第二届山东省人工智能大会Spotlight论文汇报

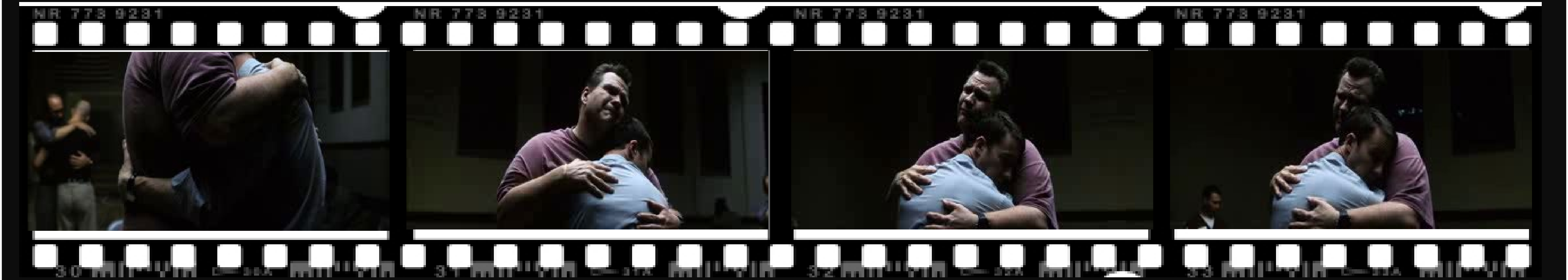
Recursive Narrative Alignment for Movie Narrating
（循环情节配准的电影叙述）

韩忠义[1,3], Hongbo WU[2], 魏本征[3,*], 尹义龙[1,*], *and* Shuo LI[4]

1. 山东大学 软件学院
2. St. Lawrence College, Kingston, Canada
3. 山东中医药大学 医学人工智能研究中心
4. Digital Imaging Group of London, Western University

Definition

Movie narrating aims to capture not only the subject matter but also the emotive essence of film frames into a story.



He finally asks, staring at me, his legs still wrapped with tears. He gives me get away from the situation with his face out. Allow repeating my question furthermore and find ways to comfort me. Lay a while, then, by the end.

Significance

- Human-like expressivity of movie shots (i.e. sets of film frames) into a cohesive and coherent narrative
- Human-like understanding of images
- Opens up numerous new applications
- It is an interdisciplinary field spanning computer vision, natural language processing, and philosophy

Challenges

- **The latent relatedness**

  For instance, in the sentence “The sky was illuminated with a brilliance of orange hues”, it is imperative to correlate the key words “brilliance” and “orange hues” to the imagery of a picturesque sunset.

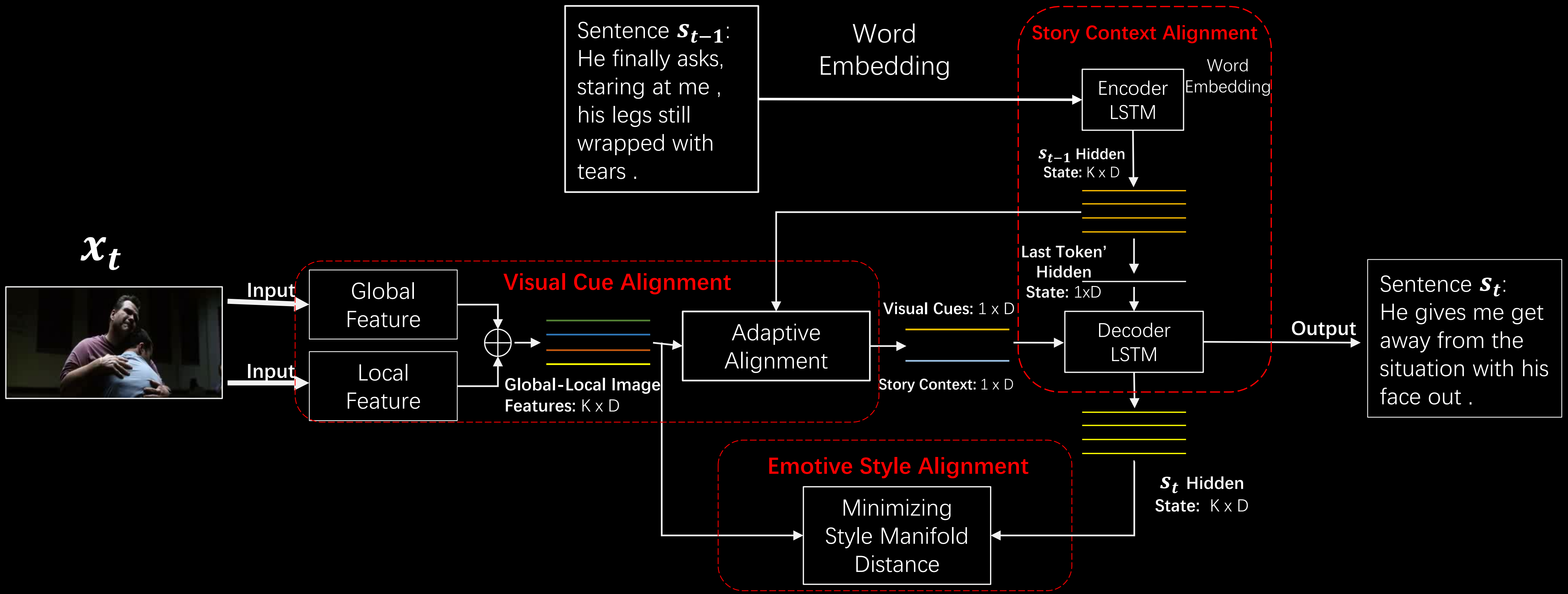
- **The weak consistency**

  Movie narrating requires film frames be described with cohesive sentences to reason about the semantic content of images.

- **The emotive state conflict**

  Movie narrating requires consistency in the emotive state between the frames and narrative. For example, frames depicting the brilliance of sunset should correspond to positive sentiment whereas images of war should be described with a negative connotation.

Recursive Narrative Alignment Framework



x_t

Input

Input

Visual Cue Alignment

Global Feature

Local Feature

Global-Local Image **Features:** K x D

Adaptive Alignment

Sentence s_{t-1}: He finally asks, staring at me , his legs still wrapped with tears .

Word Embedding

Story Context Alignment

Encoder LSTM

Word Embedding

s_{t-1} **Hidden State:** K x D

Last Token' Hidden State: 1xD

Visual Cues: 1 x D

Story Context: 1 x D

Decoder LSTM

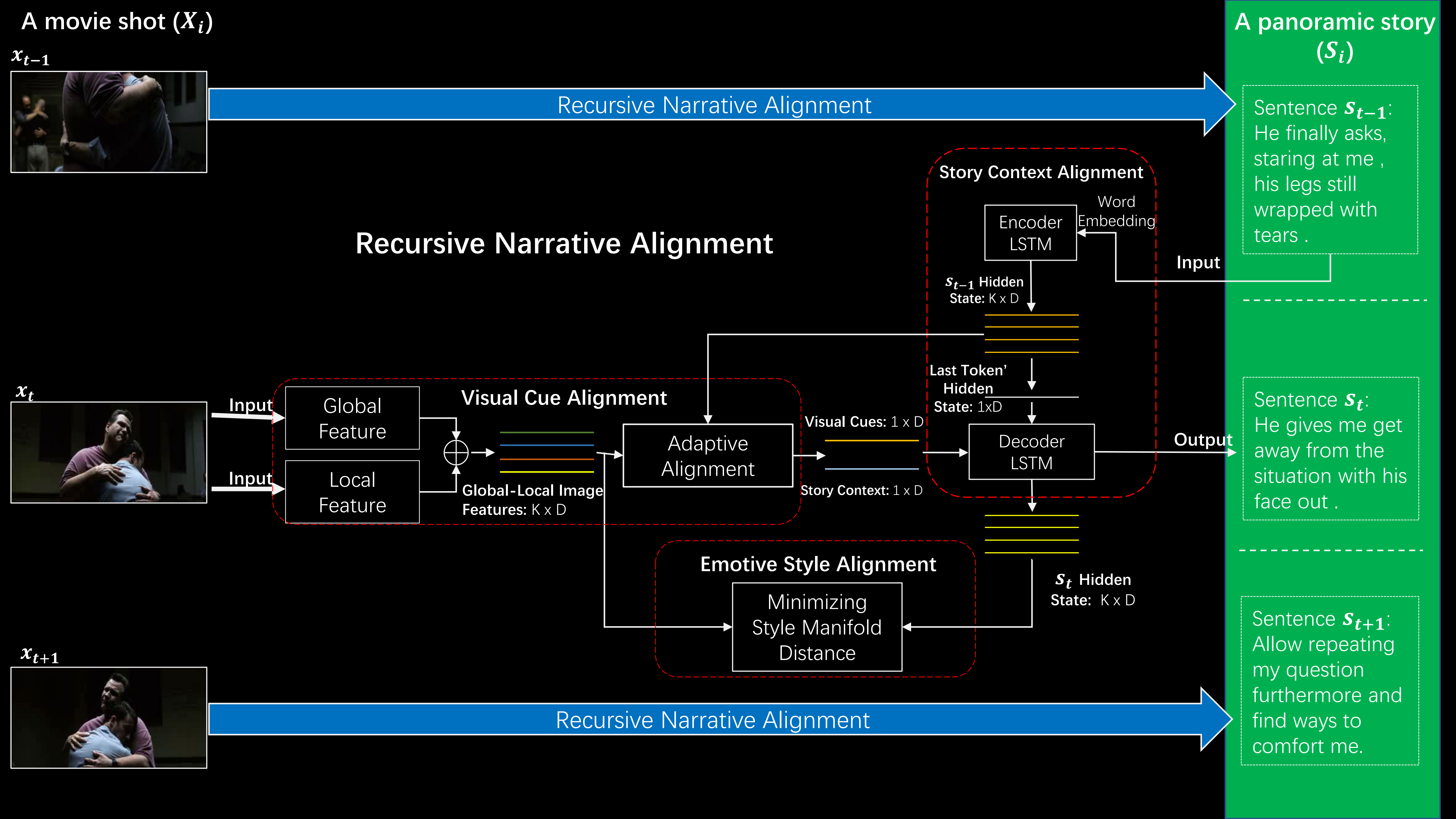
Output

Sentence s_t: He gives me get away from the situation with his face out .

s_t **Hidden State:** K x D

Emotive Style Alignment

Minimizing Style Manifold Distance



A movie shot (X_i)

x_{t-1}

Recursive Narrative Alignment

x_t

Input

Input

Visual Cue Alignment

Global Feature

Local Feature

Global-Local Image Features: K x D

Adaptive Alignment

Visual Cues: 1 x D

Story Context: 1 x D

Recursive Narrative Alignment

Story Context Alignment

Encoder LSTM

Word Embedding

Input

s_{t-1} Hidden State: K x D

Last Token' Hidden State: 1xD

Decoder LSTM

Output

s_t Hidden State: K x D

Emotive Style Alignment

Minimizing Style Manifold Distance

x_{t+1}

Recursive Narrative Alignment

A panoramic story (S_i)

Sentence s_{t-1}: He finally asks, staring at me , his legs still wrapped with tears .

Sentence s_t: He gives me get away from the situation with his face out .

Sentence s_{t+1}: Allow repeating my question furthermore and find ways to comfort me.

Recursive Narrative Alignment

- Visual Cue Alignment
  - It uses a semantic attention mechanism to adaptively align visual cues with keywords for a better frame-narrative coherence.
  - The attention mechanism comprises global and local information from each frame. It is adaptively combined into objective keywords and subjective emotive words.
- Story Context Alignment
  - The framework then recursively applies the contextual expression of previous frames into the current frame to improve the cohesion of the narrative.
- Emotive Style Alignment
  - The emotive conflict between frame and story is resolved by our newly-designed style regularize, which minimizes the style manifold distance between the file frame and story.

Visual Cue Alignment

$$V_t^{GL} = f(W^G V_t^G + b^G) + f(W^L V_t^L + b^L),$$

$$C_t^V = \sum_{k=1}^{K} \alpha_t V_{t_k}^{GL}, \qquad C_t^T = \sum_{k=1}^{K} \alpha_t H_{(t-1)\quad k},$$

$$\alpha_t = \frac{exp(z_{t_k})}{\sum_{k=1}^{K} \quad exp(z_{t_k})}, \forall t \in (1, \cdots, T), \quad \text{in which } z_{t_k} = W^z \tan \mathrm{h}(W^V V_{t_k}^{GL} + W^H H_{(t-1))\quad k}).$$

Story Context Alignment

- Encoder LSTM

$$h_k = LSTM^{enc}(w_1, \cdots, w_{k-1}).$$

- Decoder LSTM

$$p(y_k) = LSTM^{dec}(y_0, y_1, \cdots, y_{k-1}; C^T; C^V),$$

Emotive Style Alignment

$$\mathcal{L}_{style} = ||\mathcal{G}_V - \mathcal{G}_S||^2,$$

$$\mathcal{G}_V(V^{GL}) = V^{GL}(V^{GL})^T, \mathcal{G}_S(H) = HH^T.$$

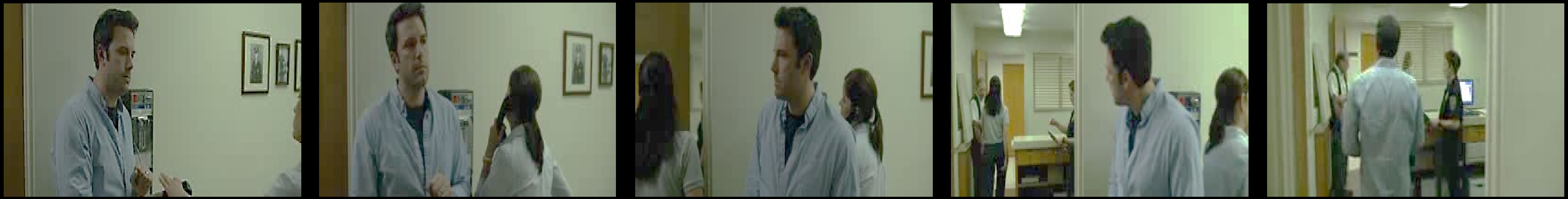
Hybrid Training Strategy

$$\mathcal{L}(x_t; s_t, s_{t-1}) = \mathcal{L}_{text}(x_t; s_{t-1}) + \phi\mathcal{L}_{style}(x_t; s_t),$$

$$\theta^* = \text{argmin}_\theta - \sum_{t=1}^{T} \log p(s_t | LSTM(C_t^V(x_t), C_t^T(s_{t-1})); \theta) + ||\mathcal{G}_V - \mathcal{G}_S||^2,$$

Results

Movie narratives generated using the Gone Girl film of the Movie-Book Data



RNA: He finally asks , as if he'd been putting off the whole week. He had some hard to be any of this way to pay the other. He looks up in a low , young , while we don't see. The door was pressed against the wall , he looked at the room. “She has a good job ”, he said uplifting his arms.

RNA-noattention: And she'd got it off a man's his *** stare , either in. To his moment going right with him pulled at all with laughter? Nobody bang bang used over 22 about - or slow down wall just chuckled. “Warn counter he will never strikes at the explanation room so'' he said. Swings please , huh ? open light beautiful *** by this other cup yet.

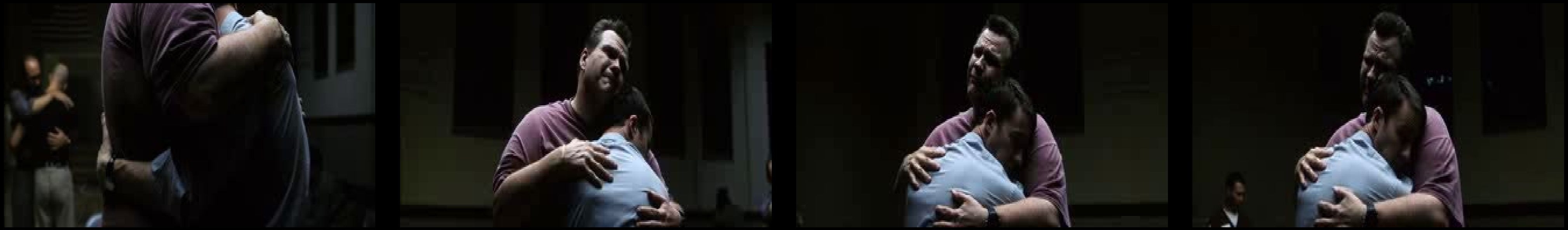
RNA-nocontext: Them months he says why wanted stay over , sat grinning how followed *****. Not very enough given quickly opened the money card back in and other eye. `` annoyed in hell shook roof here or ? Stock Her standing an sort of ice else to know both of here who here. People good lot in on an side was about more to keep years ?

RNA-nostyle: Of the magazines be road ?'' upward a good fifth white upward. Manager Feet , and he looked at the boy said , a good fathern’t. `` , he looked up at the boy , and then he said. The road and then , and then looked at the road the boy said. And stood looking at the road , and then he said the road road.

Reference: He asks the girl if he should wait another whole week. He is so depressed that he want to say anymore. He looks up a young girl, while we cannot see. He looked at the room through the door. “She has a good job and please to find her”, he said throwing up his arms in despair.

Results

Movie narratives generated using the Fight Club film of the Movie-Book Data



He finally asks , staring at me , her legs still wrapped in tears. Doing the riverbank in a faded and lay a while , then , by the end. she gives me get away from the situation with his face out. Allowed to rephrase my question and found ways to spend out of her face.

Thanks