Visual Storytelling

Ting-Hao (Kenneth) Huang

Penn State University

Guest Lecture for CIS 700 - Interactive Fiction and Text Generation

April 19, 2022

Flickr CC Search with keyword “life”, April 17, 2022.

About **Me**

Ting-Hao 'Kenneth' Huang 黃挺豪

Assistant Professor
College of Information Sciences and Technology (IST)
Pennsylvania State University (University Park)
Affiliation: Center for Social Data Analytics (C-SoDA)
Affiliation: Center for Socially Responsible Artificial Intelligence (CSRAI)

We are hiring at all levels! Come work with us!

I combine AI with crowdsourcing to create systems that are usable, robust, and intelligent.

- **Office:** E357 Westgate Building
- **Email:** txh710@psu.edu
- Google Scholar
- Curriculum Vitae (CV)
- Twitter: @windx0303
- Website: KennethHuang.cc
- ORCID iD: 0000-0001-7021-4627

Outline

- The **birth** of the Visual Storytelling (VIST) task
- The **evolvement** of VIST technologies
- The **applications** of VIST

Outline

- The **birth** of the Visual Storytelling (VIST) task
- The **evolvement** of VIST technologies
- The **applications** of VIST

Huang*, T. H., Ferraro*, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., ... & Mitchell, M. (2016, June). **Visual storytelling.** In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1233-1239).

2013-2016: Vision-and-Language Explosion

VQA: Visual Question Answering

Stanislaw Antol*[1] Aishwarya Agrawal*[1] Jiasen Lu[1] Margaret Mitchell[2]
Dhruv Batra[1] C. Lawrence Zitnick[2] Devi Parikh[1]
[1]Virginia Tech [2]Microsoft Research
[1]{santol, aish, jiasenlu, dbatra, parikh}@vt.edu [2]{memitc, larryz}@microsoft.com

VQA: 2015

Microsoft COCO Captions: Data Collection and Evaluation Server

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam
Saurabh Gupta, Piotr Dollár, C. Lawrence Zitnick

Abstract—In this paper we describe the Microsoft COCO Caption dataset and evaluation server. When completed, the dataset will contain over one and a half million captions describing over 330,000 images. For the training and validation images, five independent human generated captions will be provided. To ensure consistency in evaluation of automatic caption generation algorithms, an evaluation server is used. The evaluation server receives candidate captions and scores them using several popular metrics, including BLEU, METEOR, ROUGE and CIDEr. Instructions for using the evaluation server are provided.

COCO Caption: 2015

Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations

Ranjay Krishna[1] · Yuke Zhu[1] · Oliver Groth[2] · Justin Johnson[1] · Kenji Hata[1] · Joshua Kravitz[1] · Stephanie Chen[1] · Yannis Kalantidis[3] · Li-Jia Li[4] · David A. Shamma[5] · Michael S. Bernstein[1] · Li Fei-Fei[1]

Visual Genome: 2016

Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models

Bryan A. Plummer[†] Liwei Wang[†] Chris M. Cervantes[†] Juan C. Caicedo*
Julia Hockenmaier[†] Svetlana Lazebnik[†]
[†]Univ. of Illinois at Urbana-Champaign *Fundación Univ. Konrad Lorenz
[bplumme2,lwang97,ccervan2,juliahmr,slazebni]@illinois.edu
juanc.caicedor@konradlorenz.edu.co

Flickr30k Entities: 2015

Ferraro, et al. (2015, September). **A Survey of Current Datasets for Vision and Language Research.** In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 207-213).

Direct, Literal Descriptions

		
A group of people that are sitting next to each other.	Adult male wearing sunglasses lying down on black pavement.	The sun is setting over the ocean and mountains.

Evaluative, Figurative Language

		
A group of people that are sitting next to each other.	Adult male wearing sunglasses lying down on black pavement.	The sun is setting over the ocean and mountains.
Having a good time bonding and talking.	[M] got exhausted by the heat.	Sky illuminated with a brilliance of gold and orange hues.

"Sitting in a Room" vs. "Bonding"

uralistic interactions. There is a significant difference, yet unexplored, between remarking that a visual scene shows "sitting in a room" – typical of most image captioning work – and that the same visual scene shows "bonding". The latter description is grounded in the visual signal, yet it brings to bear information about social relations and emotions that can be additionally inferred in context (Figure 1). Visually-grounded stories facilitate more evaluative and figurative language than has previously been seen in vision-to-language research: If a system can recognize that colleagues look *bored*, it can remark and act on this information directly.

Concrete vs. Abstract Terms

		Size(k)		Language								Vision	
	Dataset	Img	Txt	Frazier	Yngve	Vocab Size (k)	Sent Len.	#Conc	#Abs	%Abs	Ppl	(A)bs/ (R)eal	BB
Balanced	Brown	-	52	18.5	77.21	47.7	20.82	40411	7264	15.24%	194	-	-
User-Gen	SBU	1000	1000	9.70	26.03	254.6	13.29	243940	9495	3.74%	346	R	-
	Deja	4000	180	4.13	4.71	38.3	4.10	34581	3714	9.70%	184	R	-
Crowd-sourced	Pascal	1	5	8.03	25.78	3.4	10.78	2741	591	17.74%	123	R	-
	Flickr30K	32	159	9.50	27.00	20.3	12.98	17214	3033	14.98%	118	R	-
	COCO	328	2500	9.11	24.92	24.9	11.30	21607	3218	12.96%	121	R	Y
	Clipart	10	60	6.50	12.24	2.7	7.18	2202	482	17.96%	126	A	Y
Video	VDC	2	85	6.71	15.18	13.6	7.97	11795	1741	12.86%	148	R	-
Beyond	VQA	10	330	6.50	14.00	6.2	7.58	5019	1194	19.22%	113	A/R	-
	CQA	123	118	9.69	11.18	10.2	8.65	8501	1636	16.14%	199	R	Y
	VML	11	360	6.83	12.72	11.2	7.56	9220	1914	17.19%	110	R	Y

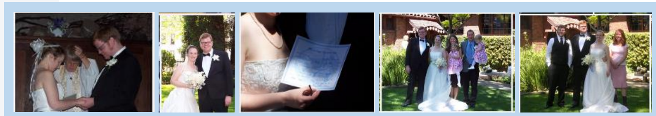
Table 1: Summary of statistics and quality metrics of a sample set of major datasets. For Brown, we report Frazier and Yngve scores on automatically acquired parses, but we also compute them for the 24K sentences with gold parses: in this setting, the mean Frazier score is 15.26 while the mean Yngve score is 58.48.

- **Abstract terms** = Ideas or concepts, e.g., ‘love’ or ‘think’.
- **Concrete terms** = All the objects or events.

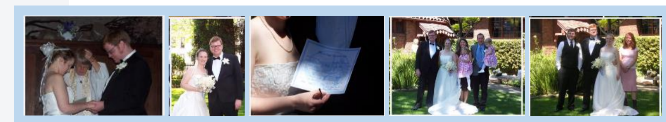
How to collect **figurative text** for images?

- **Design Constraints**
  - **Vision-and-language** data (multiple images → a story)
  - Real-world **human activities**
  - Temporal relations
  - Need **multiple references** for each instance (learned from MT)

A photo album of an event



Form a photo sequence



+

[A Short Story About the Photo Seq]

Write a story for it

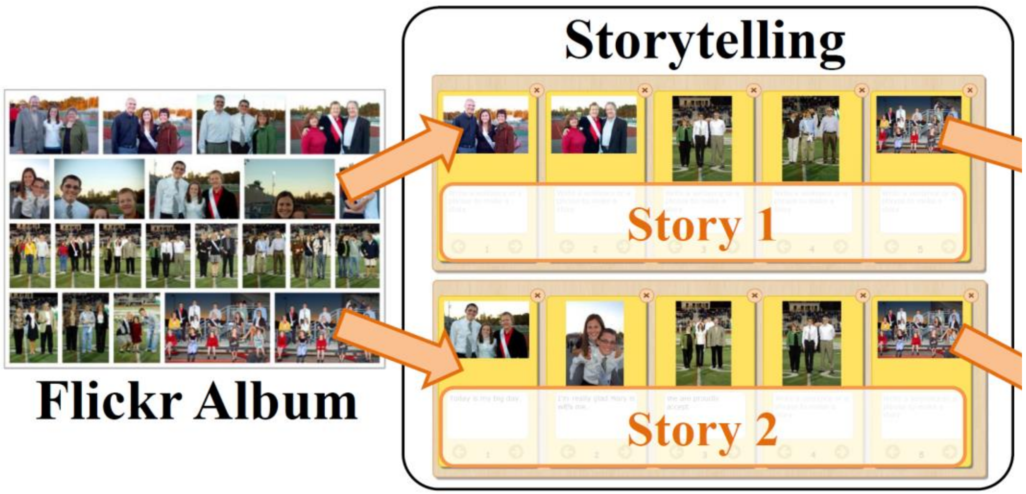
Dataset Construction **Workflow**

These terms are then used to collect albums using the Flickr API.[3] We only include albums with 10 to 50 photos where all album photos are taken within a 48-hour span and CC-licensed. See Table 1 for the query terms with the most albums returned.

beach (684)	breaking up (350)	easter (259)
amusement park (525)	carnival (331)	church (243)
building a house (415)	visit (321)	graduation ceremony (236)
party (411)	market (311)	office (226)
birthday (399)	outdoor activity (267)	father's day (221)

Table 1: The number of albums in our tiered dataset for the 15 most frequent kinds of stories.

Dataset Construction **Workflow** (Cont.)



Form a photo seq + write a short story

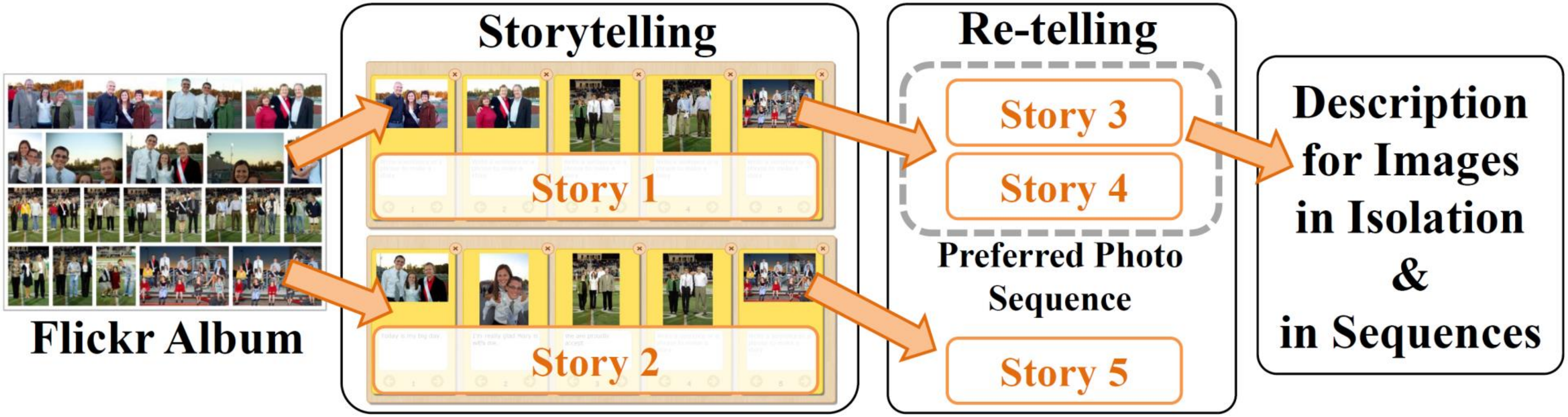
(2 crowd workers)

Dataset Construction **Workflow** (Cont.)

Pick a photo seq + write a short story

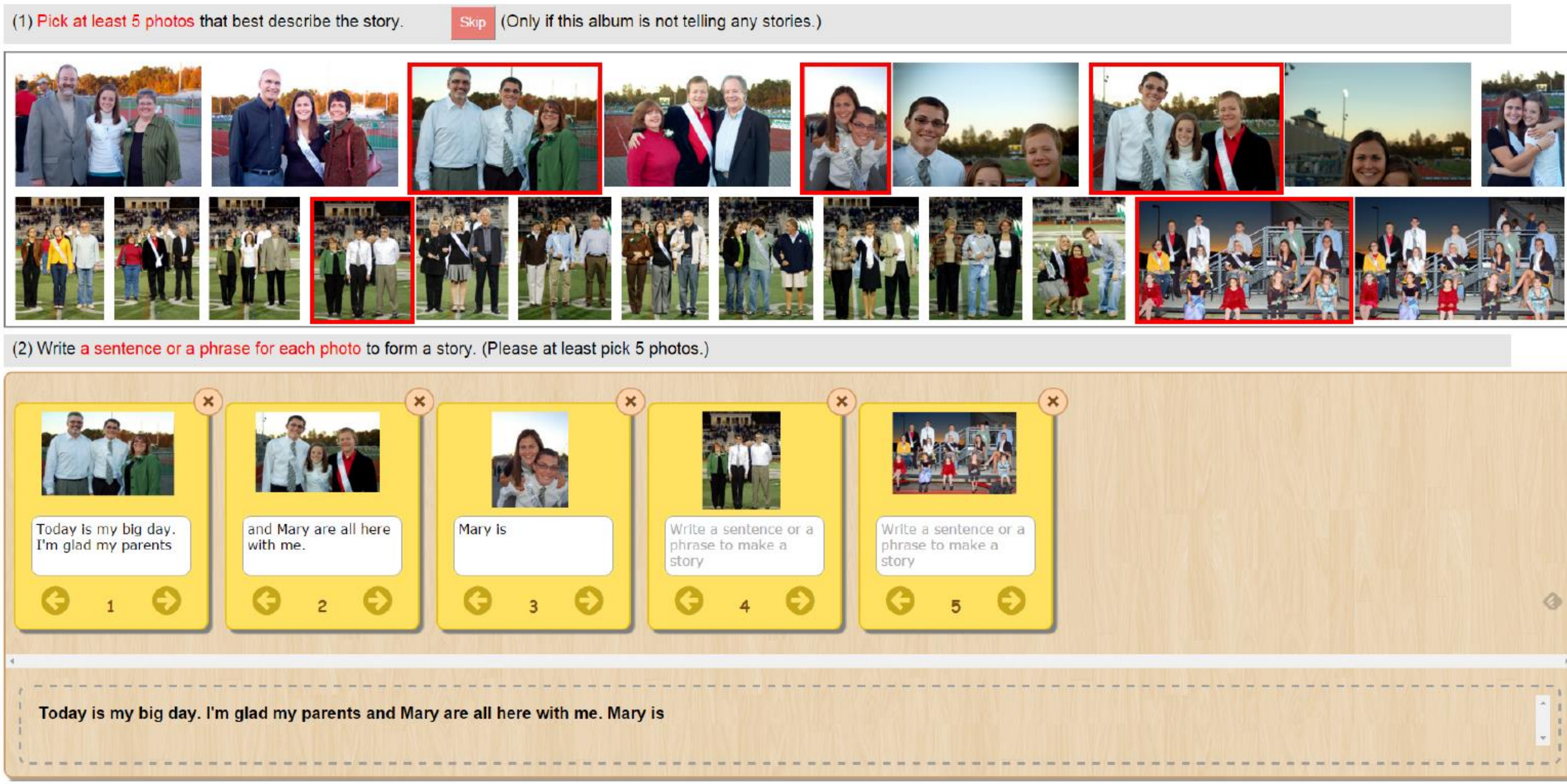
(3 crowd workers)

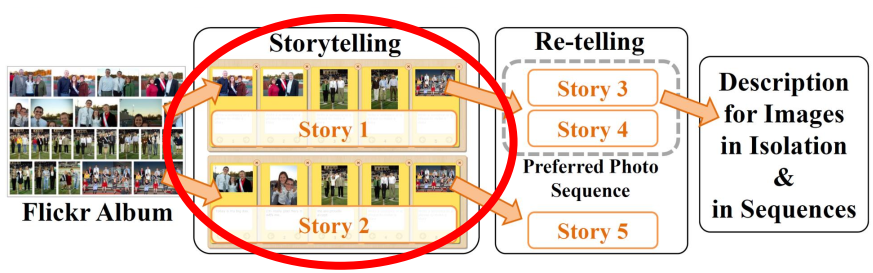
Dataset Construction **Workflow** (Cont.)



DII: Description for Images in Isolation
DIS: Description for Images in Seq
SIS: Story for Images in Seq

Worker Interface





What do the stories look like?

"A discus got stuck up on the roof. Why not try getting it down with a soccer ball? Up the soccer ball goes. It didn't work so we tried a volley ball. Now the discus, soccer ball, and volleyball are all stuck on the roof."

Compare with **Image Captions**

A black frisbee is sitting on top of a roof.
A man playing soccer outside of a white house with a red door.
The boy is throwing a soccer ball by the red door.
A soccer ball is over a roof by a frisbee in a rain gutter.
Two balls and a Frisbee are on top of a roof.

10k+ Flickr Albums Included

- **10k+ x 2** unique photo sequences
- **10k+ x 5** unique short stories

Our dataset includes 10,117 Flickr albums with 210,819 unique photos. Each album on average has 20.8 photos ($\sigma = 9.0$). The average time span of each album is 7.9 hours ($\sigma = 11.4$). Further details of each tier of the dataset are shown in Table 2.[6]

VIST Has More **Abstract Terms**

Data Set	#(Txt, Img) Pairs (k)	Vocab Size (k)	Avg. #Tok	%Abs	Frazier	Yngve	Ppl
Brown	52.1	47.7	20.8	15.2%	18.5	77.2	194.0
DII	151.8	13.8	11.0	21.3%	10.3	27.4	147.0
DIS	151.8	5.0	9.8	24.8%	9.2	23.7	146.8
SIS	252.9	18.2	10.2	22.1%	10.5	27.5	116.0

Table 2: A summary of our dataset, following the proposed analyses of Ferraro et al. (2015), including the Frazier and Yngve measures of syntactic complexity. The balanced Brown corpus (Marcus et al., 1999), provided for comparison, contains only text. Perplexity (Ppl) is calculated against a 5-gram language model learned on a generic 30B English words dataset scraped from the web.

Closer to **Modern, Internet English**

Data Set	#(Txt, Img) Pairs (k)	Vocab Size (k)	Avg. #Tok	%Abs	Frazier	Yngve	Ppl
Brown	52.1	47.7	20.8	15.2%	18.5	77.2	194.0
DII	151.8	13.8	11.0	21.3%	10.3	27.4	147.0
DIS	151.8	5.0	9.8	24.8%	9.2	23.7	146.8
SIS	252.9	18.2	10.2	22.1%	10.5	27.5	116.0

Table 2: A summary of our dataset, following the proposed analyses of Ferraro et al. (2015), including the Frazier and Yngve measures of syntactic complexity. The balanced Brown corpus (Marcus et al., 1999), provided for comparison, contains only text. Perplexity (Ppl) is calculated against a 5-gram language model learned on a generic 30B English words dataset scraped from the web.

Format of VIST Task

Input: A sequence of 5 photos

Output: A short story describing the photo sequence

How to **Generate** Stories (in 2015)?

To train the story generation model, we use a sequence-to-sequence recurrent neural net (RNN) approach, which naturally extends the single-image captioning technique of Devlin et al. (2015) and Vinyals et al. (2014) to multiple images. Here, we encode an image *sequence* by running an RNN over the `fc7` vectors of each image, in reverse order. This is used as the initial hidden state to the story decoder model, which learns to produce the story one word at a time using softmax loss over the training data vocabulary. We use Gated Recurrent Units (GRUs) (Cho et al., 2014) for both the image encoder and story decoder.

Example Outputs

The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water.

Example Outputs

+*Viterbi*	This is a picture of a family. This is a picture of a cake. This is a picture of a dog. This is a picture of a beach. This is a picture of a beach.
+*Greedy*	The family gathered together for a meal. The food was delicious. The dog was excited to be there. The dog was enjoying the water. The dog was happy to be in the water.
-*Dups*	The family gathered together for a meal. The food was delicious. The dog was excited to be there. The kids were playing in the water. The boat was a little too much to drink.
+*Grounded*	The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water.

Table 5: Example stories generated by baselines.

How about **evaluation?**

- Evaluating story quality is hard.
  - Not easy for humans.
  - Very hard for computers.

VIST uses **Human Evaluation**

For the human judgements, we again use crowd-sourcing on MTurk, asking five judges per story to rate how strongly they agreed with the statement "If these were my photos, I would like using a story like this to share my experience with my friends".[7] We take the average of the five judgments as the final score for the story. For the automatic metrics, we use

[7] Scale presented ranged from "Strongly disagree" to "Strongly agree", which we convert to a scale of 1 to 5.

Human Evaluation on **Different Aspects**

- Visual Storytelling Challenge (2018)

Human Evaluation on **Different Aspects** (Cont.)

- **Focus** ("This story is focused.")
- **Structure and Coherence** ("The story is coherent."):
- **I Would Share** ("If these were my photos, I would like using a story like this to share my experience with my friends.")
- **Written by a Human** ("This story sounds like it was written by a human.")
- **Visually Grounded** ("This story directly reflects concrete entities in the photos.")
- **Detailed** ("This story provides an appropriate level of detail.")

Mitchell, M., Huang, T. H., Ferraro, F., & Misra, I. (2018, June). **Proceedings of the First Workshop on Storytelling.** In Proceedings of the First Workshop on Storytelling.

Humans are still pretty good…

- Results of VIST Challenge 2018

Team	Focused	Coherent	Willing to Share	Written by A Human	Visually Grounded	Detailed	Total Score
DG-DLMX	3.347	3.278	2.871	3.222	2.886	2.893	18.498
SnuBiVtt (Late)	3.548	3.524	3.075	3.589	3.236	3.323	20.295
NLPSA501	3.111	2.870	2.769	2.870	3.072	2.881	17.574
UCSB-NLP	3.236	3.065	2.767	3.029	3.032	2.867	17.995
Human (Public Test Set)	4.025	3.975	3.772	4.003	3.965	3.857	23.596

Automatic Evaluation

- **METEOR** aligns better with human ratings.

	METEOR	BLEU	Skip-Thoughts
r	0.22 (2.8e-28)	0.08 (1.0e-06)	0.18 (5.0e-27)
ρ	0.20 (3.0e-31)	0.08 (8.9e-06)	0.16 (6.4e-22)
τ	0.14 (1.0e-33)	0.06 (8.7e-08)	0.11 (7.7e-24)

Table 4: Correlations of automatic scores against human judgements, with p-values in parentheses.

However …

No Metrics Are Perfect:
Adversarial Reward Learning for Visual Storytelling

Xin Wang*, Wenhu Chen*, Yuan-Fang Wang , William Yang Wang
University of California, Santa Barbara
{xwang,wenhuchen,yfwang,william}@cs.ucsb.edu

Wang, X., Chen, W., Wang, Y. F., & Wang, W. Y. (2018, July). **No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling.** In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 899-909).

And…

Reference: Human-**Written** Stories				
	BLEU4	**METEOR**	**ROUGE**	**Skip-Thoughts**
GLAC	0.03	0.30	0.26	0.66
GLAC Edited By Human	0.02	0.28	0.24	0.65

Table 4: Average evaluation scores on GLAC stories, using human-written stories as references. All the automatic evaluation metrics generate lower scores even when the editing was done by human.

Hsu, T. Y., Huang, C. Y., Hsu, Y. C., & Huang, T. H. (2019, July). **Visual Story Post-Editing.** In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6581-6586).

VISTRank (ACL’22)

Learning to Rank Visual Stories from Human Ranking Data

Chi-Yang Hsu[1*], Yun-Wei Chu[2*], Vincent Chen[3*], Kuan-Chieh Lo[3], Chacha Chen[4],
Ting-Hao (Kenneth) Huang[1], Lun-Wei Ku[3]
Pennsylvania State University [1], Purdue University [2],
Institute of Information Science, Academia Sinica[3], University of Chicago[4]
{cxh5437, txh710}@psu.edu, {chu198}@purdue.edu,
{vincent0110, kclo7898, lwku}@iis.sinica.edu.tw, {chacha}@uchicago.edu

Abstract

Visual storytelling (VIST) is a typical vision and language task that has seen extensive development in the natural language generation research domain. However, it remains unclear whether conventional automatic evaluation metrics for text generation are applicable on VIST. In this paper, we present the VHED (VIST Human Evaluation Data) dataset, which first re-purposes human evaluation results for automatic evaluation; hence we develop Vrank (VIST ranker), a novel reference-free VIST metric for story evaluation.[1] We first show that the results from commonly adopted automatic

Reference: i decided my dog would like a train ride. off to the train station we go. this is the train we will be taking our short trip on. my friend is the conductor. he is getting ready to attach the cars. here is the train all together. as you can see, my dog had a fantastic time.

Model 1 (BLEU-1: 0.605, Human Rankers: 👍)
the city was very busy. there were many different kinds
some were very unique. they were

Model 2 (BLEU-1: 0.354, Human Rankers
i went to the park station. it was a train trip to
was very long. we had to go on our w
is so happy to see us.

COMING SOON

Outline

- The **birth** of the Visual Storytelling (VIST) task
- The **evolvement** of VIST technologies
- The **applications** of VIST

Other Interesting V&L Work

What's it going to take to get you in this car today?
Relax! It just smells the other car on you.
It runs entirely on legs.
Just don't tailgate during mating season.
It's only been driven once.
He even cleans up his road kill.
The spare leg is in the trunk.
Comfortably eats six.
She runs like a dream I once had.

Inside Jokes: Identifying Humorous Cartoon Captions

Dafna Shahaf
Microsoft Research
dshahaf@microsoft.com

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

Robert Mankoff
The New Yorker Magazine
bob_mankoff@newyorker.com

Shahaf, D., Horvitz, E., & Mankoff, R. (2015, August). **Inside jokes: Identifying humorous cartoon captions.** In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1065-1074).

Other Interesting V&L Work (Cont.)

(a) **Generated**: a poll (pole) on a city street at night.
Retrieved: the light knight (night) chuckled.
Human: the knight (night) in shining armor drove away.

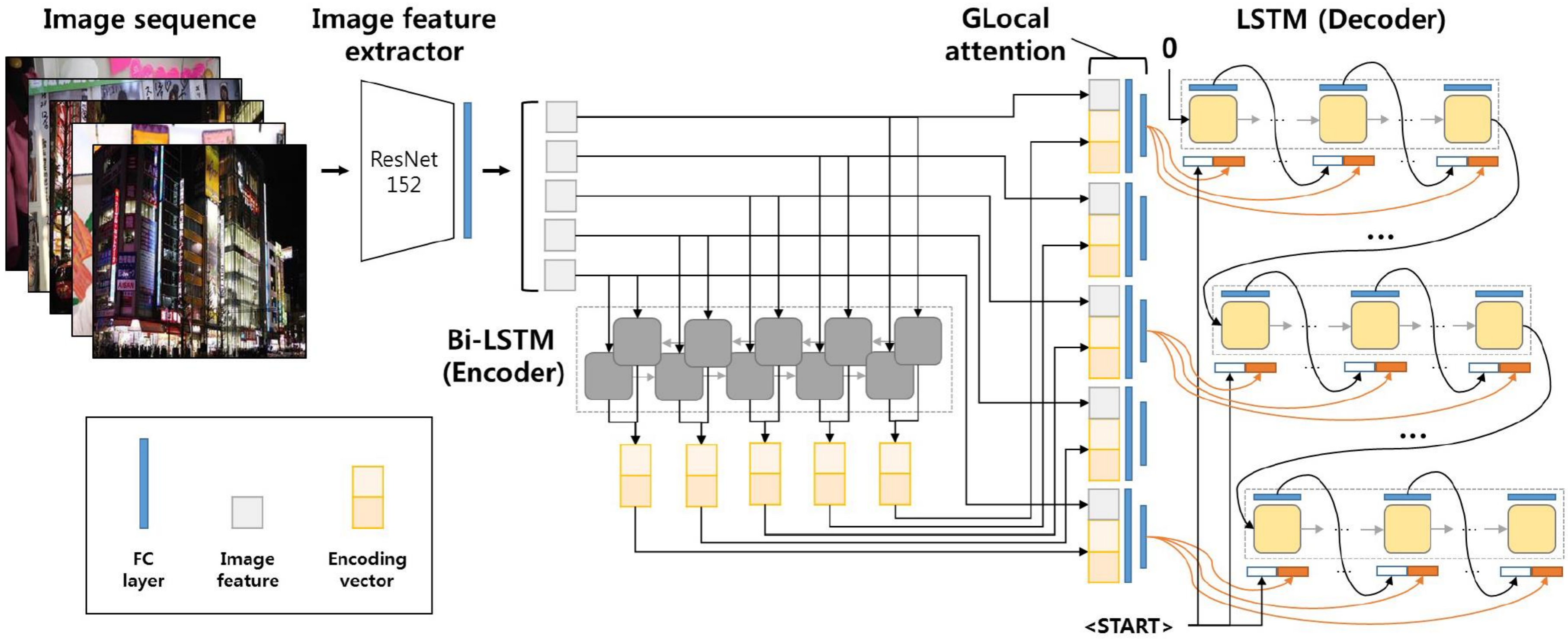
(b) **Generated**: a bare (bear) black bear walking through a forest.
Retrieved: another reporter is standing in a bare (bear) brown field.
Human: the bear killed the lion with its bare (bear) hands.

Chandrasekaran, A., Parikh, D., & Bansal, M. (2018, June). **Punny Captions: Witty Wordplay in Image Descriptions.** In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 770-775).

Visual Storytelling Challenge (2018)

GLACNet

- It received the highest human ratings in the VIST Challenge 2018.



Kim, T., Heo, M. O., Son, S., Park, K. W., & Zhang, B. T. (2018). **Glac net: Glocal attention cascading networks for multi-image cued story generation.** arXiv preprint arXiv:1805.10973.

AREL: Learning to Reward

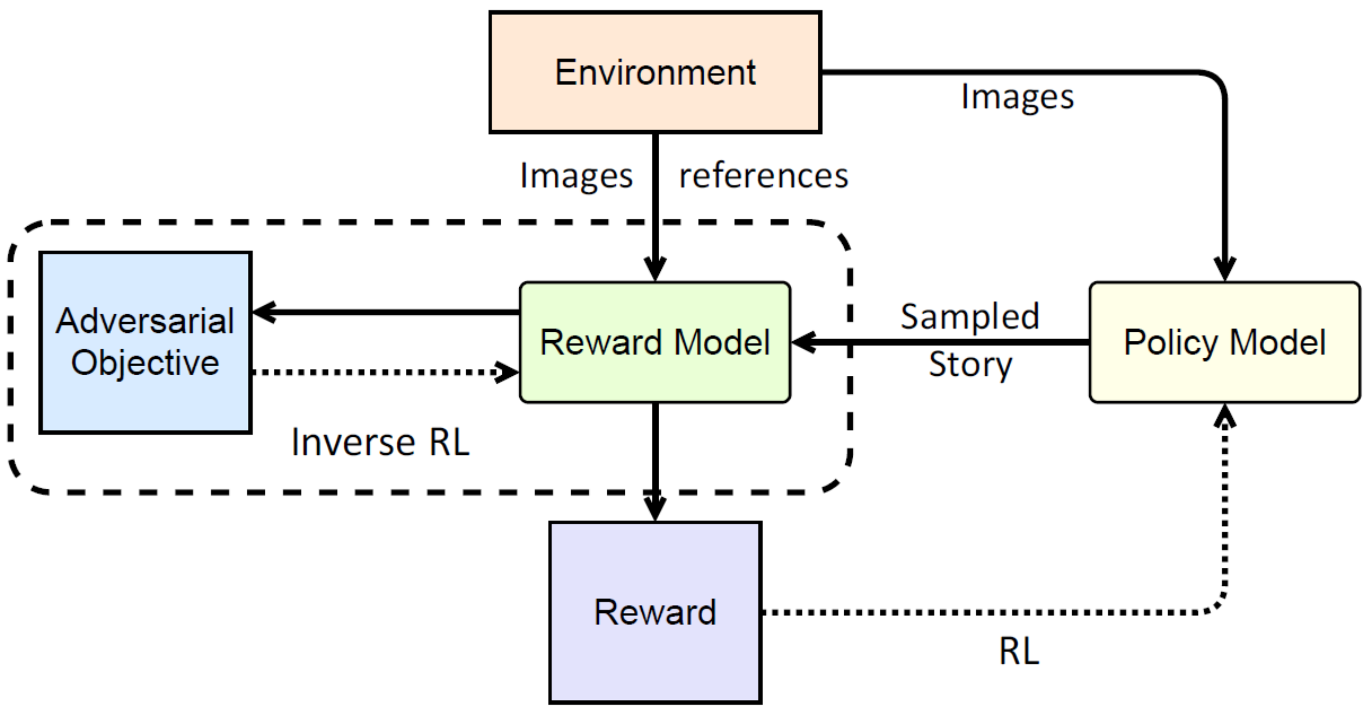


Figure 2: AREL framework for visual storytelling.

Wang, X., Chen, W., Wang, Y. F., & Wang, W. Y. (2018, July). **No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling.** In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 899-909).

Composite Rewards for VIST

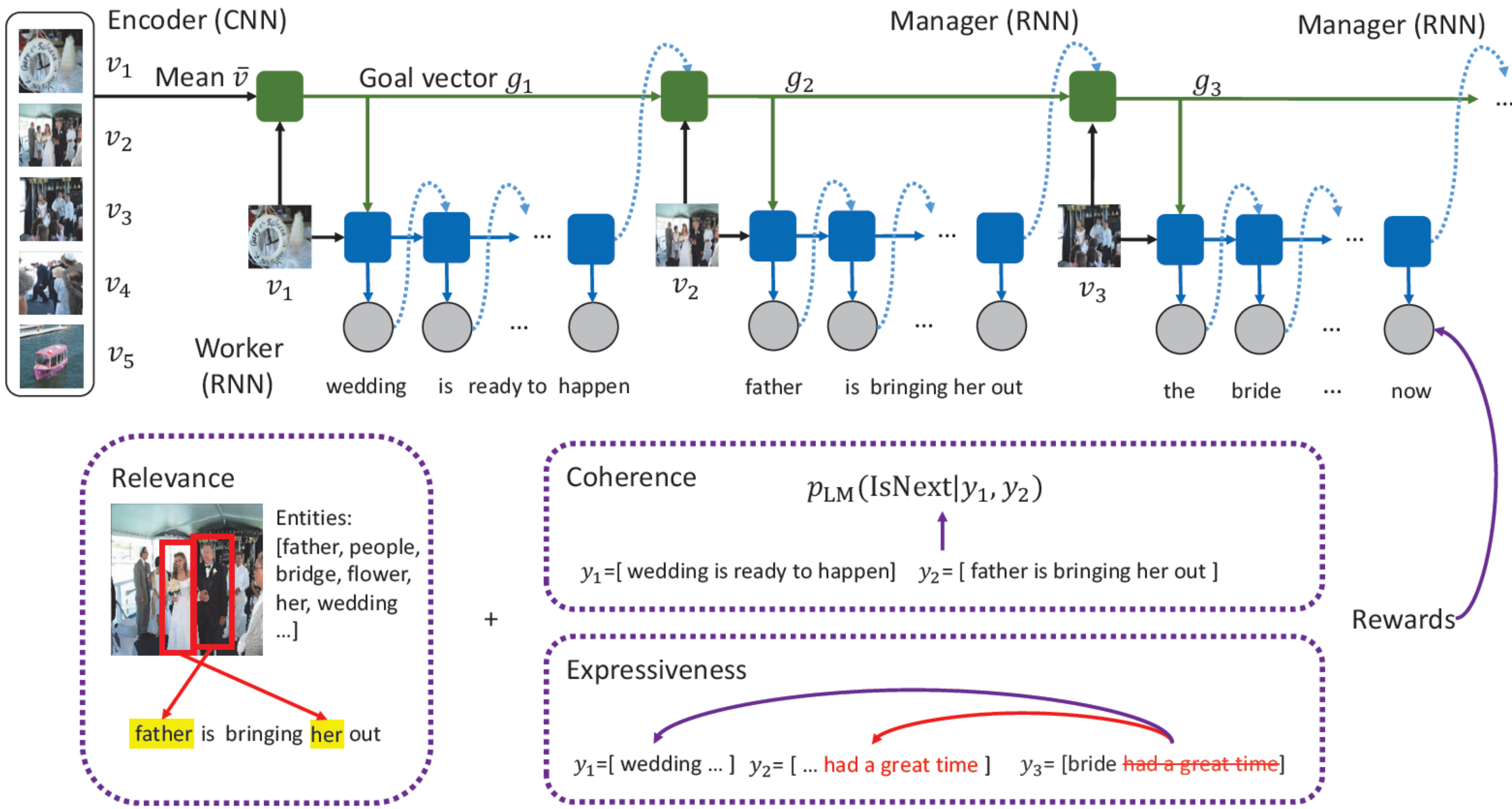


Figure 2: Model architecture and three rewards. Words highlighted in yellow show relevant concepts in the image.

Hu, J., Cheng, Y., Gan, Z., Liu, J., Gao, J., & Neubig, G. (2020, April). **What makes a good story? designing composite rewards for visual storytelling.** In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 7969-7976).

Stories Became More Coherent

- Results of VIST Challenge 2018

Team	Focused	Coherent	Willing to Share	Written by A Human	Visually Grounded	Detailed	Total Score
DG-DLMX	3.347	3.278	2.871	3.222	2.886	2.893	18.498
SnuBiVtt (Late)	3.548	$\boxed{3.524}$	3.075	3.589	3.236	3.323	20.295
NLPSA501	3.111	2.870	2.769	2.870	3.072	2.881	17.574
UCSB-NLP	3.236	3.065	2.767	3.029	3.032	2.867	17.995
Human (Public Test Set)	4.025	$\boxed{3.975}$	3.772	4.003	3.965	3.857	23.596

But, machine-generated stories are still monotonous ...

GLAC: the city was lit up at night . the buildings were tall and bright . the skyline was beautiful . the streets were busy with people . the streets were empty .

Human: the skyscrapers are some of the tallest buildings across the country . at night , the city hosted a nightly carnival . the bridge is much more convenient at night . we decided to use the bridge to get to the city carnival in record breaking time . many vendors had great food to offer at the carnival . the carnival had many inner city people show up .

Why?

- VIST dataset is relatively **small**
  - MS COCO Caption: 995k+ captions
  - VQA dataset: 760k+ questions + 10M+ answers
  - ROCStory dataset: 98k+ stories
  - VIST dataset: ~50k+ stories

- **Relations between images** were not used/modeled

What can we do?

- VIST dataset is relatively **small**
  - MS COCO Caption: 995k+ captions
  - VQA dataset: 760k+ questions + 10M+ answers
  - ROCStory dataset: 98k+ stories
  - VIST dataset: ~50k+ stories

→ ***Use external resources***

- **Relations between images** were not used/modeled

→ ***Connect neighbor images***

KG-Story

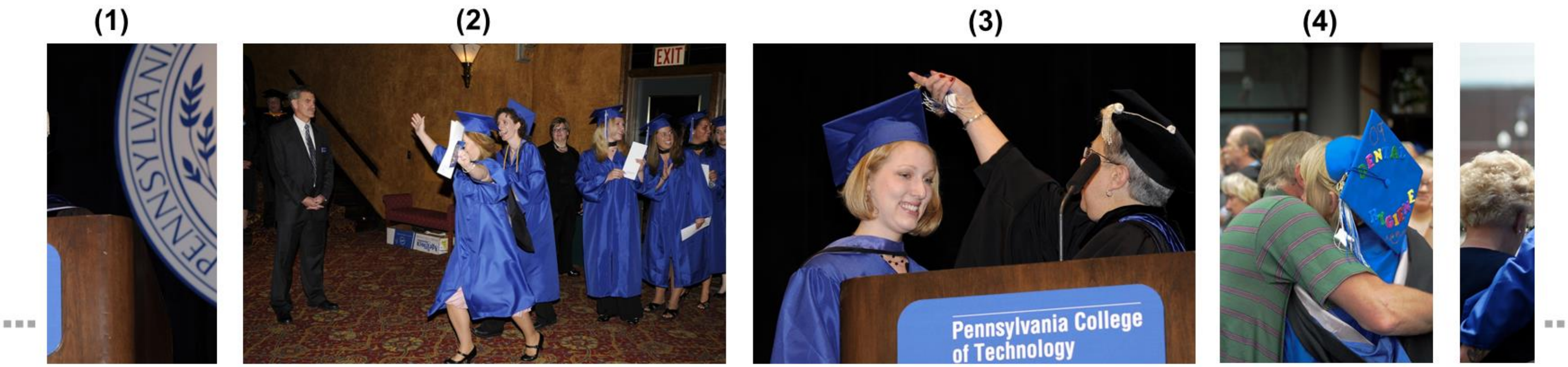
Knowledge-Enriched Visual Storytelling

Chao-Chun Hsu[1*] Zi-Yuan Chen[2*] Chi-Yang Hsu[3]
Chih-Chia Li[4] Tzu-Yuan Lin[5] Ting-Hao (Kenneth) Huang[3] Lun-Wei Ku[2,6]
[1]University of Colorado Boulder, [2]Academia Sinica, [3]Pennsylvania State University,
[4]National Chiao Tung University, [5]National Taiwan University,
[6]Most Joint Research Center for AI Technology and All Vista Healthcare
chao-chun.hsu@colorado.edu, {zychen, lwku}@iis.sinica.edu.tw, {cxh5437, txh710}@psu.edu

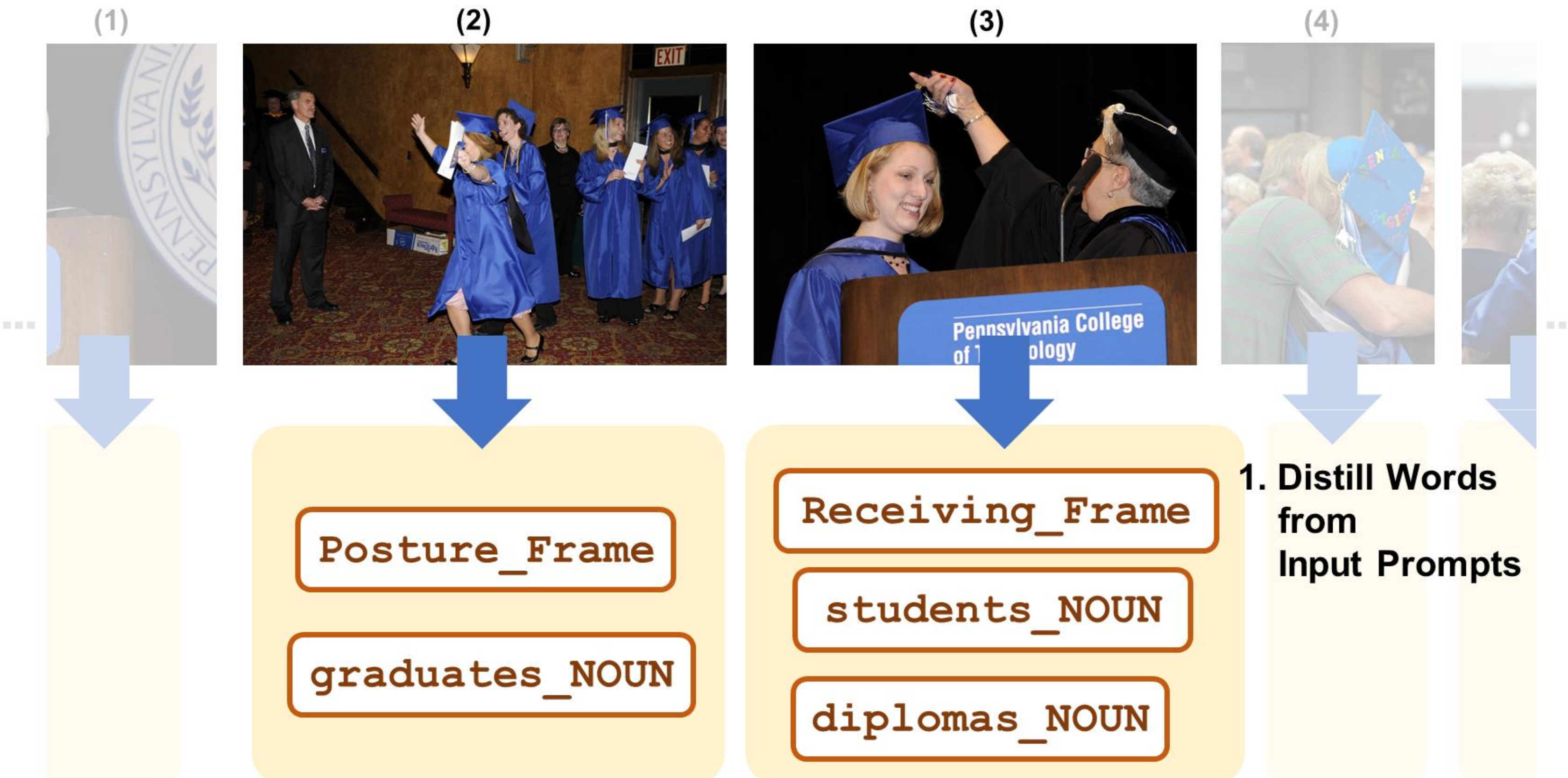
- Modular pipeline
- Image → Words → Story
- Explicitly connect two neighbor images
- Use external knowledge graphs and datasets

Hsu, C. C., Chen, Z. Y., Hsu, C. Y., Li, C. C., Lin, T. Y., Huang, T. H., & Ku, L. W. (2020, April). **Knowledge-enriched visual storytelling.** In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 7952-7960).

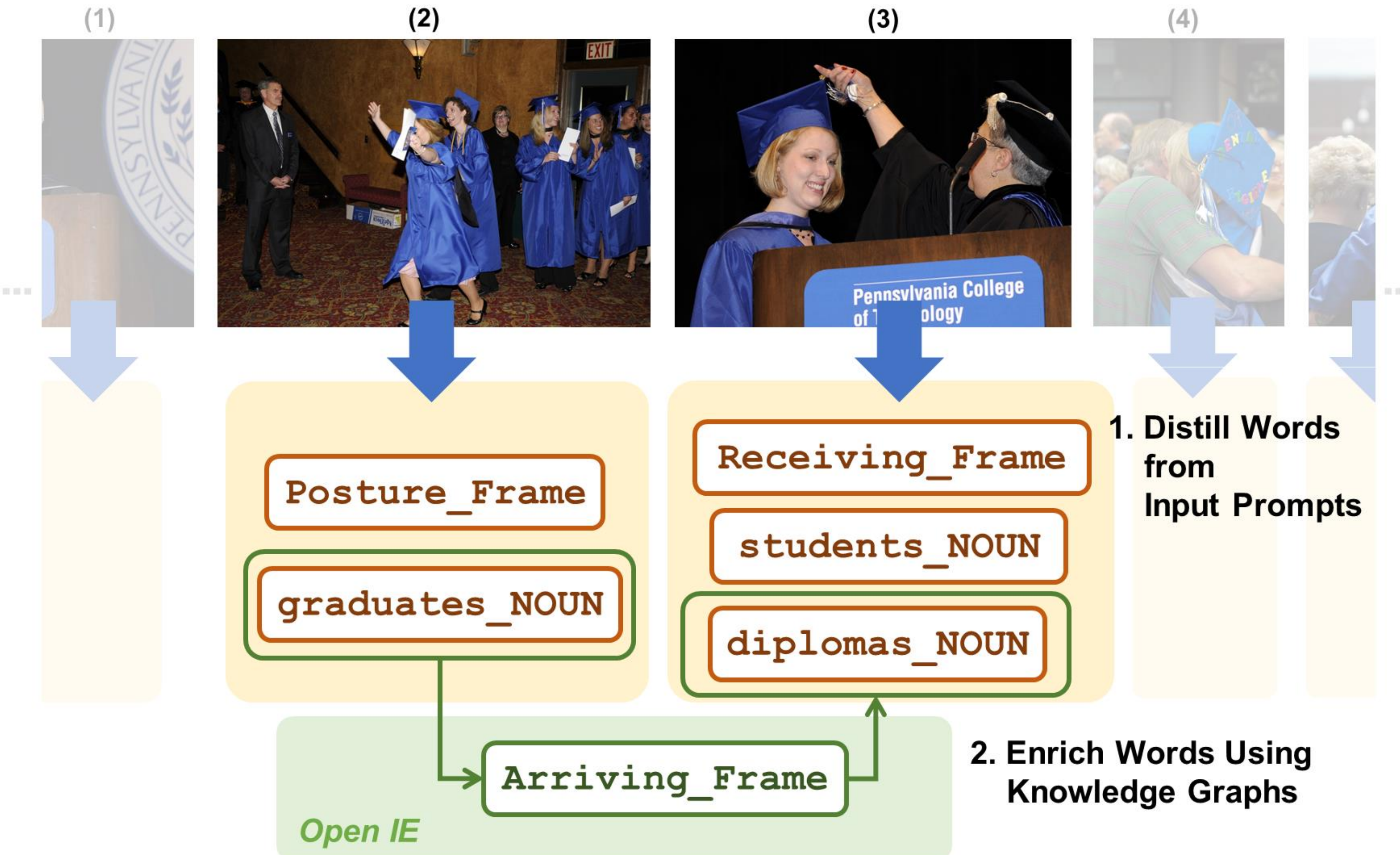
Input Photo Sequence



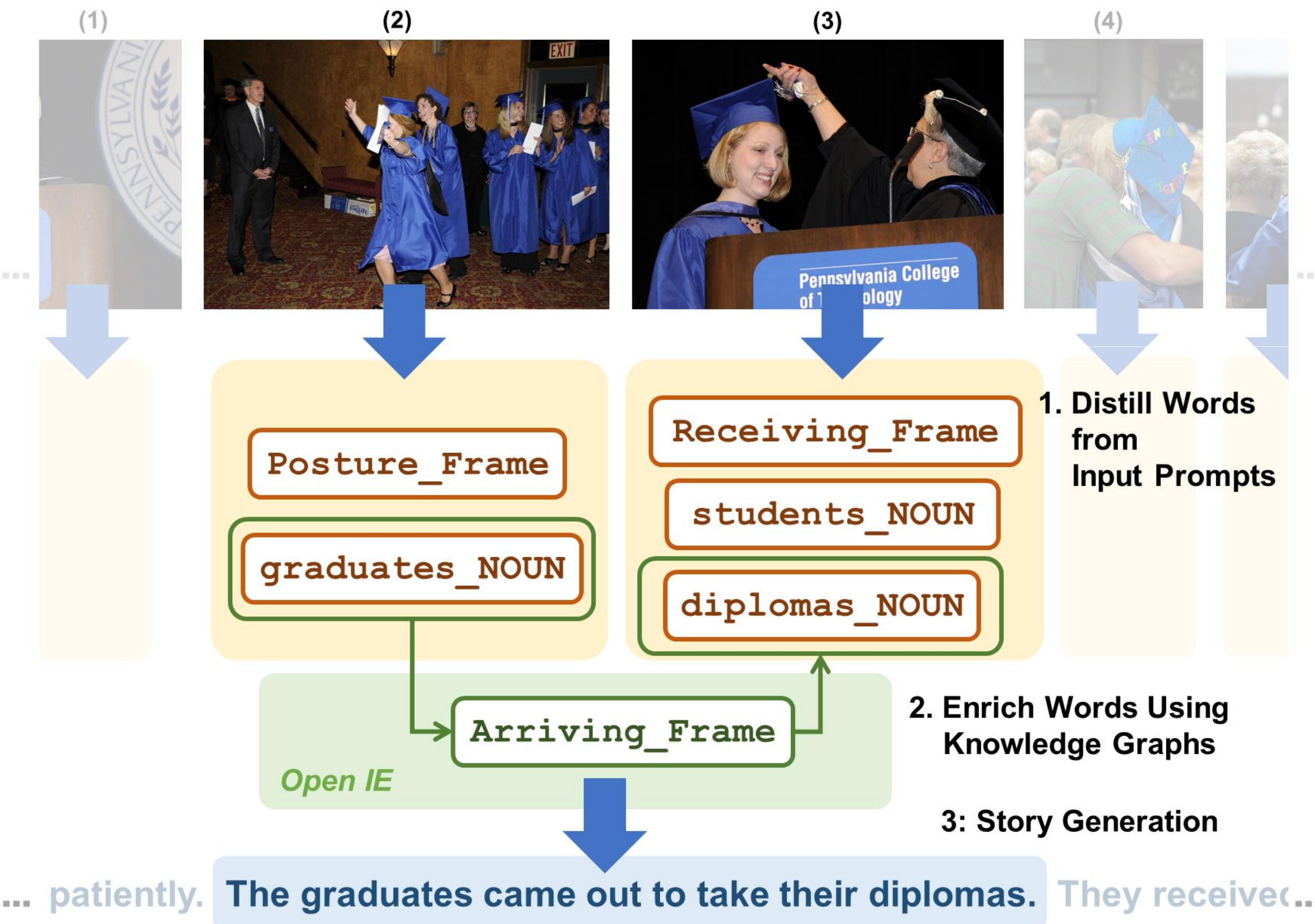
Step 1: Image to Terms



Step 2: Enrich Terms



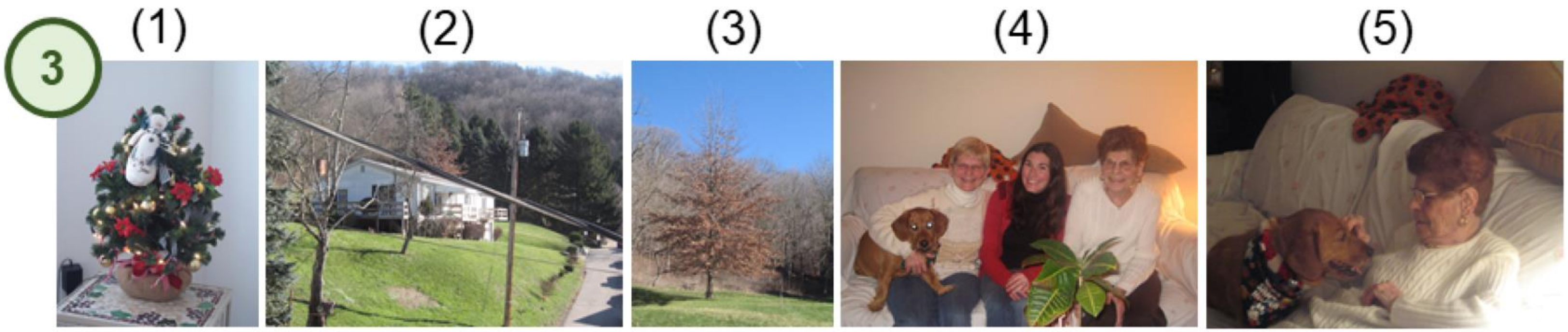
Step 3: Term to Story



Pros and Cons of KG-Story

- **Pros**
  - Easy to use external extractors (image to terms)
  - Easy to use external KGs (word enrichment)
  - Easy to use external story datasets (story generation)
  - Can technically be applied to text-only story generation
- **Cons**
  - Modular pipelines can be harder to work with
  - Propagation of error

Example Output



OpenIE: the wedding reception was very special . it was a beautiful house . there were so many trees . everyone had a great time . ***even the dog had a great time !*** the dog was very well behaved .

GLAC: the family was having a great time at the christmas party . the tree was covered in snow . the trees were beautiful . the kids were very excited . the baby was happy to be there .

Human: we visited family for christmas . they live out in the country far from the city . the trees lost their leaves because it is so cold outside . they were so happy that we had arrived . even the dog had a marry christmas .

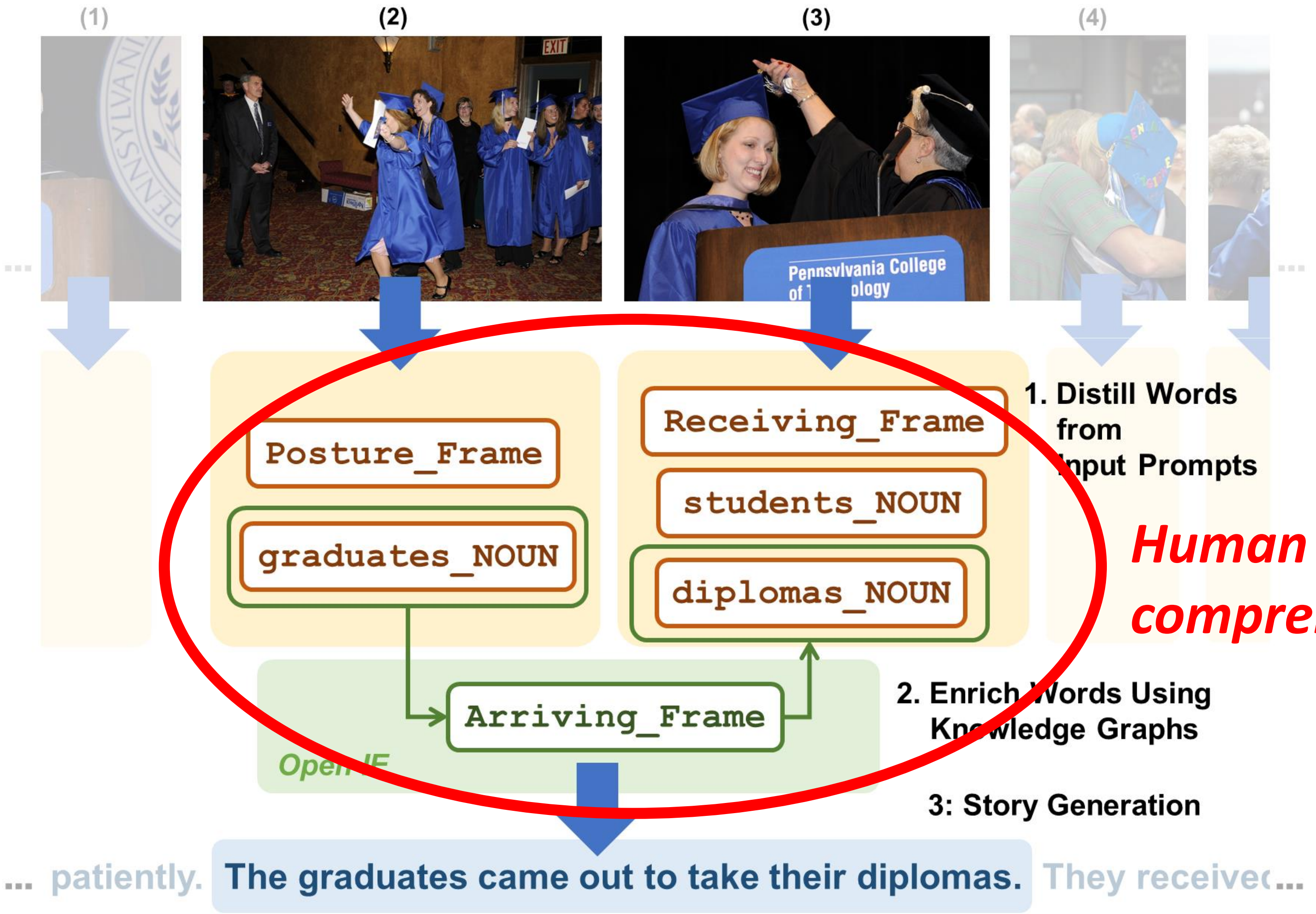
Human Evaluation (Rank)

Human Evaluation (Story Displayed **with** Images)

	GLAC (Kim et al. 2018)	**No KG**	**OpenIE**	**Visual Genome**	**Human**
Avg. Rank (1 to 5)	3.053	3.152	**2.975***	**2.975***	2.846

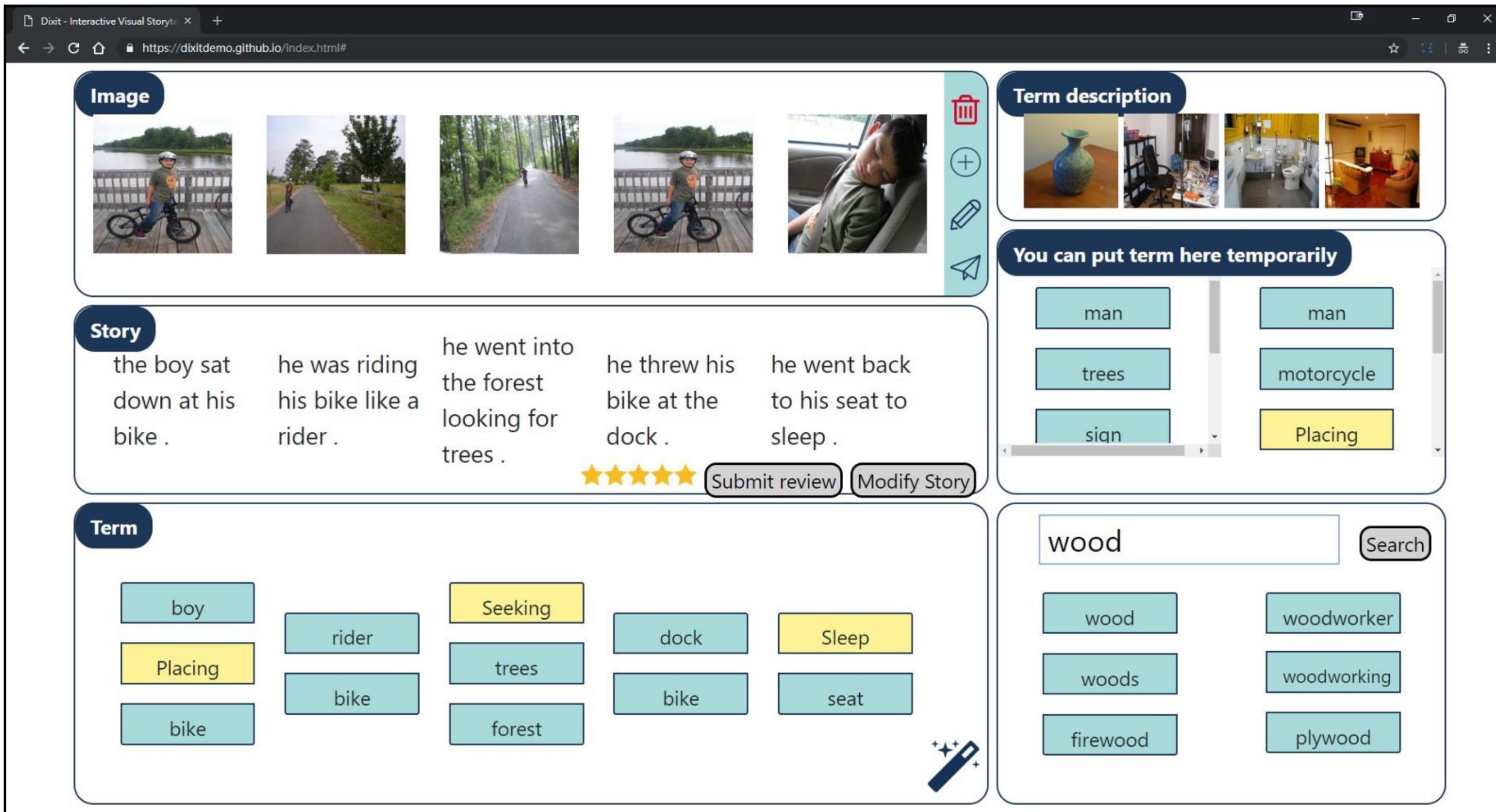
Table 2: Direct comparison evaluation of KG-Story model. Numbers indicate average rank given to stories (from 1 to 5, lower is better.) Stories generated by KG-Story using either OpenIE or Visual Genome are on average ranked significantly better (lower) than that of GLAC (unpaired t-test, $p < 0.05$, N=2500).

Bonus: Allow User Control



Human comprehensible!

Interactive Visual Storytelling via Term Manipulation



Hsu, C. C., Chen, Y. H., Chen, Z. Y., Lin, H. Y., Huang, T. H., & Ku, L. W. (2019, May). **Dixit: Interactive visual storytelling via term manipulation.** In The World Wide Web Conference (pp. 3531-3535).

Why stop at adding only *1* edge?

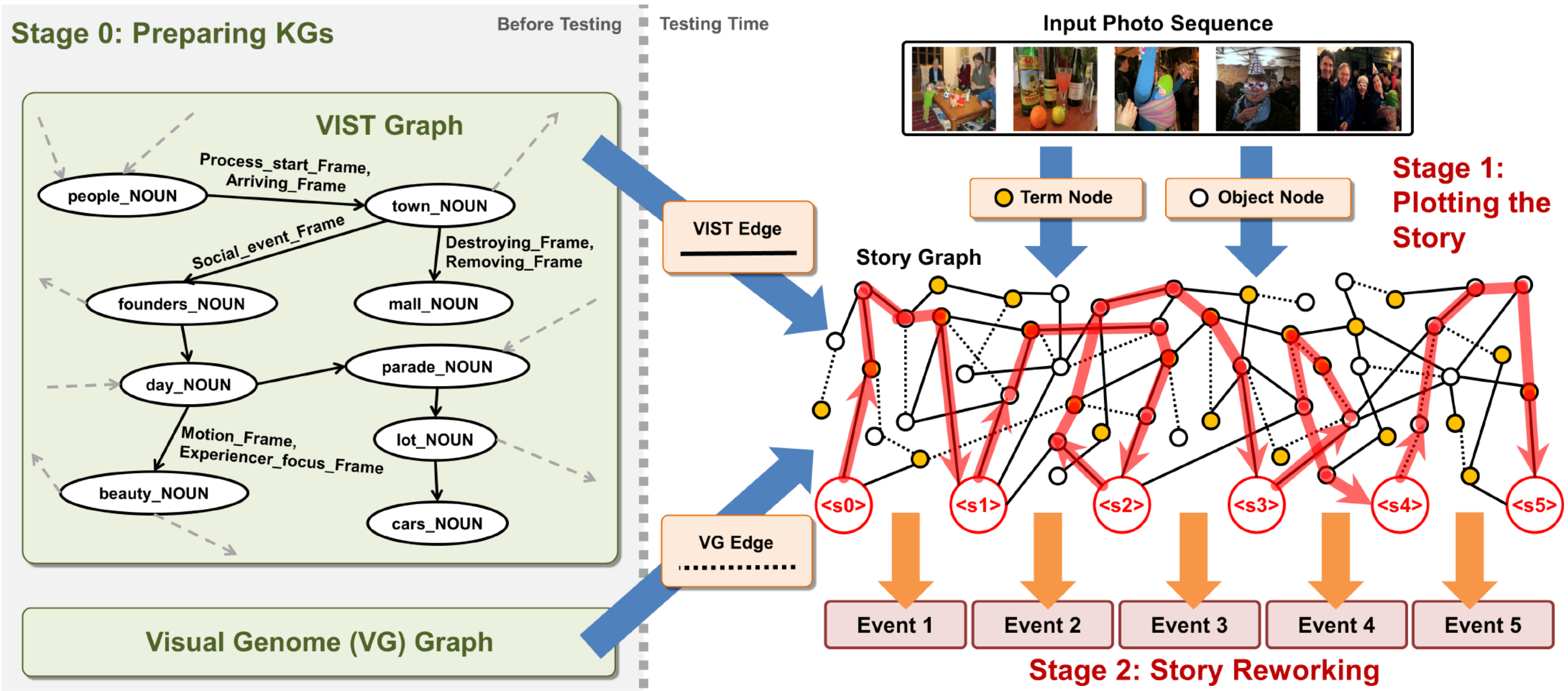


Figure 1: Overview of PR-VIST. In **Stage 1 (Story Plotting)**, PR-VIST first constructs a graph that captures the relations between all the elements in the input image sequence and finds the optimal path in the graph that forms the best storyline. In **Stage 2 (Story Reworking)**, PR-VIST uses the found path to generate the story. PR-VIST uses a story generator and a story evaluator to realize the "rework" process. In **Stage 0 (Preparation)**, a set of knowledge graphs that encode relations between elements should be prepared for the uses in Stage 1.

Hsu, C. Y., Chu, Y. W., Huang, T. H., & Ku, L. W. (2021, August). **Plot and Rework: Modeling Storylines for Visual Storytelling.** In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 4443-4453).

Outline

- The **birth** of the Visual Storytelling (VIST) task
- The **evolvement** of VIST technologies
- The **applications** of VIST

What are the **applications** of VIST?

Input: A sequence of photos





Output: A short story describing the photo sequence

Narrating the **Environment**

Figure 1: Creative Visual Storytelling Pipeline: T1 (Object Identification), T2 (Single Image Inferencing), T3 (Multi-Image Narration)

Lukin, S., Hobbs, R., & Voss, C. (2018, June). **A Pipeline for Creative Visual Storytelling.** In Proceedings of the First Workshop on Storytelling (pp. 20-32).

Content Creation

Kenneth Huang 🔥 Survival Game IN-PERSON
@windx0303

.@tknearem is defending her PhD thesis NOW! @ISTatPENNSTATE #PennStateHCI

11:19 AM · Nov 19, 2021 from IST Building · Twitter for Android

Gaps in Text Quality

OpenIE: the wedding reception was very special . it was a beautiful house . there were so many trees . everyone had a great time . ***even the dog had a great time !*** the dog was very well behaved .

GLAC: the family was having a great time at the christmas party . the tree was covered in snow . the trees were beautiful . the kids were very excited . the baby was happy to be there .

Human: we visited family for christmas . they live out in the country far from the city . the trees lost their leaves because it is so cold outside . they were so happy that we had arrived . even the dog had a marry christmas .

Human Editing is Needed

Machine-Generated Story (a): *visual storytelling*
the family got together for a dinner. the food was delicious. everyone was having a great time. the meal was delicious. the kids had a great time.

Machine-Generated (a) -> Human-Edited Story (b):
the whole family got together for thanksgiving. the food was delicious! everyone had a lot of fun, and the kids played the entire time.

Visual Story Post-Editing

(1) (2) (3) (4) (5)

Machine-Generated Story (a): *visual storytelling*

the family got together for a dinner. the food was delicious. everyone was having a great time. the meal was delicious. the kids had a great time.

Machine-Generated (a) -> Human-Edited Story (b):

the whole family got together for thanksgiving. the food was delicious! everyone had a lot of fun, and the kids played the entire time.

visual story post-editing

Machine-Generated (a) -> Machine-Edited Story (c):

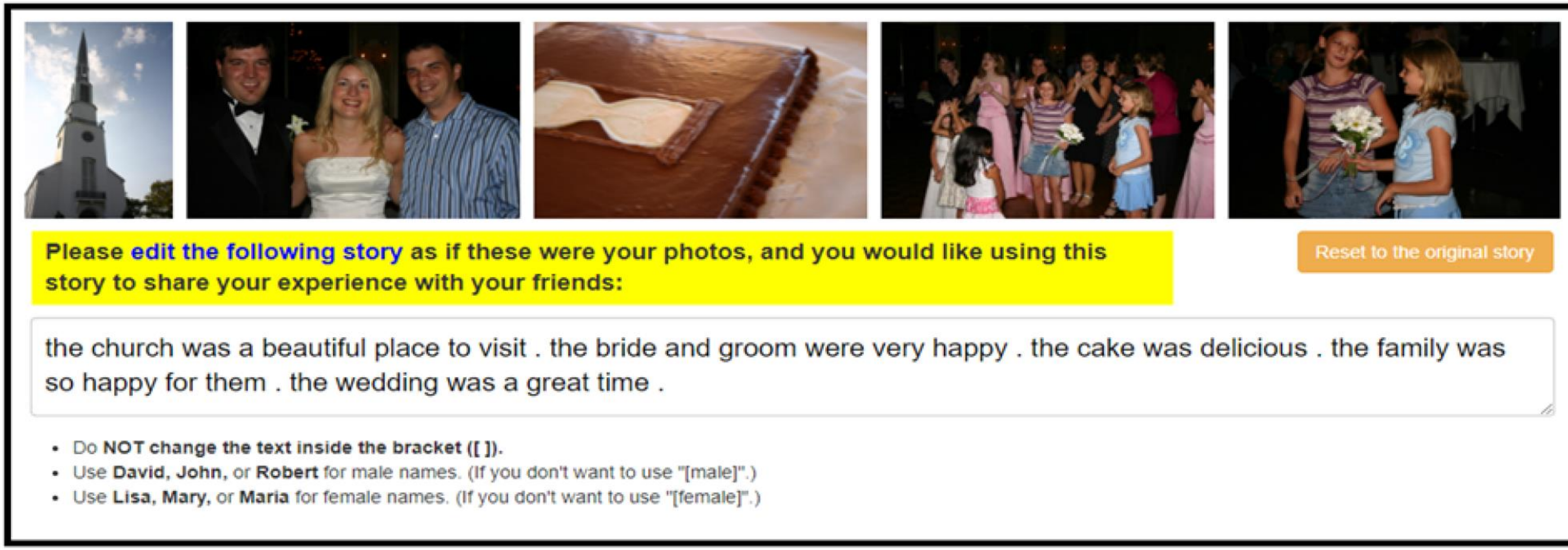
the family got together for a nice dinner. the food was delicious. the guys enjoyed the food since they had never eaten there before. the food was presented well. the dessert was delicious.

Hsu, T. Y., Huang, C. Y., Hsu, Y. C., & Huang, T. H. (2019, July). **Visual Story Post-Editing.** In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6581-6586).

Post-Editing (APE) Task

- Often used in MT
- Treat the text generation model as a **black box**.
- **Pre-** and **post-edited** parallel data are often collected.

Data Collection



Editing Increased **Lexical Diversity**

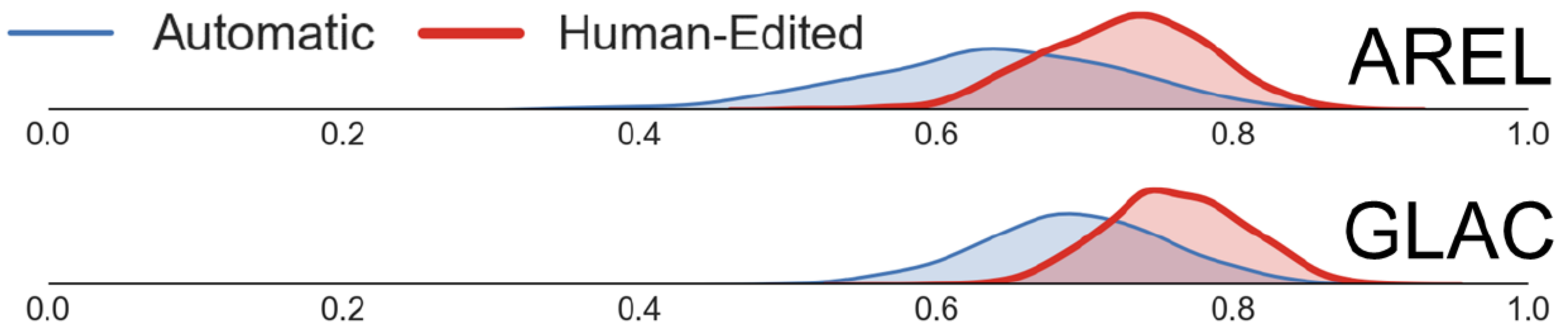
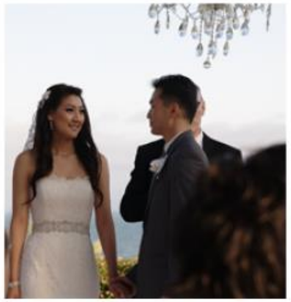


Figure 3: KDE plot of type-token ratio (TTR) for pre-/post-edited stories. People increase lexical diversity in machine-generated stories for both AREL and GLAC.

Post-Editing Improved the Stories

	AREL					
Edited By	Focus	Coherence	Share	Human	Grounded	Detailed
N/A	3.487	3.751	3.763	3.746	3.602	3.761
TF (T)	3.433	3.705	3.641	3.656	3.619	3.631
TF (T+I)	**3.542**	3.693	3.676	3.643	3.548	3.672
LSTM (T)	**3.551**	**3.800**	**3.771**	**3.751**	**3.631**	**3.810**
LSTM (T+I)	**3.497**	3.734	3.746	3.742	3.573	3.755
Human	3.592	3.870	3.856	3.885	3.779	3.878

Example Output

we had a great time at the wedding today. the bride and groom were very happy to be married. the bride and groom were very happy to be married. the bride and groom pose for a picture. at the end of the wedding, the bride and groom pose for a picture.

the wedding was held in a beautiful church. the bride and groom walked down the aisle. they were very happy to be married. the couple looked so lovely together. the bride and groom danced the night away at the reception.

Outline

- The **birth** of the Visual Storytelling (VIST) task
- The **evolvement** of VIST technologies
- The **applications** of VIST

1 Introduction

Beyond understanding simple objects and concrete scenes lies interpreting causal structure; making sense of visual input to tie disparate moments together as they give rise to a cohesive narrative of events through time. This requires moving from reasoning about single images – static moments, devoid of context – to sequences of images that depict events as they occur and change. On the vision side, progressing from single images to images in context allows us to begin to create an artificial intelligence (AI) that can reason about a visual moment given

Huang*, T. H., Ferraro*, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., ... & Mitchell, M. (2016, June). **Visual storytelling.** In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1233-1239).

Are we there yet?

Meta Takeaways

- A good dataset **sets an interesting and rich agenda** for the research community.

- A good **summer intern project** could shape your career!

In2Writing Workshop (@ACL'22)

In2Writing

The First Workshop on Intelligent and Interactive Writing Assistants

The purpose of this interdisciplinary workshop is to bring together researchers from the natural language processing (NLP) and human-computer interaction (HCI) communities as well as industry practitioners and professional writers to discuss innovations in building, improving, and evaluating intelligent and interactive writing assistants. We plan to alternate our workshop venue between an NLP conference and a HCI conference every year to facilitate collaboration.

The first 100 participants get a free premium subscription to Grammarly and Wordtune.

This year the workshop will be held at ACL 2022 in Dublin, Ireland on the 26th of May, 20

The workshop is expected to be hybrid, unless the pandemic situation dictates otherwis

outside our control.

COMING SOON

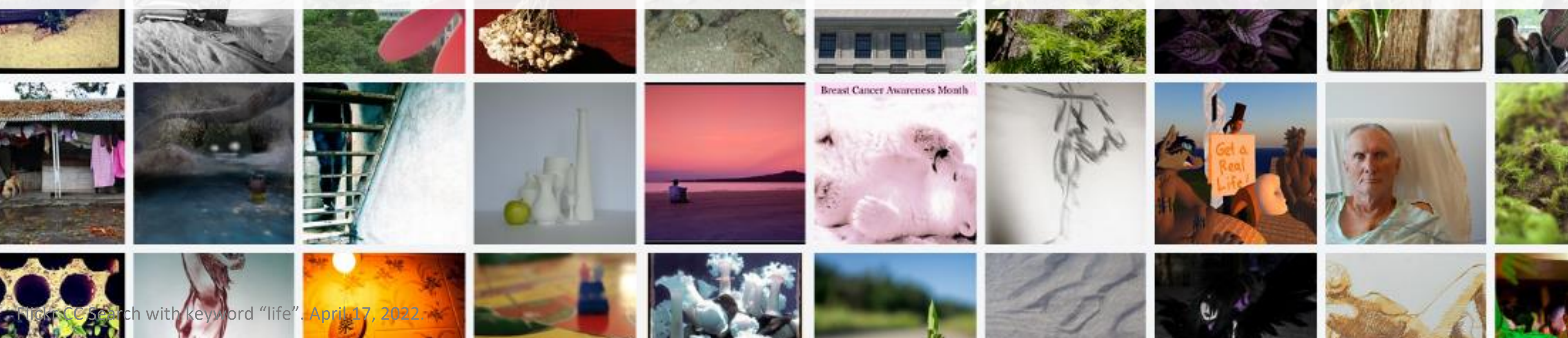
Visual Storytelling

Ting-Hao (Kenneth) Huang

Penn State University

Guest Lecture for CIS 700 - Interactive Fiction and Text Generation

April 19, 2022



Flickr CC Search with keyword “life”. April 17, 2022.