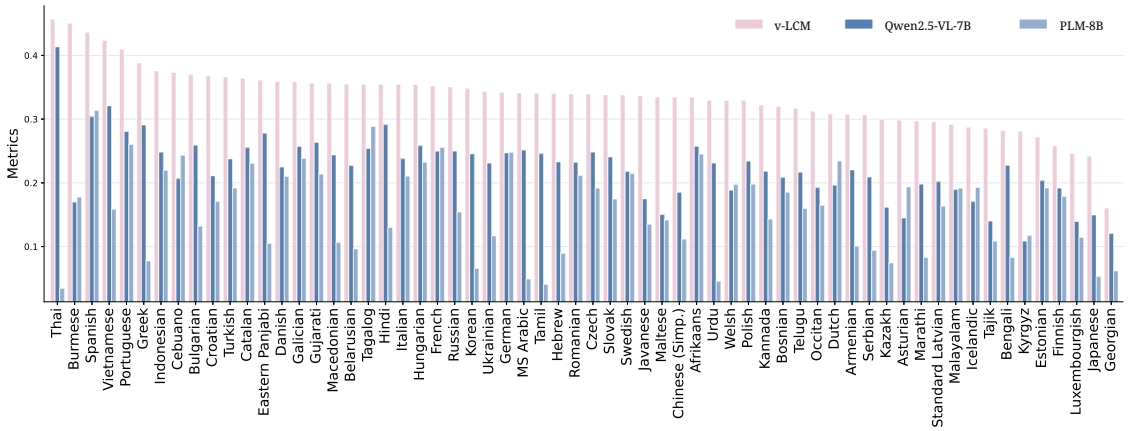


图6 针对MSRVTT-QA的M3IT 61种语言评估。

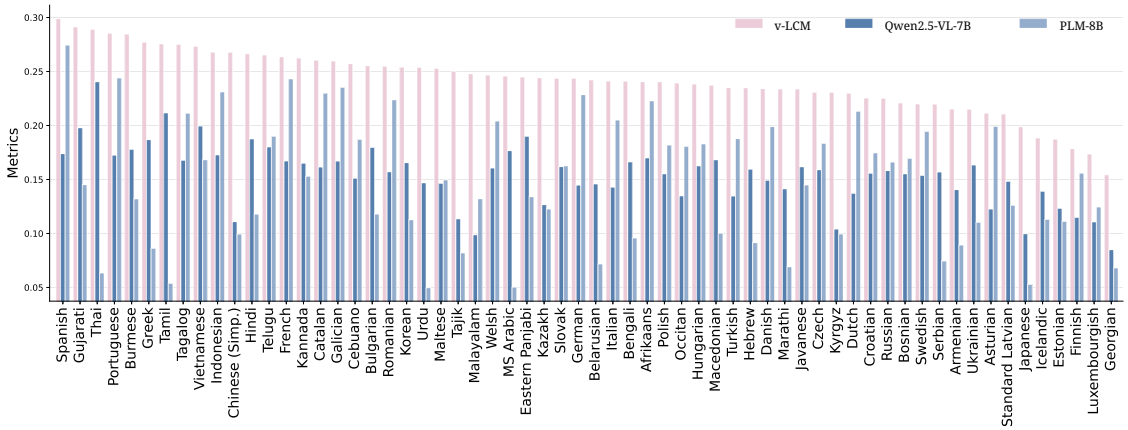


图7 针对OKVQA的M3IT 61种语言评估。

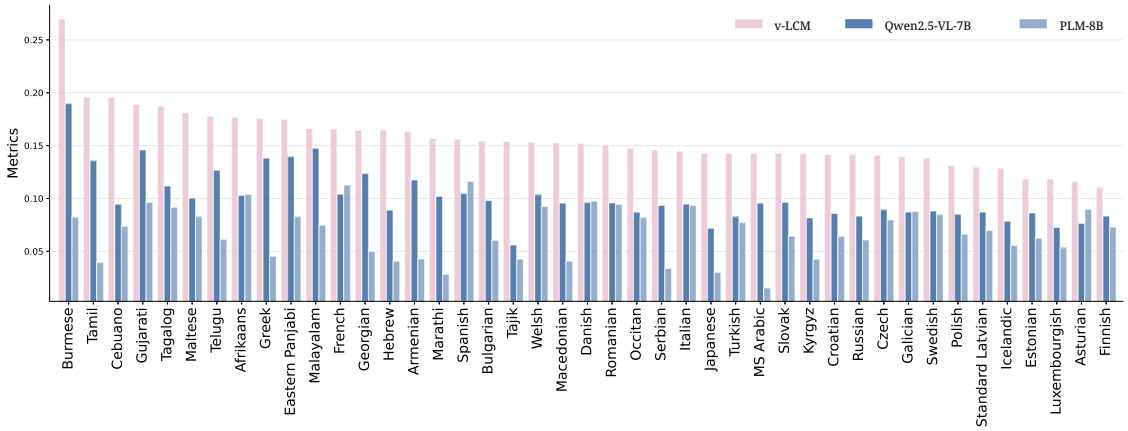


图8 M3IT在61种语言上的VIST评估。

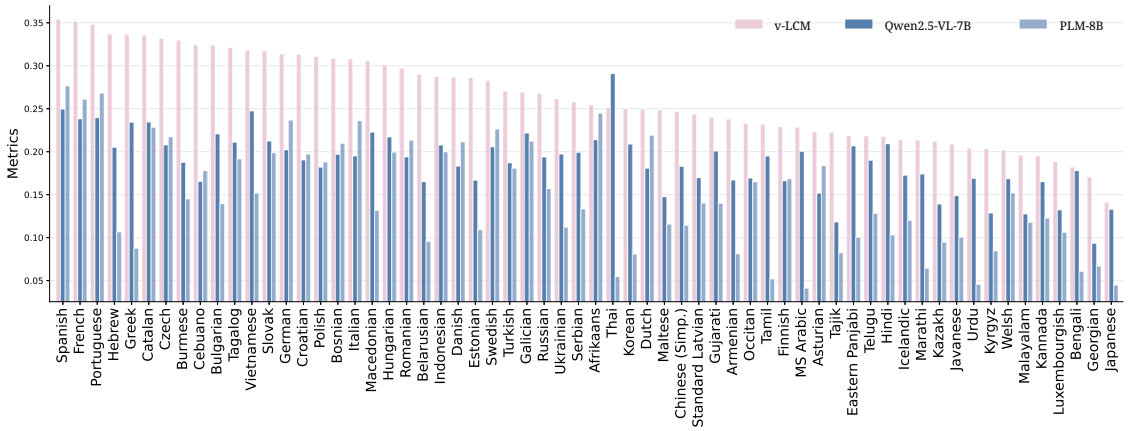


图9 M3IT在61种语言上的VQA-V2评估。

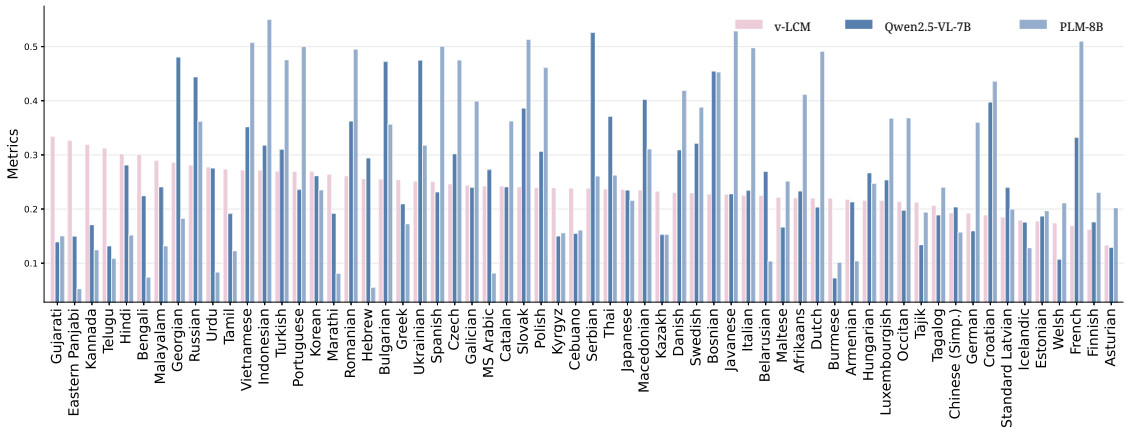
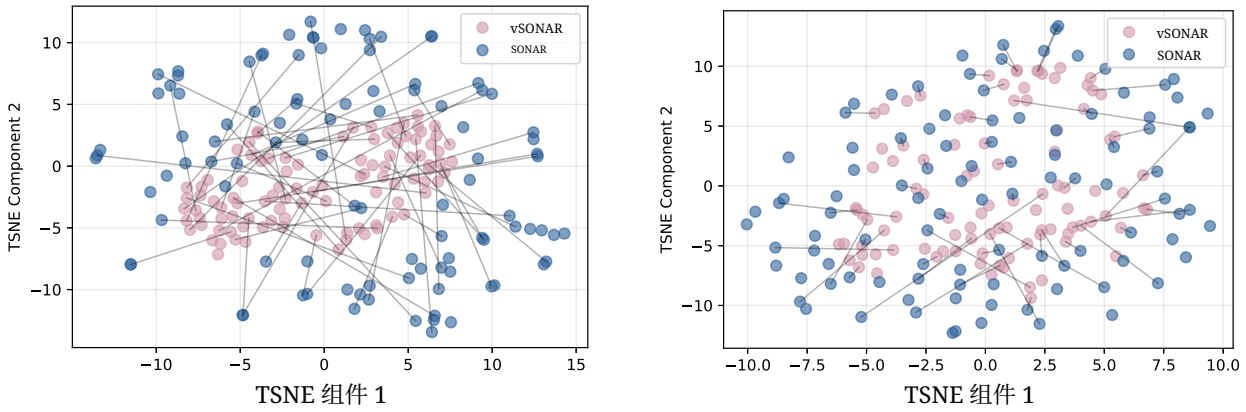


图10 M3IT在61种语言上的图像描述评估。



(a) 未对齐。 **(b)** 阶段 1 后。

(c) 阶段 2 后。 **(d)** 阶段 3 后。

图 11 使用 t-SNE 可视化课程各阶段后 Sonar 和 v-Sonar 嵌入的变化。v-Sonar 对视频进行编码，Sonar 对图像描述进行编码。我们从 PE-Video 测试集中随机抽取 200 个样本用于 t-SNE，并明确绘制 50 个样本中配对视频与图像描述之间的连线。

G v-Sonar 和 v-LCM 的跨模态漂移分析

在本节中，我们研究 v-Sonar 和 v-LCM 的解码过程中是否可能发生跨模态漂移。我们旨在回答这样一个问题：Sonar 解码器和 v-LCM 生成的语言，是否忠实地表达了 v-Sonar 嵌入空间中的同一语义区域。我们在 PE-Video 测试集上进行了三项分析，以实证证明 v-Sonar 和 v-LCM 在处理来自不同模态的嵌入时，解码过程中产生的语义漂移极小。

	语义相似度		**往返检索**			
	余弦	距离	R@1	R@5	R@10	MRR
真实值	0.666	0.197	87.00%	95.90%	97.10%	0.9084
Sonar 解码器	0.689	0.175	82.50%	97.00%	98.70%	0.8883
v-LCM	0.562	0.219	82.30%	96.70%	97.90%	0.8867

表 11 左：视频嵌入与真值字幕、Sonar 解码器生成的字幕以及 v-LCM 生成字幕之间的语义相似度和距离。右：往返检索消融研究的结果，其中我们使用真值字幕、Sonar 解码器和 v-LCM 生成的字幕来检索视频。

嵌入级语义保真度。对于每个视频嵌入 v，我们首先比较其与以下对象的相似度和距离：(i) 真值字幕嵌入 t_{gt}，(ii) SONAR解码器字幕嵌入 (t_{sonar}，以及 (iii) LCM

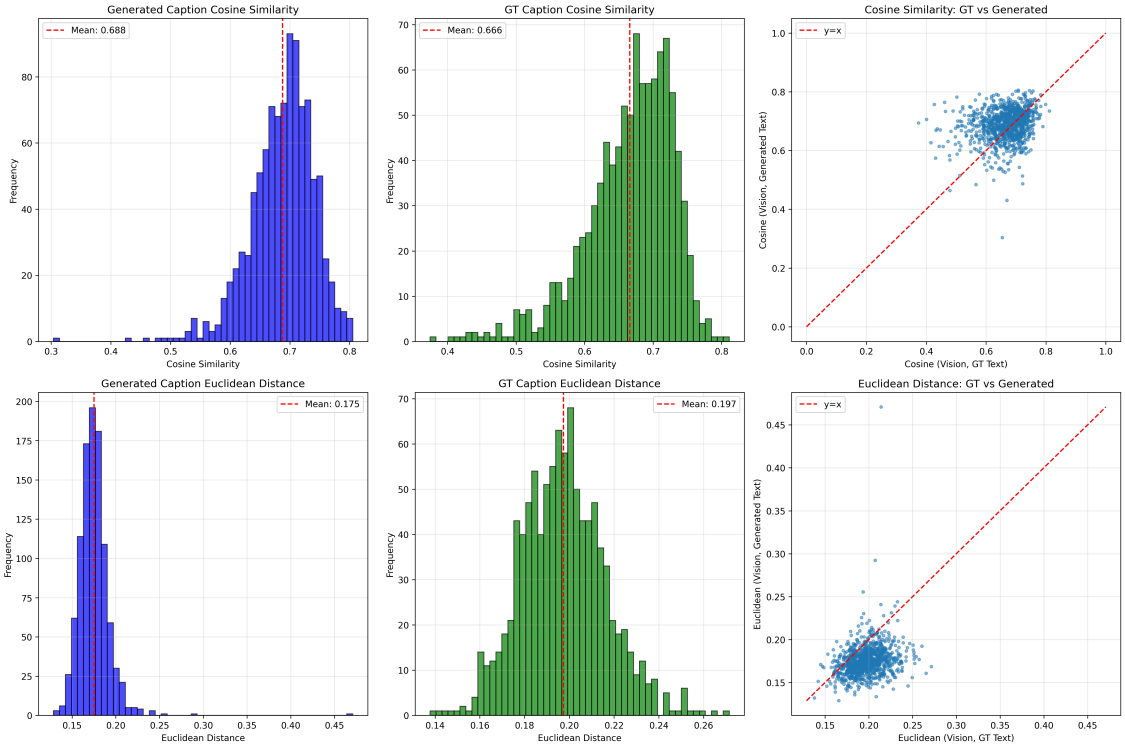


图 12 vSONAR 的跨模态语义漂移可视化。

字幕嵌入 t_{lcm}。此分析直接比较 Sonar 解码器和 v-LCM 生成的字幕与真值字幕。

如 表 11所示，我们发现，与真实值相比，SONAR 解码字幕的余弦相似度或距离几乎完全相同，这表明跨模态漂移可以忽略不计。vLCM 字幕的偏差略大；我们认为这是因为 vLCM 的指令遵循训练引入了风格化改写，而非语义漂移。

往返检索消融研究。 我们进一步针对往返检索开展消融研究，使用三组字幕（真实值、v-Sonar 解码器，以及 v-LCM）作为查询，在 PE-Video 中检索源视频。在 表 8中，我们发现，由 SONAR 或 LCM 解码的字幕能够以极高的准确率检索到正确视频。值得注意的是，LCM 的 R@1 仅比 SONAR 低 0.2%。如果跨模态漂移显著，检索准确率就会大幅下降；然而，准确率仍然很高，这证实了语义保留。

可视化跨模态漂移。最后，我们在v-Sonar和LCM 生成的字幕图 13和图 12中绘制视觉嵌入与真实图像描述嵌入之间的相似度，以及它们与 $y = x$ 直线的关系。Sonar 解码字幕的余弦相似度/距离与真实标注几乎相同（或略好），表明跨模态漂移可以忽略不计。vLCM 字幕的偏差稍大；我们认为这是因为 vLCM 的指令遵循训练引入了风格化改写，而非语义漂移（已在下一项实验中验证）。各数据点总体上也聚集在

H 定性案例

H.1 v-LCM 的图像定性案例

我们展示定性　　图像描述的定性案例见图 14，图像问答的定性案例见图1　　5.

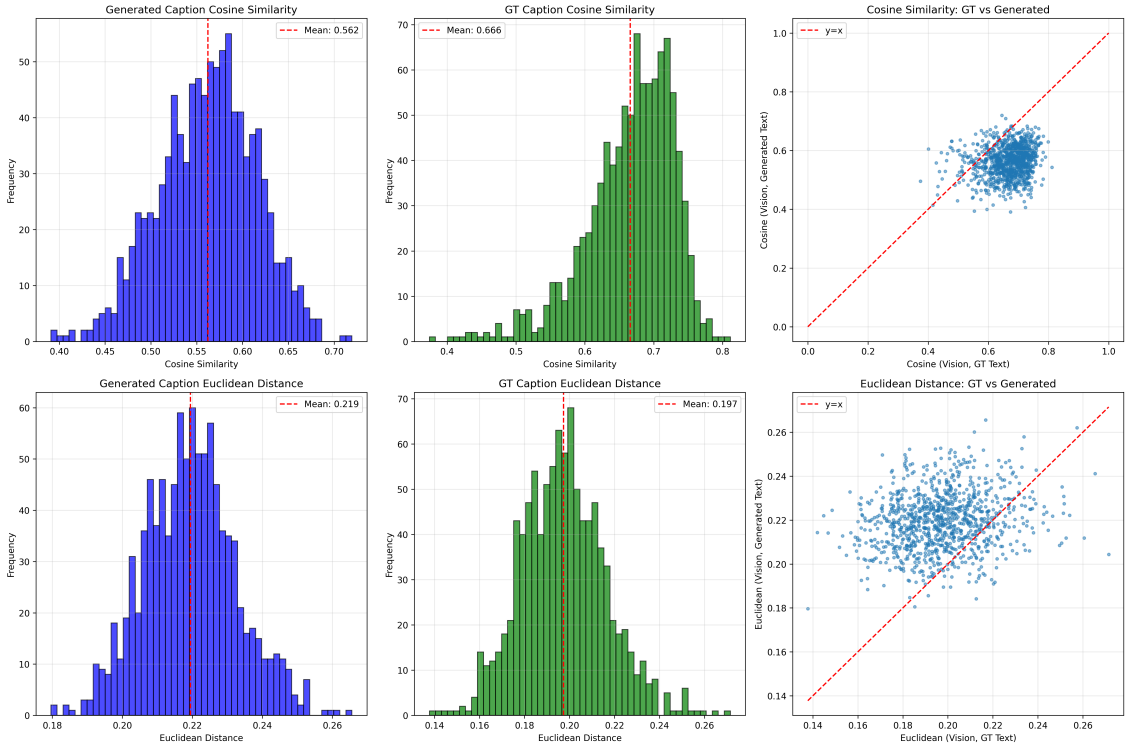


图 13 vLCM 对跨模态语义漂移的可视化。

H.2 v-LCM 的视频定性案例

视频描述和问答任务的定性案例见图16。

H.3 v-Sonar 的视频描述定性案例

除 LCM 的定性案例外，我们还展示了v-SONAR 与 PVD-Bench 和 DREAM-1K 中其他视觉语言模型的视频描述对比定性案例，见图17 至图20。

H.4 v-Sonar 的视频描述定性案例

我们展示了v-LCM 与其他视觉语言模型在图20中的视觉常识推理（VCR）定性案例。该定性示例表明，这些对齐编码器虽然使用语义级描述进行训练，仍能够实现布局定位并具备空间推理能力。

视觉输入	提示	输出
	撰写一段简短且信息丰富的描述 that 亮点 the 主要主体及正在发生的动作，出现在给定图像中。 <IMAGE>	**LCM**：“视频显示一名男子骑着一辆红色自行车骑着一辆红色摩托行驶在黑色山峰上，旁边有一辆摩托车和一块地毯，还有一名红色男子站立着在一座有桥和草的山后面，还有一座屋顶。” **v-LCM**：“一名男子骑着一辆红色摩托车在山上。 ”
	撰写一段描写性文字，让它将图像呈现出来通过细致描绘生活叙事，将各种视觉元素整合为统一的叙事。 <IMAGE>	**LCM**：“视觉始于明亮的视频建筑站在一栋铺有大片地面的建筑物前，而一群人们走在街道上。太阳照耀着建筑物时，建筑物开始沿着街道向下延伸。车辆逐渐靠近街道 , *with* 建筑物沿着街道缓缓向下延伸。 车辆正在驶向街道，另一侧有人站在人行道上。最后，画面以车辆驶上街道结束，而建筑物沿着街道滑动。” **v-LCM**：“一栋大型建筑正在一个拐角附近行走，街道。建筑物前有白色窗户。建筑物前有一根很大的灰色杆子。 有人们在杆子前的街道上行走。”

图 14 M3IT基准测试中图像描述任务的定性示例。每一行展示输入图像、给定的提示和生成的模型输出。

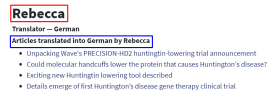
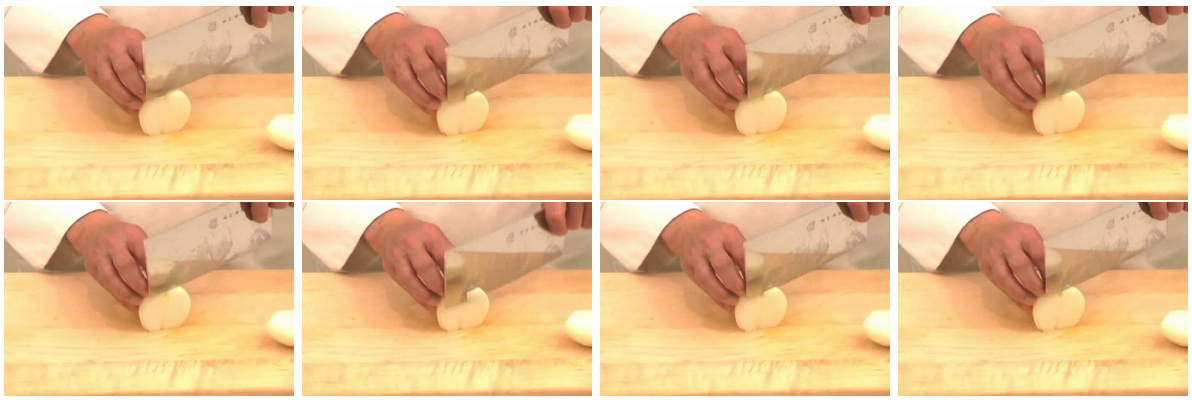
视觉输入	提示	输出
	仔细观察图像，识别 其中出现的命名实体，并运用你的知 识回答与.....相关的问题 那些实体。 <图像> 谁在 这个体育场比赛?	**LCM**：“在 这个体育场比赛的实体是红色体育场。” **v-LCM**： “在这座体育场比赛的 *Team at this stadium is Manchester United.*”
	分析图像，找出其中任何命名实 体，并针对问题给出恰当的回答 视觉开始。 <图像> 视觉结束。 在这个游戏中， “V”牌值多少分?	**LCM**：“提供的信息中并未说明 到底有多少 “*V*”牌值多少分。” **v-LCM**：“一 ’*V*’ 方块值 *4* 分。”
	仔细阅读问题及相关 图像/文本上下文，确定准确的答- 案，并从多项选择中选出， 请提供详细、合乎逻辑的解释， 以支持你的选择，并展示多重 庞大的推理过程。 <图像> 问题：以下哪项可能是戈尔 登的测试结果? [...] 选项： (A) 航天器是否在使用时受损—— 使用出口为 1 米、速度为 200 的降落伞 千米/小时。 (B) 出口为 1 米的降落伞有多稳定? 米通风口，时速200公里。 (C) 带有1米通风口的降落伞是否 会在400公里/小时时摆动过大。	**LCM**：“*(B)* 一个有多稳定 带有*1*米通风口的降落伞在*200* 公里/小时。” **v-LCM**：“答案是 *(B)*，稳定 的是一顶带有 *1* 米 漏气口的降落伞，速度为每小时 *200* 千米。”
	请参考以下内容回答该查询 图片及其有限实体来回答查询。 <IM- AGE> 译者的名字是什么?	**LCM**：“未提供译者的姓名 未提供。 ” **v-LCM**：“译者的姓 名是 *Rebecca Herting*。”

图 15 图像问答任务的定性示例。每一行展示输入图像、给定的提示以及生成的模型输出。



提示：你将看到一段视频，并被问及你所看到的内容。你的回答应是一句简短的句子，且符合视频内容。 <视频> 男子在切什么？**LCM：**一名男子正在用刀切洋葱。**v-LCM：**视频中的男子正在切洋葱。

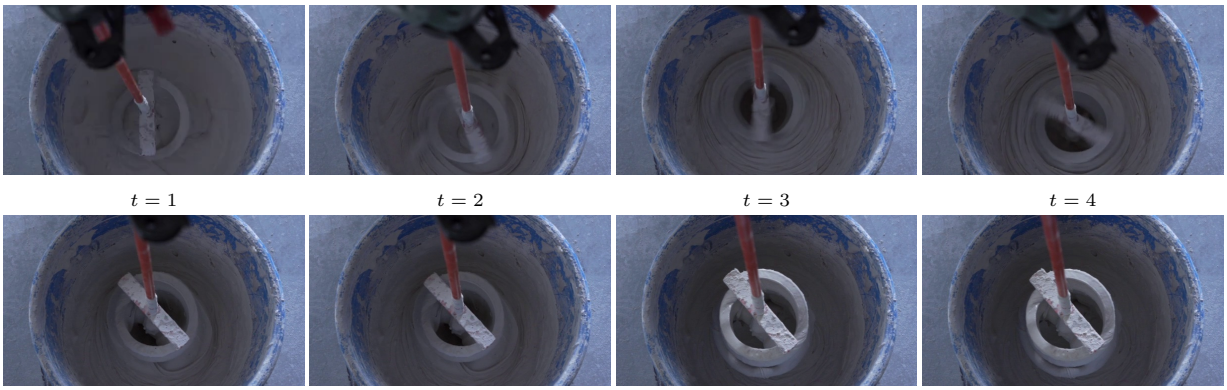


提示：请运用你的观察能力，借助一段简短的视频片段回答给定问题。视觉开始。 <视频> 视觉结束。所示的是哪种面包？**LCM：**所示面包的类型是三明治面包片。**v-LCM：**视频中所示的面包是贝果。



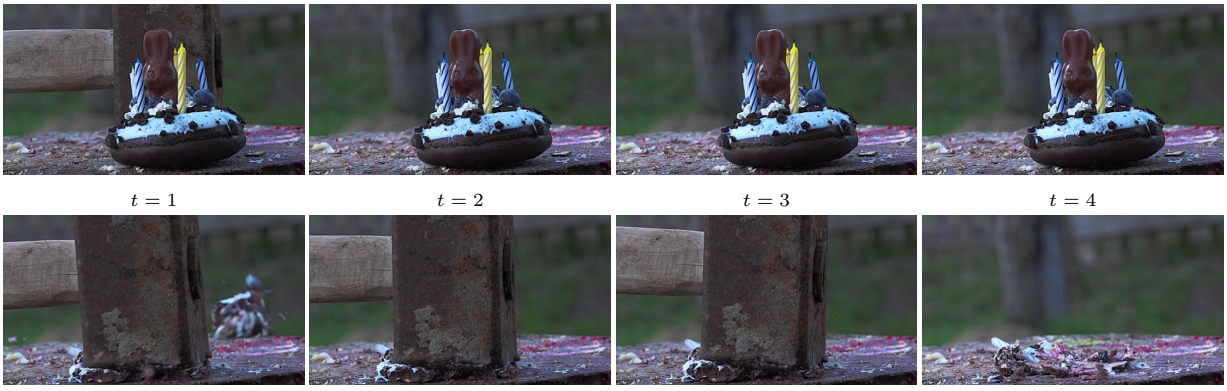
提示：观看视频，并根据视觉线索生成图像描述。 <视频>**LCM：** /回答结束。/
v-LCM：有人正在驾驶一辆红色汽车

图16 视频描述与问答任务的定性示例。帧从左到右、从上到下呈现。



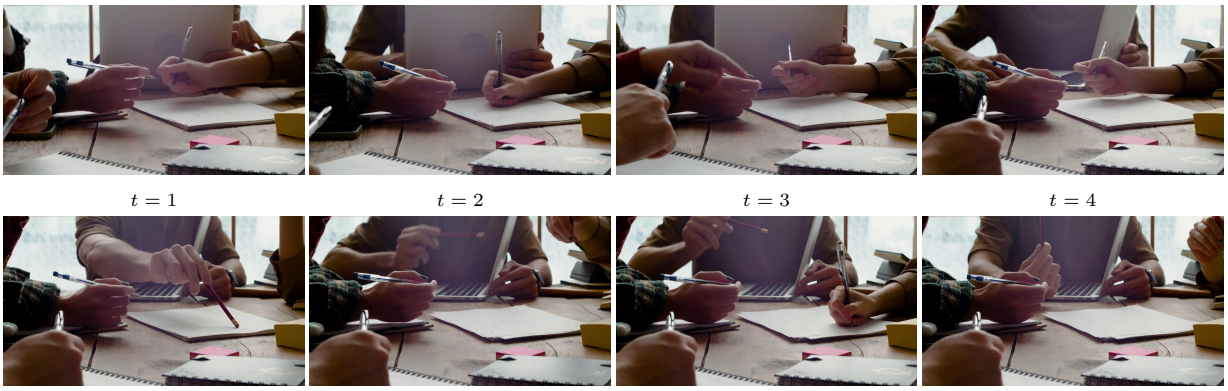
模型	视频字幕
v-Sonar（我们的方法）	视频中显示一个**蓝色搅拌机**正在圆形桶中搅拌**白色粉末**。桶正在移动，做旋转运动，并连接在一个蓝色圆柱体上。搅拌机正在缓慢地在中间移动，产生白色粉末。
PLM-3B	背景中有一个桶、一根棍子和一块灰色表面。搅拌机的声音清晰可闻。 /模糊；提到了音频/
Qwen2.5-VL-3B-Instruct	一个装满白色油漆的蓝色桶的特写镜头。一把**红柄画笔**蘸入油漆中，然后提起，留下了一道油漆痕迹。接着，画笔再次放入油漆中，并**重复数次**。摄像机保持静止不动。
InternVL2.5-1B	一只**黄色的手**伸入白色塑料桶中，抓起红色圆柱形物体，它是一种用于**浇筑水泥**的工具。

图17 PE-Video 中细粒度动作理解的定性比较。尽管当前最先进的视觉语言模型（通义千问-2.5、InternVL-2.5）会产生幻觉或图像描述错误（以**红色**标出），例如误认为是“画笔”或“黄色的手”，而 PLM 的描述仍然含糊，**SONAR** 却能准确捕捉搅拌机和蓝色圆柱形附件的工作机制（以**绿色**标出）。



模型	视频字幕
v-Sonar（我们的方法）	视频中有一个蛋糕，上面放着一根糖果棒。画面放大，聚焦在一只兔子和一根巧克力棒上，随后糖果棒从画面中被移走。视频最后，兔子躺在地面上，摄像机则推近蛋糕。
PLM-3B	背景中，**有一个棕色表面、一个棕色物体、一面棕色墙壁、一片绿色草地表面，还有一个棕色物体**。可以听到音乐声。 /描述模糊；提到了音频/
Qwen2.5-VL-3B-Instruct	**一个巧克力甜甜圈**上面涂有白色糖霜，点缀着蓝莓，放在木质表面上。这个甜甜圈上插着三根蜡烛，两根蓝色，一根黄色。蜡烛的光照亮了甜甜圈，并且**可以看到火焰**。甜甜圈周围散落着碎屑和其他残渣。背景画面模糊，但看起来是有草和树木的户外环境。
InternVL2.5-1B	可以看到桌子上的小蛋糕正被一个沉重的棍棒状物体猛烈击打，从中**烟开始升起**。蛋糕在桌子上被近距离展示，**周围环绕着纸**。在蛋糕被猛烈击打后，所用的锤子被放回桌子上，并继续击打大理石杯垫。

图18 PE-Video 中的定性比较。我们用**红色**标出错误。



模型	视频字幕
v-Sonar（我们的方法）	视频中，一群人围坐在桌子旁，手上拿着笔记本。 人们在纸上书写和绘画，有些人拿着铅笔，其他人则 拿着笔记本。桌子位于前景，人们在桌子周围走动， 他们正在纸上写字。
PLM-3B	视频中，一群人围坐在桌子旁，正在进行一项协作 活动。这些人手里拿着钢笔和铅笔，桌上摆着各种物体， 包括笔记本电脑、笔记本和便利贴。场景看起来像是一场会议 或头脑风暴会议，参与者正齐心协力地创作某种作品。 视频采用慢动作拍摄 , 为场景增添了宁静与专注的氛围。总体而言 , the 视频传达出协作、创造力和生产力的氛围。
Qwen2.5-VL-3B-Instruct	视频展示了一群人围坐在桌子旁，共同开展一个项目。 他们使用笔记本电脑，并在纸上记录笔记。氛围充满协作精神， 每个人都专注于当前任务并积极参与其中。灯光温暖而舒适， 营造出舒适宜人的环境。人们专注于工作， 整体氛围积极而高效。
InternVL2.5-1B	在这段视频中，一群人协作，共同参与设计过程，围绕一张 桌子展 开工作。

图19 DREAM-1K 中的定性比较。



$t = 1$ $t = 2$ $t = 3$ $t = 4$



$t = 5$ $t = 6$ $t = 7$ $t = 8$

模型	视频字幕
v-Sonar（我们的方法）	视频中，一群男子穿着白衬衫跳舞。一名男子站在田野里，手里拿着一根绳子，身后有一辆红色卡车。男子正在拉绳子，而卡车正朝田野驶去。
PLM-3B	一名穿着白色T恤和黑色裤子的男子站在草地上，然后开始跑起来并踢白色绳子。另一名穿着红色T恤和黑色裤子的男子站在草地上，然后开始跑起来并踢白色绳子。第三名穿着白色T恤和黑色裤子的男子站在草地上，然后开始跑起来并踢白色绳子。在背景中，有一片草地，树木、一辆红色车辆、一根白色绳子和天空。可以听到人们的声音。
Qwen2.5-VL-3B-Instruct	视频中，一名男子站在田野里，穿着白色T恤和黑色裤子。他正握着一根绳子，似乎在玩拔河游戏。背景中有树木和晴朗的蓝天。
InternVL2.5-1B	一名穿着白衬衫和黑腰带的表演者在公园里踩着一根长绳保持平衡，旁边有一辆红色背景中有一辆车和一棵茂盛的红树。表演者改变姿势，最终从绳子上摔下来并落在地面上，随后迅速起身，准备迎接下一次尝试或战斗。

Dream-1k 中的定性比较（图20）。



模型	
提示	给定一张图像和一些标定的物体，请回答问题并解释答案为何 正确。 物体：第 0 个物体是红色框中的人物。第 1 个物体是 蓝色框中的人物。第 2 个物体是绿色框中的人物。第 3 个物体是人物 在粉色方框中。第 4 个物体是浅蓝色方框中的狗。 问题：第 3 个物体正在和第 4 个物体做什么？
PLM-8B	年幼的儿童喜欢带着宠物到户外散步。
Qwen2.5-VL-8B-Instruct	答案 (A) 第 3 个物体正在带着第 4 个物体到户外散步。理由：第 3 个 个物体从体型和姿势来看似乎是一个儿童，而第 4 个物体是一只狗。 图像的背景表明这是一个户外环境，有一个儿童和一只狗，这与 儿童带宠物散步的情景相符。狗通常喜欢散步，而儿童带宠物散步也很常见， 这也是他们日常生活或玩耍的一部分。 其他选项与图像背景不太相符；没有迹象表明有人在调情、 看电脑或布置什么东西。
InternVL2.5-8B	第 3 个物体正在设置第 4 个物体。
v-LCM	第 3 个物体正带着第 4 个物体在外面散步。因为小孩子喜欢带 他们的宠物散步。

图21 视觉常识推理（VCR）中的定性比较，这需要结合常识进行落地和空间推理。