VinaBench: Benchmark for Faithful and Consistent Visual Narratives

Supplementary Material

The supplementary materials contain the following information and materials:

- Data construction details (Section S1).
- Evaluation details (Section S2).
- Experimental setup details (Section S3).
- Full experimental results (Section S4)

S1. VinaBench Data Construction Details

The visual-textual narrative pairs in our benchmark are sampled from three diverse visual storytelling datasets, including Visual Writing Prompts (VWP) [3], Storyboard20K [16] and StorySalon [8]. The VWP dataset contains ∼12K narrative samples, whose visual narrative scenes are extracted and curated from MovieNet [4] frames, with corresponding textual narratives crafted by Amazon Mechanical Turk (AMT) workers. The Storyboard20K dataset covers a broader set of visual narrative scenes sampled from MovieNet and also LSMDC [13], with real movie synopses collected by a two-stage approach of automatic tagging and manual calibration. We filter the narrative samples in Storyboard20K to keep ∼10K of them, which have aligned shot-by-shot movie synopses, serving as the textual narratives. Different from the movie-based narratives in VWP and Storyboard20K, the StorySalon dataset is oriented to animation-style visual narratives, whose images and aligned narrative texts are extracted from diverse YouTube videos and E-books. We use the Google Translation API[1] to translate non-English narrative texts collected in StorySalon into English. To ensure accurate translation, we only apply the API to ∼26K StorySalon scenes (or images) whose associated narrative texts are in the 19 common languages shown in Table S1, and then exclude the narrative samples whose texts are not fully translated into English. Besides, we filter the StorySalon samples whose textual narratives are poor-annotated, *i.e.*, >10% of the sample's scenes are annotated with uninformative texts containing less than 5 words. Finally, ∼2K narrative samples from StorySalon are included.

Based on the sampled visual-textual narrative pairs, VinaBench further annotates the commonsense and discourse constraints underlying each narrative sample, by prompting advanced VLMs and LLMs instead of relying on human annotators. Table S2 summarizes the number of few-shot prompting examples used for each step of our VinaBench constraint annotation. For each annotation step, we tune the number of few-shot examples on a scale of 1 to 3, and select the number that leads to the best annotation results in our pilot study on 10 narrative samples. Figure S4 - S7 list the specific few-shot examples and instructions that we finally used for annotating the image captions, commonsense links, global and scene features in VinaBench, respectively.

Language	**# Scenes**	**Language**	**# Scenes**
Hindi (hi)	8213	Hausa (ha)	926
French (fr)	2503	Spanish (es)	758
Indonesian (id)	2197	Italian (it)	386
Arabic (ar)	2053	Dutch (nl)	198
Marathi (mr)	1544	German (de)	187
Nepali (ne)	1521	Portuguese (pt)	137
Afrikaans (af)	1464	Finnish (fi)	113
Swahili (sw)	1311	Welsh (cy)	82
Vietnamese (vi)	1220	Polish (pl)	78
Uzbek (uz)	1150	**Total**	**26041**

Table S1. Statistics of StorySalon scenes (or images) whose associated non-English narrative texts are translated into English.

Cap.	**Ent.**	**CL**	**Sty.**	**List**	**Attr.**	**Num.**	**Name**	**Time**	**Loc.**
2	3	3	1	2	3	3	2	3	2

Table S2. Number of few-shot examples used for VinaBench data annotation, including dense image captioning (**Cap.**), visual entity extraction from dense captions (**Ent.**), commonsense link construction (**CL**), and the parsing of image appearance style (**Sty.**), global character list (**List**) and attributes (**Attr.**), and each scene's presented character number (**Num.**) and name (**Name**), time of day (**Time**) and location (**Loc.**).

Source	**Set**	**# Nar.**	**# Sce.**	**Avg. # Char.**		**# CL**	**# Label Types**		
				per Nar.	**per Sce.**		**Sty.**	**Time**	**Loc.**
VWP	train	11652	66632	3.07	1.71	274861	9	6	545
	test	834	4901	3.00	1.73	22434			191
Story-board20K	train	9252	92520	3.73	1.40	316356	9	6	551
	test	1194	11940	4.07	1.59	40576			341
Story-Salon	train	1593	21827	6.49	2.14	71489	9	6	518
	test	85	1181	6.64	2.03	4374			111

Table S3. Statistics of VinaBench data samples and annotations, including total number of narratives (**# Nar.**), total number of scenes or images (**# Sce.**), average number of distinct characters per narrative (**Avg. # Char. per Nar.**), average number of presented characters per scene (**Avg. # Char. per Sce.**), total number of commonsense links (**# CL**), total types of appearance style (**# Sty.**), time of day (**# Time**) and location (**# Loc.**) labels.

According to VinaBench annotations, we also exclude the narrative samples that contain no character or commonsense link. Table S3 shows the final statistics of VinaBench narrative samples and annotations. Each VinaBench narrative sample contains ∼8.09 scenes (or images) in average, which is longer than prior image sequences (with a length of 5) studied in visual narrative generation, *i.e.*, VIST [5],

[1] https://github.com/ssut/py-googletrans

Aspect	Metric	Demonstration
Alignment	**Non-Character**	{*generated image for a scene*} Does this image contain or imply {*each non-character visual entity in the scene's gold commonsense links*}? Only answer yes or no.
	Character Number	{*generated image for a scene*} How many characters are in this image? Only answer an Arabic number.
	Character Attribute	{*generated image for a scene*} Character descriptions: {*gold character 1 presented in the scene features*}: {*profile of character 1 in the global features*} {*gold character 2 presented in the scene features*}: {*profile of character 2 in the global features*} ... Do characters in this image fit into their descriptions? Only answer yes or no.
	Time of Day	{*generated image for a scene*} Is this image taken in (or at) the {*gold time of day labeled in the scene features*}? Only answer yes or no.
	Location	{*generated image for a scene*} Is this image taken at a (or an) {*gold location labeled in the scene features*}? Only answer yes or no.
Consistency	**Style**	{*generated image for scene 1*} {*generated image for scene 2*} ... {*generated image for scene N*} Are all these images in the same style? Only answer yes or no.
	Character	{*generated image for scene X*} {*generated image for scene Y*} ... Do all these images contain the same character {*each overlapped character across the scenes X, Y, ..., indicated by their scene features*}: {*profile of the overlapped character in the global features*}? Only answer yes or no.
	Location	{*generated image for scene X*} {*generated image for scene Y*} ... Are all these images taken at the same {*gold location label shared by the scenes X, Y, ..., indicated by their scene features*}? Only answer yes or no.

Table S4. VQA demonstrations used for the fine-grained alignment and consistency metrics in VinaBench. For Alignment of Character Number, we record the average probability of the VLMs (MiniCPM-V-2.6 or LLaVA-OneVision-72B) outputting the correct character number as its first decoded token (or if characters are more than 9, the same number of leading tokens as the correct number of digits). For other metrics, we report the average probability of the VLM outputting *Yes* as its first decoded token. The spans labeled by "{}" in the demonstrations are replaced by their corresponding texts or images.

PororoSV [7] and FlintstonesSV [10]. Besides, VinaBench incorporates new annotations of fine-grained visual narrative constraints, which are not involved in previous visual narrative studies.

S2. VinaBench Evaluation Details

We adopt zero-shot prompting to implement all of our proposed VQA-based fine-grained alignment and consistency metrics in VinaBench. Table S4 lists the specific demonstrations used for our VQA-based metrics. The VQA score of non-character alignment metric is averaged across each non-character visual entity labeled in gold commonsense links. While for other fine-grained alignment metrics, we calculate the average VQA score across each scene in the testing narrative samples. For the style consistency metric, since it is based on all scenes of a narrative, we average the VQA score across each testing narrative sample. In terms of the character and location consistency metrics, the VQA score is averaged across each gold character or location labeled in the narrative that is shared by multiple scenes.

S3. Experimental Setup Details

For the setting of training visual narrative models with LLM-generated constraints (w/ LLM Cons.), we preprocess our annotated commonsense and discourse constraints in VinaBench, to enable training the auto-regressive LLM (Llama3.1-70B-Instruct [2]) to generate those constraints. First, we merge the commonsense links into the dense image caption. Specifically, for each entity in the image caption, if it appears in one of the commonsense links, we insert its linked textual narrative phrase right after the entity (in parentheses). For example, if the image caption is *A woman wearing a green shirt*, and its entity *woman* is linked to the character *Samantha* in the textual narrative, the caption will be converted to *A woman (Samantha) wearing a green shirt*. Second, we use a template to serialize the scene features, and insert presented characters' attributes in the global features. For instance, if the scene features indicate that the presented character, time of day and location are *Samantha*, *afternoon* and *kitchen*, respectively, and *Samantha* has the profile *adult female, wife* in the global features, the scene features will be serialized into the text sequence: *[Characters] Samantha (adult female, wife) [Time of Day] afternoon [Location] kitchen*. We train the LLM to auto-regressively generate the concatenation of image caption (with commonsense links inserted) and serialized scene features, as the narrative constraints used for augmenting the visual narrative generation.

We test three representative visual narrative generation models on VinaBench, which cover diverse model structures, as described below:

- **ARLDM** [11] trains a Stable Diffusion [14] module to auto-regressively generate each visual narrative image, which is conditioned on the BLIP [6] embeddings of previous scenes' generated images and input textual con-

Model	Setting	Ranking		Non-Character		Character Number		Character Attribute		Time of Day		Location	
		CLIP-T-MRR	VQA-MRR	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava
ARLDM	w/o Constraint	0.1096	0.1435	0.5640	0.5419	0.3980	0.3858	0.3199	0.3176	0.4429	0.3942	0.3759	0.4031
	w/ LLM Cons.	**0.1508**	**0.2423**	**0.6741**	**0.6344**	0.4434	0.4345	0.4107	0.3785	**0.5119**	**0.4810**	0.5835	0.5825
	w/ Gold Cons.	0.1551	0.2503	0.6823	0.6420	0.6188	0.5607	0.5464	0.5573	0.5183	0.4945	0.6899	0.5650
StoryGen	w/o Constraint	0.1003	0.1158	0.4708	0.4707	0.3352	0.3236	0.2846	0.2167	0.2788	0.2804	0.3153	0.3791
	w/ LLM Cons.	0.1056	0.1503	0.5950	0.5764	0.4236	0.4028	0.3412	0.3191	0.3673	0.3444	0.5041	0.5460
	w/ Gold Cons.	0.1151	0.1728	0.6138	0.5873	0.5474	0.5081	0.4443	0.3749	0.3930	0.3467	0.5982	0.6325
MM-Inter.	w/o Constraint	0.0660	0.1126	0.4990	0.4856	0.4088	0.3982	0.3259	0.3265	0.4632	0.4373	0.4489	0.4713
	w/ LLM Cons.	0.1107	0.2074	0.6434	0.5942	**0.4578**	**0.4407**	**0.4118**	**0.3915**	0.4856	0.4745	**0.5998**	**0.6016**
	- w/o CL	0.1090	0.2037	0.6422	0.5934	0.4546	0.4344	0.4092	0.3870	0.4748	0.4681	0.5968	0.5944
	- w/o DS	0.1074	0.1983	0.6238	0.5872	0.4489	0.4355	0.4005	0.3887	0.4742	0.4635	0.5642	0.5734
	- Random	0.0476	0.0861	0.4149	0.4152	0.3986	0.3904	0.3180	0.3135	0.4120	0.3849	0.4116	0.4335
	w/ Gold Cons.	0.1179	0.2105	0.6521	0.6054	0.6226	0.5634	0.5462	0.5736	0.4965	0.4841	0.7276	0.7157
Gold Ref.	-	0.1586	0.2662	0.7755	0.7163	0.8127	0.7652	0.7581	0.7157	0.7555	0.7196	0.8632	0.8100

Table S5. Full evaluation results of our ranking-based and fine-grained **Alignment** metrics on VWP narratives. *MiniCPM* and *Llava* denote our fine-grained VQA-based metrics deployed on MiniCPM-V-2.6 and LLaVA-OneVision-72B. *Gold Ref.* denotes gold references. Best results under *w/ LLM Cons.* and *w/ Gold Cons.* settings are **bolded** and underlined, respectively.

Model	Setting	Style		Character		Location	
		MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava
ARLDM	w/o Constraint	0.4664	0.5857	0.3793	0.4102	0.3759	0.1788
	w/ LLM Cons.	0.8586	0.7434	0.5507	0.5215	0.6888	0.3472
	w/ Gold Cons.	0.8539	0.7326	0.5687	0.5280	0.6972	0.4359
StoryGen	w/o Constraint	0.2379	0.4936	0.2305	0.3809	0.3106	0.2020
	w/ LLM Cons.	0.4523	0.5390	0.4177	0.5014	0.4649	0.3192
	w/ Gold Cons.	0.4747	0.5421	0.4233	0.5105	0.5272	0.3800
MM-Inter.	w/o Constraint	0.9470	0.8077	0.5823	0.5631	0.4489	0.4831
	w/ LLM Cons.	**0.9859**	**0.8672**	**0.6780**	**0.6375**	**0.7642**	**0.6151**
	- w/o CL	0.9829	0.8664	0.6431	0.6290	0.7577	0.6113
	- w/o DS	0.9776	0.8604	0.6443	0.6095	0.6842	0.5937
	- Random	0.9453	0.7933	0.5763	0.5768	0.4471	0.4769
	w/ Gold Cons.	0.9764	0.8542	0.6880	0.6399	0.8558	0.6931
Gold Ref.	-	0.9706	0.8790	0.7797	0.7077	0.8632	0.7754

Table S6. Full evaluation results of our **Consistency** metrics on VWP narratives. Notations are same as Table S5.

Model	Setting	FID	CLIP-I	CLIP-T
ARLDM	w/o Constraint	42.55	0.6384	0.1951
	w/ LLM Cons.	**37.60**	**0.6762**	**0.2036**
	w/ Gold Cons.	35.25	0.7156	0.2089
StoryGen	w/o Constraint	78.58	0.5624	0.1836
	w/ LLM Cons.	52.09	0.6003	0.1935
	w/ Gold Cons.	48.93	0.6194	0.1901
MM-Inter.	w/o Constraint	48.33	0.6337	0.1758
	w/ LLM Cons.	42.24	0.6670	0.1978
	- w/o CL	42.85	0.6660	0.1966
	- w/o DS	43.28	0.6568	0.1960
	- Random	53.74	0.6143	0.1739
	w/ Gold Cons.	39.27	0.6981	0.1997
Gold Ref.	-	-	-	0.2077

Table S7. Evaluation results of full-reference metrics on VWP narratives. Lower FID is better. Notations are same as Table S5.

straints, and the CLIP [12] embedding of current scene's input textual constraints.

- **StoryGen** [8] uses a dual-diffusion structure to perform the auto-regressive generation of narrative images. It first adds noise to each previously generated image, and then the noisy image is de-noised by a Stable Diffusion module (conditioned on the image's corresponding input textual constraints), whose latent diffusion states are used as the extracted features of the image. Conditioned on the current textual constraints and the concatenation of previous images' extracted features, a second Stable Diffusion module is trained to generate the current narrative image.
- **MM-Interleaved (MM-Inter.)** [15] trains a VLM, *i.e.*, Vicuna [17] with CLIP vision encoder, to model the interleaved sequence of previously generated images and their textual constraints, and a Stable Diffusion module to generate the current narrative image based on the output states of the VLM. Both the VLM and the diffusion module are augmented by additional layers of cross-attention to sparse image features via Deformable Attention [18].

S4. Full Experimental Results

Table S5 - S13 present the full evaluation results of visual narrative generation on VinaBench. All results coherently indicate the same conclusion that learning with VinaBench's commonsense and discourse constraints significantly improves the consistency of visual narrative generations and their alignment to the input textual narrative. The coherent results on all types of VinaBench narratives imply the ubiquity of implicit knowledge constraints in visual narratives, which also indicate that our proposed knowledge augmentation framework is universally effective on various visual narrative domains and image styles.

Moreover, our two ranking-based metrics CLIP-T-MRR and VQA-MRR consistently show that all model generations and the gold reference score far below the maximum (1.0), supporting the fact that creating visual narratives is a considerably open-ended task, which does not possess the only feasible reference that always ranks the first. More importantly, our VQA-based metrics deployed on

Model	Setting	Ranking		Non-Character		Character Number		Character Attribute		Time of Day		Location	
		CLIP-T-MRR	VQA-MRR	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava
ARLDM	w/o Constraint	0.0954	0.1279	0.3487	0.3302	0.2682	0.2714	0.2330	0.2208	0.3250	0.2768	0.2987	0.3341
	w/ LLM Cons.	**0.1369**	**0.2273**	**0.6590**	**0.6233**	0.4702	**0.4330**	0.3694	0.3434	**0.4049**	0.3623	0.4899	0.5031
	w/ Gold Cons.	0.1415	0.2350	0.6745	0.6319	0.6067	0.5743	0.4804	0.4485	0.4689	0.4084	0.5994	0.6011
StoryGen	w/o Constraint	0.0926	0.1079	0.3051	0.3080	0.2908	0.2956	0.2064	0.1593	0.1599	0.1988	0.1710	0.2505
	w/ LLM Cons.	0.0992	0.1438	0.5259	0.5306	0.4304	0.3955	0.2684	0.2677	0.2754	0.2580	0.3739	0.4423
	w/ Gold Cons.	0.1078	0.1653	0.5273	0.5291	0.6629	0.5572	0.3709	0.3636	0.2950	0.2686	0.4281	0.4883
MM-Inter.	w/o Constraint	0.0521	0.0979	0.3286	0.3264	0.2616	0.2323	0.2290	0.1956	0.3311	0.3042	0.2294	0.2588
	w/ LLM Cons.	0.0983	0.1935	0.6030	0.5702	**0.4767**	0.4329	**0.3796**	**0.3449**	0.3733	**0.3686**	**0.4971**	**0.5170**
	w/ Gold Cons.	0.1049	0.1959	0.6375	0.5786	0.6132	0.5746	0.4877	0.4456	0.4231	0.4155	0.6167	0.6224
Gold Ref.	-	0.1657	0.2735	0.7630	0.7118	0.8682	0.8375	0.7981	0.7156	0.7620	0.7114	0.8955	0.7879

Table S8. Full zero-shot evaluation results of our ranking-based and fine-grained **Alignment** metrics on Storyboard20K narratives. Notations are same as Table S5.

Model	Setting	Style		Character		Location	
		MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava
ARLDM	w/o Constraint	0.2279	0.2089	0.2459	0.2373	0.0879	0.1133
	w/ LLM Cons.	0.6477	0.6140	0.5113	0.4531	0.3047	0.2686
	w/ Gold Cons.	0.6968	0.6167	0.5997	0.5031	0.4219	0.3679
StoryGen	w/o Constraint	0.2671	0.2613	0.0645	0.1033	0.2259	0.2108
	w/ LLM Cons.	0.5209	0.5795	0.3156	0.3494	0.3526	0.3395
	w/ Gold Cons.	0.5259	0.5848	0.3688	0.4108	0.4465	0.3975
MM-Inter.	w/o Constraint	0.8766	0.8594	0.3627	0.3667	0.4207	0.3374
	w/ LLM Cons.	**0.9324**	**0.9016**	**0.6187**	**0.5694**	**0.6958**	**0.6378**
	w/ Gold Cons.	0.9349	0.9047	0.6598	0.6283	0.7956	0.7310
Gold Ref.	-	0.9399	0.8556	0.8118	0.7665	0.8955	0.7996

Table S9. Full zero-shot evaluation results of our **Consistency** metrics on Storyboard20K narratives. Notations are same as Table S5.

Model	Setting	FID	CLIP-I	CLIP-T
ARLDM	w/o Constraint	97.91	0.5910	0.1936
	w/ LLM Cons.	**82.64**	**0.6395**	**0.1995**
	w/ Gold Cons.	77.70	0.6754	0.2057
StoryGen	w/o Constraint	161.41	0.5367	0.1690
	w/ LLM Cons.	112.03	0.5832	0.1880
	w/ Gold Cons.	107.67	0.5966	0.1837
MM-Inter.	w/o Constraint	102.42	0.5876	0.1644
	w/ LLM Cons.	95.73	0.6362	0.1893
	w/ Gold Cons.	90.82	0.6587	0.1933
Gold Ref.	-	-	-	0.2049

Table S10. Evaluation results of full-reference metrics on Storyboard20K narratives. Lower FID is better. Notations are same as Table S5.

MiniCPM-V-2.6 and LLaVA-OneVision-72B demonstrate mostly aligned preference among different models and settings. This verifies that our proposed metrics are not biased on the preference of a specific VLM used for VQA scoring.

We more closely study the correlation of our automatic evaluation metrics to the five human evaluation metrics. In particular, we consider the average of our fine-grained alignment and consistency metrics, denoted as **Alignment** and **Consistency**, and compare them to the CLIP-based metrics CLIP-I and CLIP-T. For each pair of human and automatic evaluation metrics, we calculate their Pearson correlation coefficient based on their scoring of four model[2] generations and gold references, on 100 VWP testing samples. Figure S1 presents the results of our correlation study. Compared to CLIP-I and CLIP-T, Alignment and Consistency metrics demonstrate overall better correlation with human evaluation, verifying that our proposed VQA-based evaluation gives more reliable results than CLIP-based similarity measure. We also find that the evaluation of alignment and consistency are closely correlated with each other, *e.g.*, our Alignment metric shows the highest correlation with Text-Image Alignment, while also possesses fairly high correlation with Style, Content and Character Consistency in human evaluation. This indicates that the faithfulness and self-consistency of visual narrative generation are not mutually independent, and therefore may benefit from the joint learning of these two aspects.

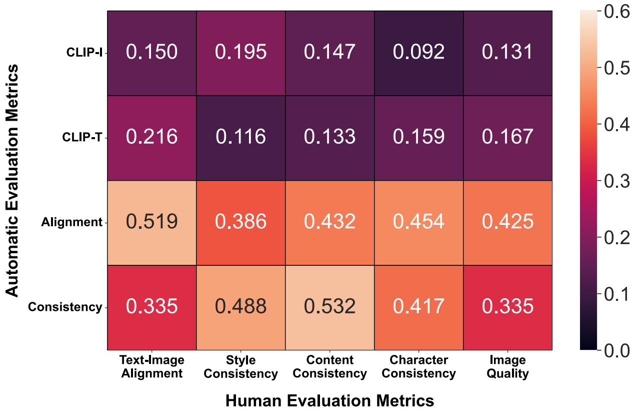


Figure S1. Pearson correlation coefficients between human and automatic evaluation metrics on VWP narratives. **Alignment** and **Consistency** in automatic evaluation metrics denote the average of our VQA-based fine-grained alignment and consistency metrics, respectively, rooted on MiniCPM-V-2.6.

[2] We consider the four models studied in the human evaluation, *i.e.*, ARLDM with and without LLM constraints, and MM-Interleaved with and without LLM constraints.

Model	Setting	Ranking		Non-Character		Character Number		Character Attribute		Time of Day		Location	
		CLIP-T-MRR	VQA-MRR	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava
ARLDM	w/o Constraint	0.1015	0.1367	0.4706	0.6048	0.2878	0.2884	0.1666	0.1764	0.4045	0.4135	0.3802	0.4041
	w/ LLM Cons.	0.1428	0.2328	**0.5685**	**0.6432**	0.3065	0.3118	0.2217	0.2787	0.4409	0.4345	0.4420	0.4468
	w/ Gold Cons.	0.1493	0.2417	0.5771	0.6519	0.3568	0.3386	0.2676	0.2984	0.4894	0.4474	0.4862	0.4839
StoryGen	w/o Constraint	0.1010	0.1347	0.4536	0.5257	0.2825	0.2851	0.1651	0.1738	0.3965	0.4043	0.3735	0.3961
	w/ LLM Cons.	**0.1443**	**0.2348**	0.5633	0.6037	0.3070	0.3148	0.2079	0.2559	0.4225	0.4267	0.3971	0.4205
	w/ Gold Cons.	0.1469	0.2410	0.5714	0.6160	0.3515	0.3376	0.2577	0.2907	0.4690	0.4400	0.4385	0.4609
MM-Inter.	w/o Constraint	0.0581	0.1062	0.4477	0.4983	0.2917	0.2946	0.1840	0.2188	0.4227	0.4158	0.3743	0.3714
	w/ LLM Cons.	0.1065	0.2015	0.5352	0.5853	**0.3662**	**0.3847**	**0.2645**	**0.2903**	**0.4727**	**0.4481**	**0.4587**	**0.4536**
	w/ Gold Cons.	0.1124	0.2032	0.5450	0.5986	0.4126	0.4242	0.3125	0.3238	0.5030	0.4624	0.5609	0.5375
Gold Ref.	-	0.1601	0.2688	0.7584	0.7432	0.8171	0.8061	0.7780	0.7655	0.7545	0.7728	0.7523	0.7635

Table S11. Full evaluation results of our ranking-based and fine-grained **Alignment** metrics on StorySalon narratives. Notations are same as Table S5.

Model	Setting	Style		Character		Location	
		MiniCPM	Llava	MiniCPM	Llava	MiniCPM	Llava
ARLDM	w/o Constraint	0.5000	0.4824	0.1461	0.1839	0.2903	0.2596
	w/ LLM Cons.	0.6563	0.5684	0.2622	0.2551	0.3296	0.2978
	w/ Gold Cons.	0.6875	0.5770	0.2890	0.2672	0.3839	0.3257
StoryGen	w/o Constraint	0.4246	0.4197	0.1041	0.1362	0.2265	0.2205
	w/ LLM Cons.	0.6073	0.5583	0.2886	0.2793	0.3191	0.2784
	w/ Gold Cons.	0.6472	0.5609	0.2911	0.2826	0.3745	0.3147
MM-Inter.	w/o Constraint	0.9450	0.8668	0.3349	0.4086	0.7022	0.6232
	w/ LLM Cons.	**0.9563**	**0.8747**	**0.3545**	**0.4449**	**0.7798**	**0.6978**
	w/ Gold Cons.	0.9688	0.8786	0.3834	0.4737	0.8034	0.7617
Gold Ref.	-	0.9688	0.9865	0.7686	0.7611	0.8135	0.8059

Table S12. Full evaluation results of our **Consistency** metrics on StorySalon narratives. Notations are same as Table S5.

Model	Setting	FID	CLIP-I	CLIP-T
ARLDM	w/o Constraint	64.69	0.6278	0.1975
	w/ LLM Cons.	56.65	0.6515	0.2001
	w/ Gold Cons.	56.51	0.6887	0.2022
StoryGen	w/o Constraint	63.63	0.6463	0.1946
	w/ LLM Cons.	**56.18**	**0.6600**	**0.2005**
	w/ Gold Cons.	55.62	0.6919	0.2021
MM-Inter.	w/o Constraint	74.92	0.6370	0.1834
	w/ LLM Cons.	72.91	0.6552	0.1879
	w/ Gold Cons.	72.03	0.6780	0.1896
Gold Ref.	-	-	-	0.2065

Table S13. Evaluation results of full-reference metrics on StorySalon narratives. Lower FID is better. Notations are same as Table S5.

Besides of MM-Interleaved, which is the best-performed model fine-tuned on VinaBench, we further test other similar interleaved image-text generative models, including **Anole** [1] and **Lumina-mGPT** [9], which however completely fail our benchmark task (with nearly zero scores on VinaBench metrics) under zero-shot or few-shot settings.[3] This indicates that supervised learning (or fine-tuning) is necessary for current interleaved image-text generative models to address our benchmark's challenging task, while the fine-tuning codes of these models are not publicly available, which hinders more experimental verifications.

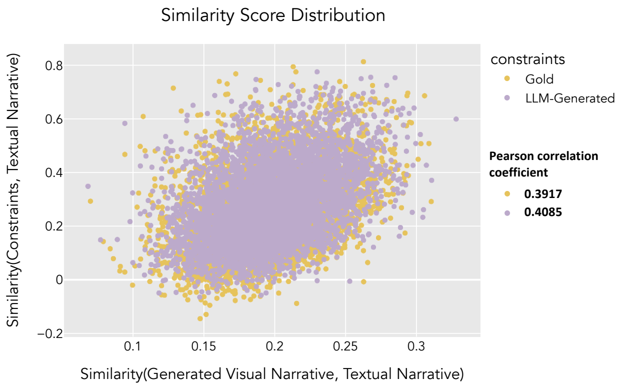


Figure S2. Correlation between generated visual narrative images and augmented narrative constraints (either from gold labels or generated by LLM, Llama3.1-70B-Instruct), w.r.t. their CLIP embedding similarity to the input textual narrative. Data samples are from MM-Interleaved generations (w/ LLM Cons. and w/ Gold Cons.) on VWP narratives.

Figure S2 shows the distribution of paired similarity scores in our correlation study between visual narrative generation and constraints, where the x-axis denotes the CLIP similarity between each visual generation and input textual narrative, and the y-axis denotes the CLIP similarity between the sample's augmented constraints and the textual narrative. The distribution demonstrates a clear positive correlation between the narrative constraints and their resulting visual narrative generations, with ~ 0.4 Pearson correlation coefficient, no matter whether the constraints are from gold labels or generated by LLM. This highlights the importance of planning faithful storytelling constraints to advance visual narrative generations.

We also evaluate MM-Interleaved model on varied settings of using LLMs to generate narrative constraints (w/ LLM Cons.), including 4-shot (**4S**) prompting Llama3.1-

[3] We verify that MM-Interleaved model would also fail our benchmark task under zero/few-shot settings, *i.e.*, without fine-tuning.

w/ LLM Cons.	FID	CLIP-I	CLIP-T	CLIP-T-MRR	Alignment	Consistency
FT Llama-70B	42.24	0.6670	0.1978	0.1107	0.5197	0.8093
4S Llama-70B	42.95	0.6625	0.1973	0.1104	0.4948	0.7936
FT Llama-8B	49.61	0.6293	0.1833	0.0570	0.3980	0.7436
FT Gemma-7B	51.69	0.6180	0.1788	0.0445	0.3751	0.7312
FT Qwen2-7B	47.83	0.6376	0.1915	0.0866	0.4507	0.7606
Gold Ref.	-	-	0.2077	0.1586	0.7930	0.8711

Table S14. Performance of MM-Interleaved model with different LLM-generated narrative constraints, evaluated on VWP narratives. Llama3.1-70B-Instruct (Llama-70B) is fine-tuned (FT) with LoRA or 4-shot (4S) prompted, while Llama3.1-8B-Instruct (Llama-8B), Gemma-7B and Qwen2-7B are fully fine-tuned. **Alignment** and **Consistency** denote the average score of our proposed fine-grained alignment and consistency metrics.

70B-Instruct (**Llama-70B**), and fine-tuning (**FT**) Llama3.1-8B-Instruct (**Llama-8B**), **Gemma-7B** and **Qwen2-7B**, compared to our adopted setting of fine-tuning Llama3.1-70B-Instruct with LoRA. Results in Table S14, based on the VWP narratives of VinaBench, show that our adopted setting best augments visual narrative generation.

Figure S3 displays several visual narratives generated by our deployed baseline methods. The model generations still contain unfaithful or inconsistent contents, even with the augmentation of narrative constraints. This reveals the challenge of developing more robust methods for the visual narrative generation, which we leave for future work.

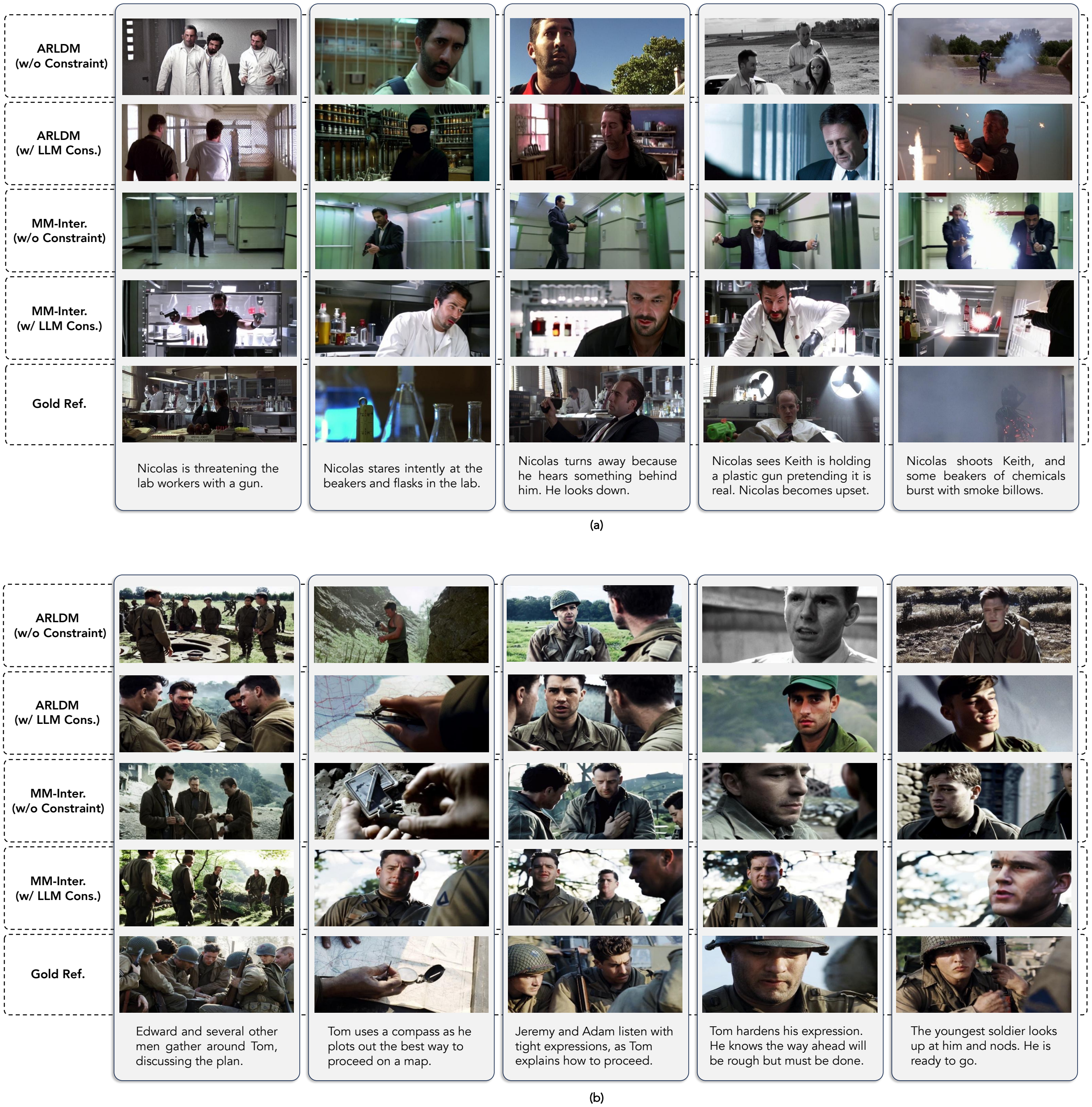
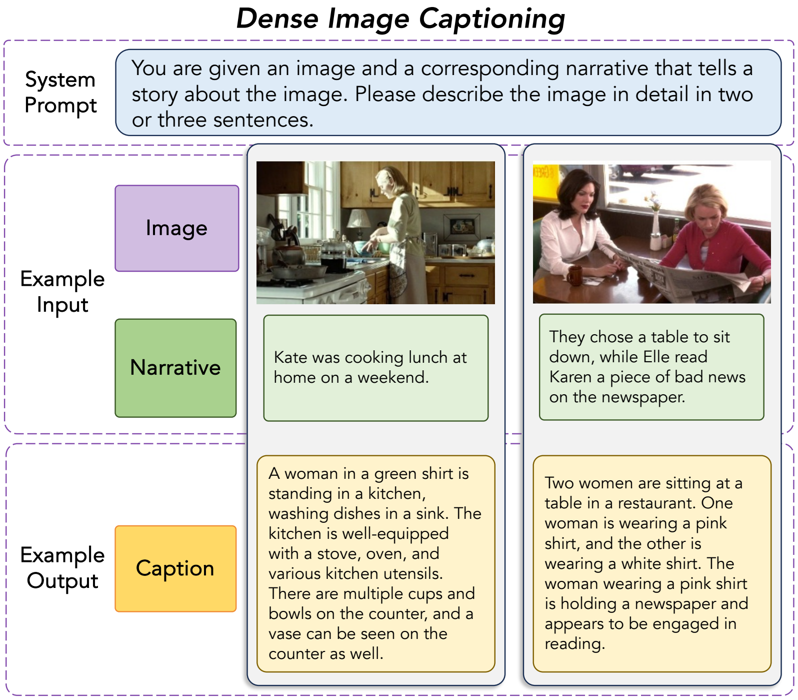

ARLDM (w/o Constraint)
ARLDM (w/ LLM Cons.)
MM-Inter. (w/o Constraint)
MM-Inter. (w/ LLM Cons.)
Gold Ref.
Nicolas is threatening the lab workers with a gun.
Nicolas stares intently at the beakers and flasks in the lab.
Nicolas turns away because he hears something behind him. He looks down.
Nicolas sees Keith is holding a plastic gun pretending it is real. Nicolas becomes upset.
Nicolas shoots Keith, and some beakers of chemicals burst with smoke billows.
(a)
ARLDM (w/o Constraint)
ARLDM (w/ LLM Cons.)
MM-Inter. (w/o Constraint)
MM-Inter. (w/ LLM Cons.)
Gold Ref.
Edward and several other men gather around Tom, discussing the plan.
Tom uses a compass as he plots out the best way to proceed on a map.
Jeremy and Adam listen with tight expressions, as Tom explains how to proceed.
Tom hardens his expression. He knows the way ahead will be rough but must be done.
The youngest soldier looks up at him and nods. He is ready to go.
(b)

Figure S3. Visual narratives generated by ARLDM and MM-Interleaved (MM-Inter.), with and without LLM-generated narrative constraints, compared to the gold reference. In narrative (a), LLM-generated constraints significantly improve MM-Interleaved, by pushing its generation more aligned with the lab setting described in textual narrative. By contrast, ARLDM fails to generate images with decent alignment to textual narrative, although the image style consistency is improved by LLM constraints, *e.g.*, avoid generating a black and white image at the fourth scene. In narrative (b), the generation of ARLDM with LLM constraints turns out to achieve improved image style consistency and alignment to textual narrative plot, *e.g.*, showing a map in the second scene. Besides, compared to MM-Interleaved without constraint, the generation of MM-Interleaved with LLM constraints displays better consistency of character (*e.g.*, Tom) facial features and background location, and comparable faithfulness to textual narrative. However, both model generations with constraints still contain unreasonable contents, *e.g.*, a sudden shift of character Nicolas's outfit in the generation of MM-Interleaved (w/ LLM Cons.) in (a), inconsistent faces of character Tom in the ARLDM (w/ LLM Cons.) generation in (b).



Dense Image Captioning

System Prompt

You are given an image and a corresponding narrative that tells a story about the image. Please describe the image in detail in two or three sentences.

Example Input

Image

Narrative

Kate was cooking lunch at home on a weekend.

They chose a table to sit down, while Elle read Karen a piece of bad news on the newspaper.

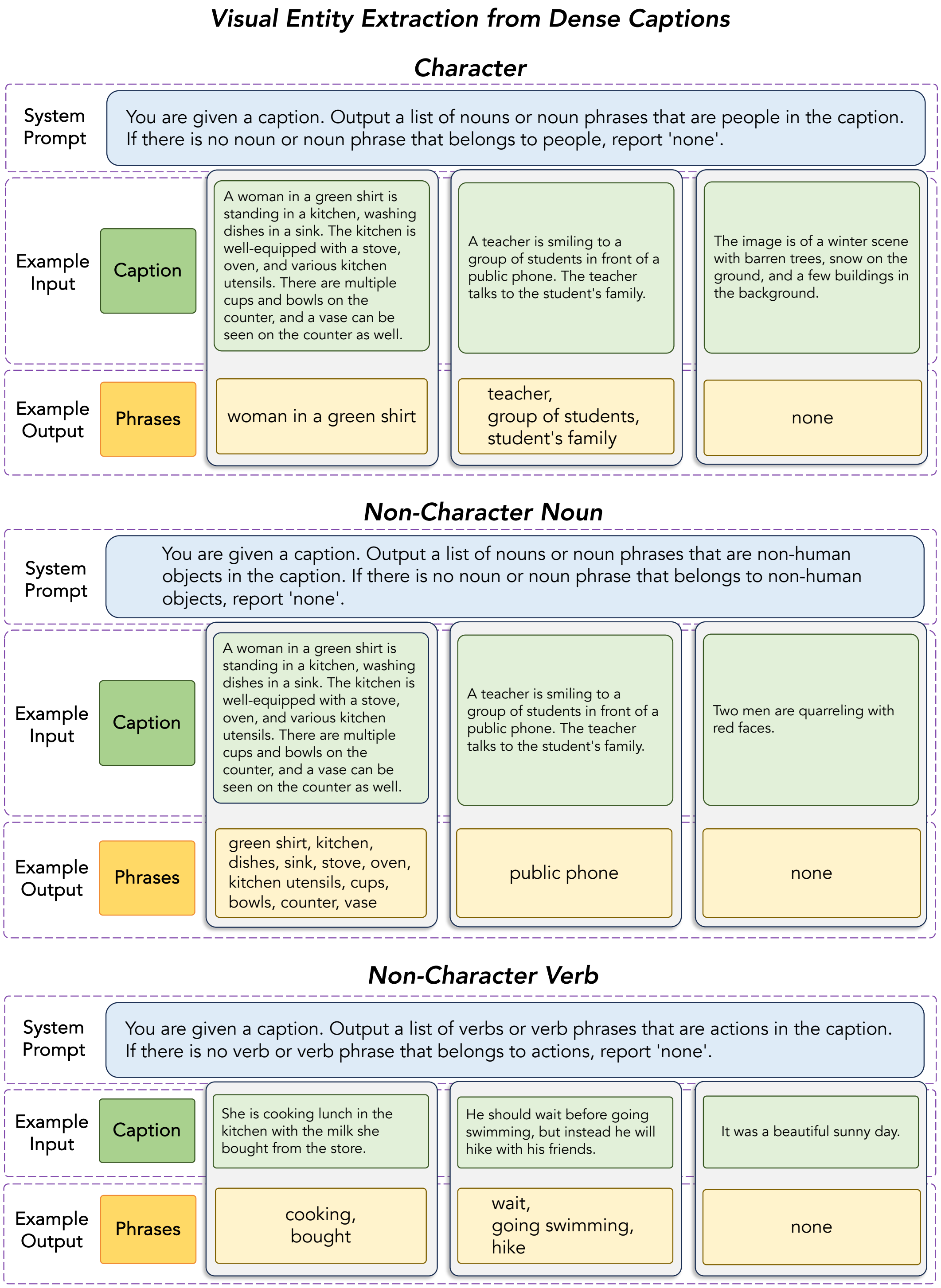
Example Output

Caption

A woman in a green shirt is standing in a kitchen, washing dishes in a sink. The kitchen is well-equipped with a stove, oven, and various kitchen utensils. There are multiple cups and bowls on the counter, and a vase can be seen on the counter as well.

Two women are sitting at a table in a restaurant. One woman is wearing a pink shirt, and the other is wearing a white shirt. The woman wearing a pink shirt is holding a newspaper and appears to be engaged in reading.

Figure S4. Few-shot prompting demonstrations for constructing the dense **image captions** in VinaBench.



Visual Entity Extraction from Dense Captions

Character

You are given a caption. Output a list of nouns or noun phrases that are people in the caption. If there is no noun or noun phrase that belongs to people, report 'none'.

Example Input – Caption: A woman in a green shirt is standing in a kitchen, washing dishes in a sink. The kitchen is well-equipped with a stove, oven, and various kitchen utensils. There are multiple cups and bowls on the counter, and a vase can be seen on the counter as well.

Example Output – Phrases: woman in a green shirt

Example Input – Caption: A teacher is smiling to a group of students in front of a public phone. The teacher talks to the student's family.

Example Output – Phrases: teacher, group of students, student's family

Example Input – Caption: The image is of a winter scene with barren trees, snow on the ground, and a few buildings in the background.

Example Output – Phrases: none

Non-Character Noun

You are given a caption. Output a list of nouns or noun phrases that are non-human objects in the caption. If there is no noun or noun phrase that belongs to non-human objects, report 'none'.

Example Input – Caption: A woman in a green shirt is standing in a kitchen, washing dishes in a sink. The kitchen is well-equipped with a stove, oven, and various kitchen utensils. There are multiple cups and bowls on the counter, and a vase can be seen on the counter as well.

Example Output – Phrases: green shirt, kitchen, dishes, sink, stove, oven, kitchen utensils, cups, bowls, counter, vase

Example Input – Caption: A teacher is smiling to a group of students in front of a public phone. The teacher talks to the student's family.

Example Output – Phrases: public phone

Example Input – Caption: Two men are quarreling with red faces.

Example Output – Phrases: none

Non-Character Verb

You are given a caption. Output a list of verbs or verb phrases that are actions in the caption. If there is no verb or verb phrase that belongs to actions, report 'none'.

Example Input – Caption: She is cooking lunch in the kitchen with the milk she bought from the store.

Example Output – Phrases: cooking, bought

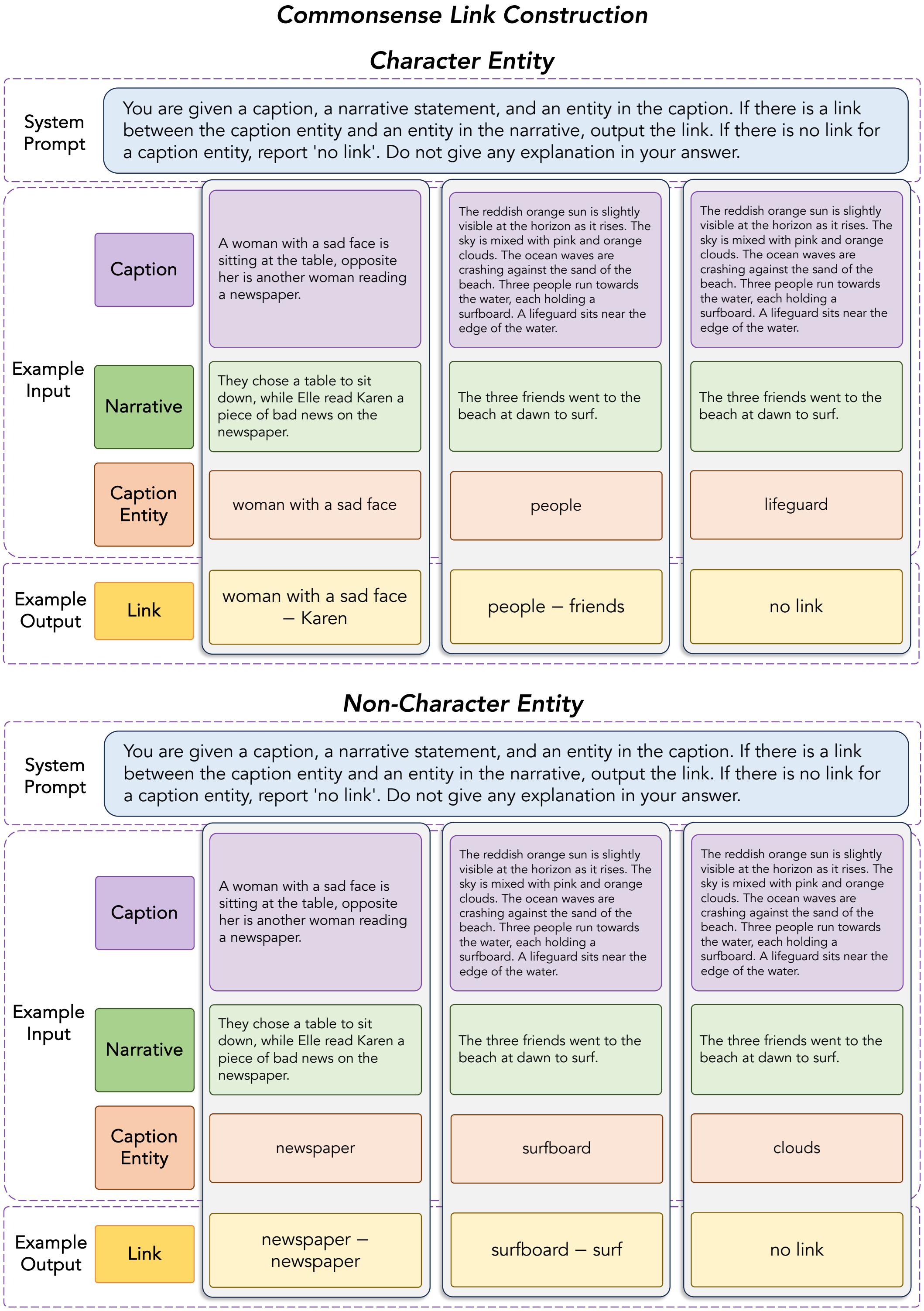
Example Input – Caption: He should wait before going swimming, but instead he will hike with his friends.

Example Output – Phrases: wait, going swimming, hike

Example Input – Caption: It was a beautiful sunny day.

Example Output – Phrases: none

(a)



Commonsense Link Construction

Character Entity

System Prompt: You are given a caption, a narrative statement, and an entity in the caption. If there is a link between the caption entity and an entity in the narrative, output the link. If there is no link for a caption entity, report 'no link'. Do not give any explanation in your answer.

Example Input	Caption	A woman with a sad face is sitting at the table, opposite her is another woman reading a newspaper.	The reddish orange sun is slightly visible at the horizon as it rises. The sky is mixed with pink and orange clouds. The ocean waves are crashing against the sand of the beach. Three people run towards the water, each holding a surfboard. A lifeguard sits near the edge of the water.	The reddish orange sun is slightly visible at the horizon as it rises. The sky is mixed with pink and orange clouds. The ocean waves are crashing against the sand of the beach. Three people run towards the water, each holding a surfboard. A lifeguard sits near the edge of the water.
	Narrative	They chose a table to sit down, while Elle read Karen a piece of bad news on the newspaper.	The three friends went to the beach at dawn to surf.	The three friends went to the beach at dawn to surf.
	Caption Entity	woman with a sad face	people	lifeguard
Example Output	Link	woman with a sad face — Karen	people — friends	no link

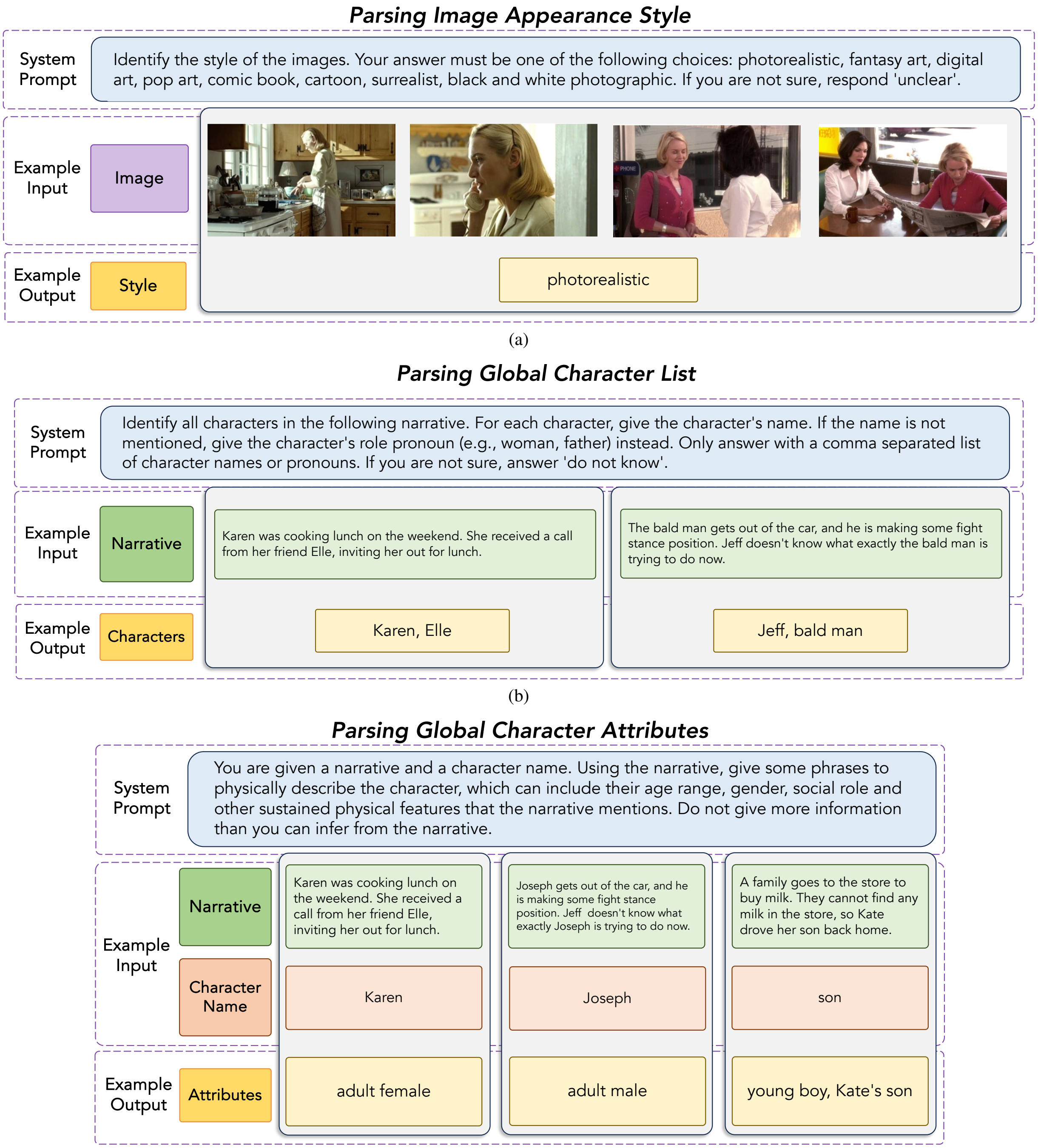
Non-Character Entity

System Prompt: You are given a caption, a narrative statement, and an entity in the caption. If there is a link between the caption entity and an entity in the narrative, output the link. If there is no link for a caption entity, report 'no link'. Do not give any explanation in your answer.

Example Input	Caption	A woman with a sad face is sitting at the table, opposite her is another woman reading a newspaper.	The reddish orange sun is slightly visible at the horizon as it rises. The sky is mixed with pink and orange clouds. The ocean waves are crashing against the sand of the beach. Three people run towards the water, each holding a surfboard. A lifeguard sits near the edge of the water.	The reddish orange sun is slightly visible at the horizon as it rises. The sky is mixed with pink and orange clouds. The ocean waves are crashing against the sand of the beach. Three people run towards the water, each holding a surfboard. A lifeguard sits near the edge of the water.
	Narrative	They chose a table to sit down, while Elle read Karen a piece of bad news on the newspaper.	The three friends went to the beach at dawn to surf.	The three friends went to the beach at dawn to surf.
	Caption Entity	newspaper	surfboard	clouds
Example Output	Link	newspaper — newspaper	surfboard — surf	no link

(b)

Figure S5. Few-shot prompting demonstrations for constructing the **commonsense links** in VinaBench, including (a) visual entity extraction (w.r.t. character, non-character noun and verb), and (b) link construction (w.r.t. each extracted character and non-character entity).



Parsing Image Appearance Style

System Prompt

Identify the style of the images. Your answer must be one of the following choices: photorealistic, fantasy art, digital art, pop art, comic book, cartoon, surrealist, black and white photographic. If you are not sure, respond 'unclear'.

Example Input: Image

Example Output: Style

photorealistic

(a)

Parsing Global Character List

System Prompt

Identify all characters in the following narrative. For each character, give the character's name. If the name is not mentioned, give the character's role pronoun (e.g., woman, father) instead. Only answer with a comma separated list of character names or pronouns. If you are not sure, answer 'do not know'.

Example Input: Narrative

Example Output: Characters

Karen was cooking lunch on the weekend. She received a call from her friend Elle, inviting her out for lunch.

Karen, Elle

The bald man gets out of the car, and he is making some fight stance position. Jeff doesn't know what exactly the bald man is trying to do now.

Jeff, bald man

(b)

Parsing Global Character Attributes

System Prompt

You are given a narrative and a character name. Using the narrative, give some phrases to physically describe the character, which can include their age range, gender, social role and other sustained physical features that the narrative mentions. Do not give more information than you can infer from the narrative.

Example Input: Narrative; Character Name

Example Output: Attributes

Karen was cooking lunch on the weekend. She received a call from her friend Elle, inviting her out for lunch.

Karen

adult female

Joseph gets out of the car, and he is making some fight stance position. Jeff doesn't know what exactly Joseph is trying to do now.

Joseph

adult male

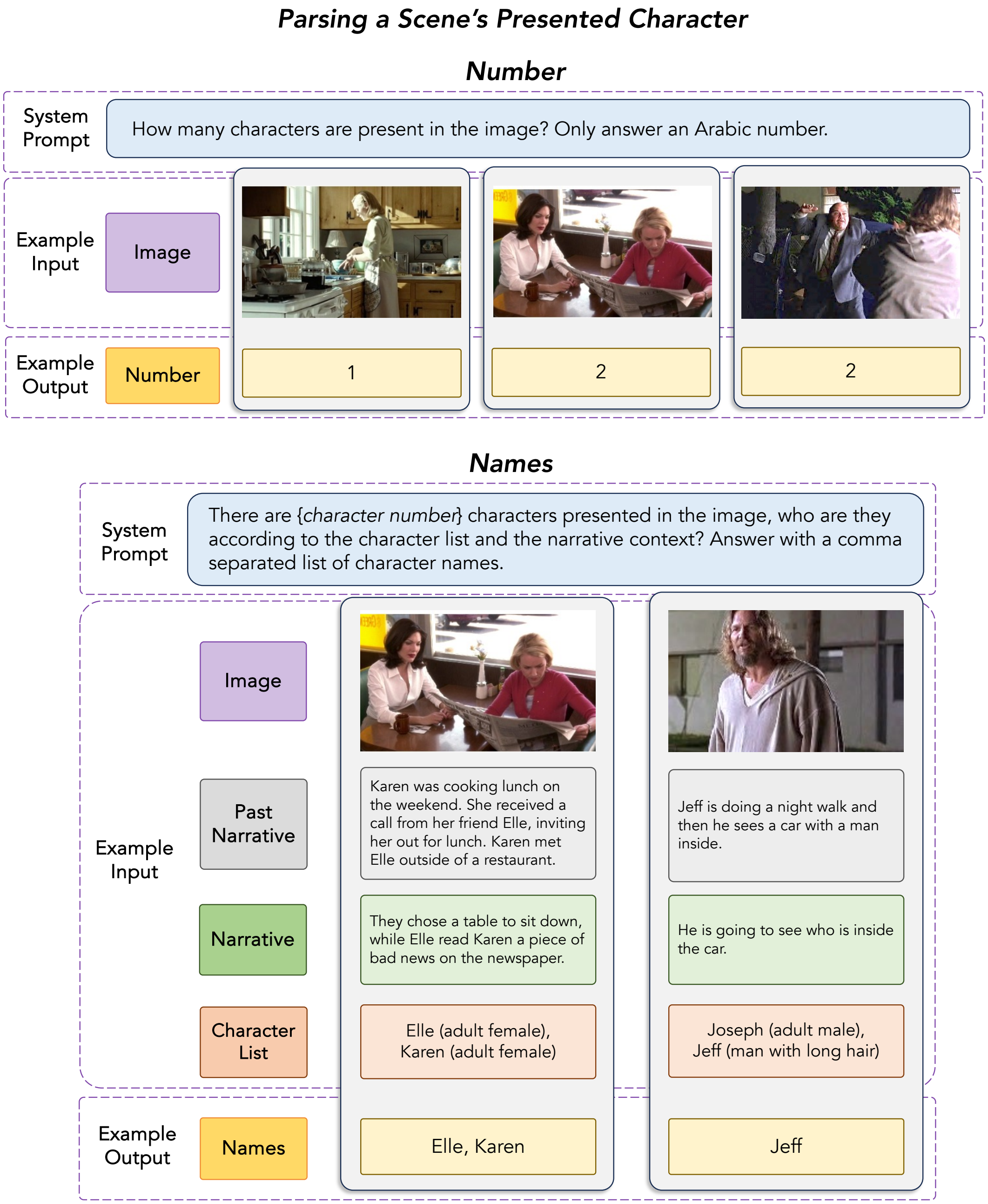
A family goes to the store to buy milk. They cannot find any milk in the store, so Kate drove her son back home.

son

young boy, Kate's son

(c)

Figure S6. Few-shot prompting demonstrations for parsing the **global features** in VinaBench, including (a) image appearance style, (b) character list, and (c) character attributes. The output features of (b) and (c) form the global profile of characters.



Parsing a Scene's Presented Character

Number

System Prompt

How many characters are present in the image? Only answer an Arabic number.

Example Input: Image

Example Output: Number

1

2

2

Names

System Prompt

There are {*character number*} characters presented in the image, who are they according to the character list and the narrative context? Answer with a comma separated list of character names.

Example Input:

Image

Past Narrative

Narrative

Character List

Karen was cooking lunch on the weekend. She received a call from her friend Elle, inviting her out for lunch. Karen met Elle outside of a restaurant.

They chose a table to sit down, while Elle read Karen a piece of bad news on the newspaper.

Elle (adult female), Karen (adult female)

Jeff is doing a night walk and then he sees a car with a man inside.

He is going to see who is inside the car.

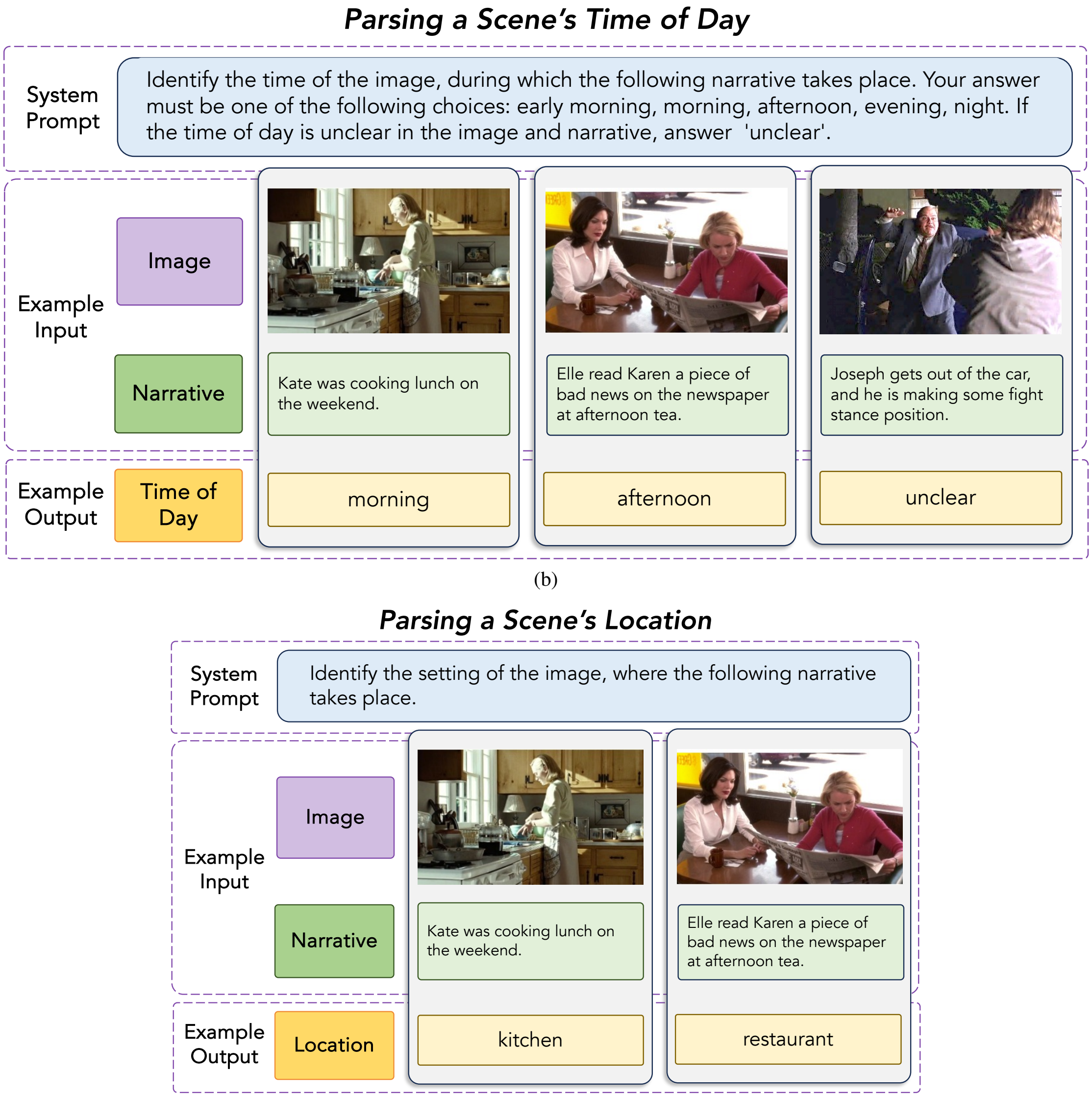
Joseph (adult male), Jeff (man with long hair)

Example Output: Names

Elle, Karen

Jeff

(a)



Parsing a Scene's Time of Day

System Prompt

Identify the time of the image, during which the following narrative takes place. Your answer must be one of the following choices: early morning, morning, afternoon, evening, night. If the time of day is unclear in the image and narrative, answer 'unclear'.

Example Input: Image, Narrative

Kate was cooking lunch on the weekend.

Elle read Karen a piece of bad news on the newspaper at afternoon tea.

Joseph gets out of the car, and he is making some fight stance position.

Example Output: Time of Day

morning

afternoon

unclear

(b)

Parsing a Scene's Location

System Prompt

Identify the setting of the image, where the following narrative takes place.

Example Input: Image, Narrative

Kate was cooking lunch on the weekend.

Elle read Karen a piece of bad news on the newspaper at afternoon tea.

Example Output: Location

kitchen

restaurant

(c)

Figure S7. Few-shot prompting demonstrations for parsing the **scene features** in VinaBench, including (a) presented character number and names, (b) time of day, and (c) location. In the step of parsing presented character names in (a), the span "{*character number*}" in the system prompt is replaced by the output in the prior step of parsing character number.

References

[1] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024. 5

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[3] Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581, 2023. 1

[4] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 1

[5] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 1

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[7] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338, 2019. 2

[8] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6190–6200, 2024. 1, 3

[9] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 5

[10] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*, 2021. 2

[11] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2920–2930, 2024. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[13] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123:94–120, 2017. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[15] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 3

[16] Jinheng Xie, Jiajun Feng, Zhaoxu Tian, Kevin Qinghong Lin, Yawen Huang, Xi Xia, Nanxu Gong, Xu Zuo, Jiaqi Yang, Yefeng Zheng, et al. Learning long-form video prior via generative pre-training. *arXiv preprint arXiv:2404.15909*, 2024. 1

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 3

[18] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3