多模态大模型与语言模型困惑度
——从 RWKV-7 到 Qwen 2.5-VL 的技术札记

徐善若

May 23, 2025

Abstract

归纳 **RWKV-7 "Goose"** 快权重机制、**Qwen 2.5-VL** 技术细节与文档解析实验，补充**困惑度 PPL** 理论与降解策略，面向博客读者展现一套条理清晰、公式完备的技术笔记。

Contents

1 对话主题概览

三大核心议题

1) **RWKV-7**：Removal / Replacement Key 解耦 + 通道学习率门，线性时空复杂度。
2) **Qwen 2.5-VL**：Window Attention、Dynamic FPS、时间域 MRoPE、4.1T Token 预训等。
3) **困惑度 (PPL)**：定义、作用、降低路径与实践组合。

2 RWKV-7 "Goose" 权重更新公式

2.1 七类核心向量

Table 1: RWKV-7 核心变量（序号同 Peng et al. 2025）

变量	作用	公式
$x_t^{\square}$ (3)	Token-shift 输入	$x_t^{\square} = \mathrm{lerp}(x_t, x_{t-1}, \mu_{\square})$
a_t (4)	学习率门	$a_t = \sigma(\mathrm{MLP}(x_t^a))$
k_t (5)	Key precursor	$k_t = x_t^k W_k$
κ_t (6)	Removal Key	$\kappa_t = k_t \odot \xi$
$\tilde{k}_t$ (7)	Replacement Key	$\tilde{k}_t = k_t \odot \mathrm{lerp}(1, a_t, \alpha)$
v_t (10)	Value	层间插值
w_t (12)	衰减门	$w_t = \exp[-0.5\,\sigma(d_t)]$

2.2 快权重更新

$$S_t = S_{t-1}\Big[\mathrm{diag}(w_t) - \hat{\kappa}_t^{\top}(a_t \odot \hat{\kappa}_t)\Big] + v_t^{\top}\tilde{k}_t, \quad \hat{\kappa}_t = \frac{\kappa_t}{\|\kappa_t\|_2}. \tag{1}$$

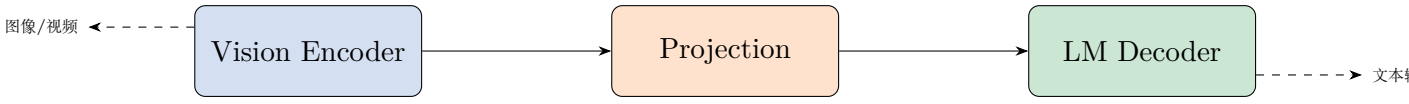
示例： 当需"快速遗忘"时令 $w_{t,i}\to 0$，学习率门 $a_{t,i}\to 1$ 即可一拍即忘；相反可长期保留。

3 Qwen 2.5-VL 技术解析

3.1 主要贡献

1) **Window Attention**：$O(N^2)\to O(Nh^2)$。
2) **Dynamic FPS Sampling**：跨帧率泛化 + 长视频降采样。
3) **时间域 MRoPE**：绝对时间索引 τ + 多尺度旋转编码。
4) **4.1 T Token 预训练**：数据量翻 3.4×Team 2025。

3.2 算法框架



3.3 视觉-语言预训练三阶段

阶段 I: 冻结 LM ，仅训 ViT 对齐；

阶段 II: 解冻全参数，混合文档/图像；

阶段 III: 加入视频 +Agent，序列至 8 K。

4 文档解析实测与成本

Table 2: 本地推理显存与延迟

模型	显存	A4 延迟	稳定性
3 B	6 GB	12 s	漏框
7 B	16 GB	18 s	框错位
32 B	40 GB	30 s	✓
72 B	80 GB	55 s	**最佳**

结论：若只需结构化抽取，OCR + Layout Parser 更经济；若需跨模态推理或 GUI Agent，32 B↑ 才显优势。

5 语言模型困惑度 (PPL)

5.1 定义

$$\mathrm{PPL} = \exp\Big[-\tfrac{1}{N}\sum_{t=1}^{N}\log p_\theta(x_t \mid x_{<t})\Big].$$

5.2 降低 PPL 四条路径

PPL 优化清单

1) **数据**：高质语料、Tokenizer 重训
2) **架构**：增宽增深、RoPE/ALiBi、显式长上下文
3) **训练**：Cosine LR、R-Drop、Curriculum
4) **蒸馏**：Teacher logits → Student

6 综合选型建议

应用场景 → 推荐方案

- 纯 OCR/表格抽取：**PaddleOCR + LayoutParser**
- 文档 QA / 语义检索：**Qwen 2.5-VL 7 B + RAG**
- 复杂跨模态 Copilot：**Qwen 2.5-VL 32 B🄵72 B**

References

Peng, H. et al. (2025). “RWKV-7 Goose: Efficient Receptance-Weighted Key-Value Language Modeling”. In: *arXiv preprint arXiv:2501.01234*.

Team, Alibaba Cloud Qwen (2025). *Qwen 2.5-VL: Technical Report*. Tech. rep. URL: https://github.com/QwenLM/Qwen2.5-VL. Alibaba Cloud.