```
┌─────────────────────────────────┐                    ┌──────────────────────┐
│           Model                 │                    │  Tokenized Captions  │
│       Clip (frozen) +           │                    │    [B, target_len]   │
│  Proj Linear Layer + Proj Model │                    └──────────────────────┘
│      (trainable)                │                               │
│       + phi2 (trainable)        │                               ▼
└─────────────────────────────────┘                    ┌──────────────────────┐
                 │                                      │  Tokenized Captions  │
                 ▼                                      │    [B, target_len]   │
┌─────────────────────────────────┐                    └──────────────────────┘
│            Inp_emb              │                               │
│ [B, (49+4+qn_len+4+tgt_len), 2560]│                             │
│       Eg : [B, 107, 2560]       │                               │
└─────────────────────────────────┘                               │
                 │                                                 │
                 ▼                                                 │
        ┌────────────────────┐                                     │
        │  Phi2 model.forward()│                                   │
        │   [B, 107, 50297]   │                                    │
        └────────────────────┘                                     │
                 │                                                 │
                 ▼                                                 ▼
┌─────────────────────────────────┐                    ┌──────────────────────┐
│    Extracts last word embed     │                    │                      │
│      [B, ans_pos:, 50297]       │───────────────────▶│  Cross Entropy Loss  │
│ where ans_pos = 49 + 4 + qn_len + 3│                 │                      │
└─────────────────────────────────┘                    └──────────────────────┘
                                                                   │
```

Backprop loss to update proj
linear layer, proj model and
phi2 adapter weights