Emanuele Vivoli,
Marco Bertini,
Dimosthenis Karatzas

CVC Computer Vision Center
MICC

CoMix : A comprehensive Benchmark for Multi-Task Comic Understanding

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

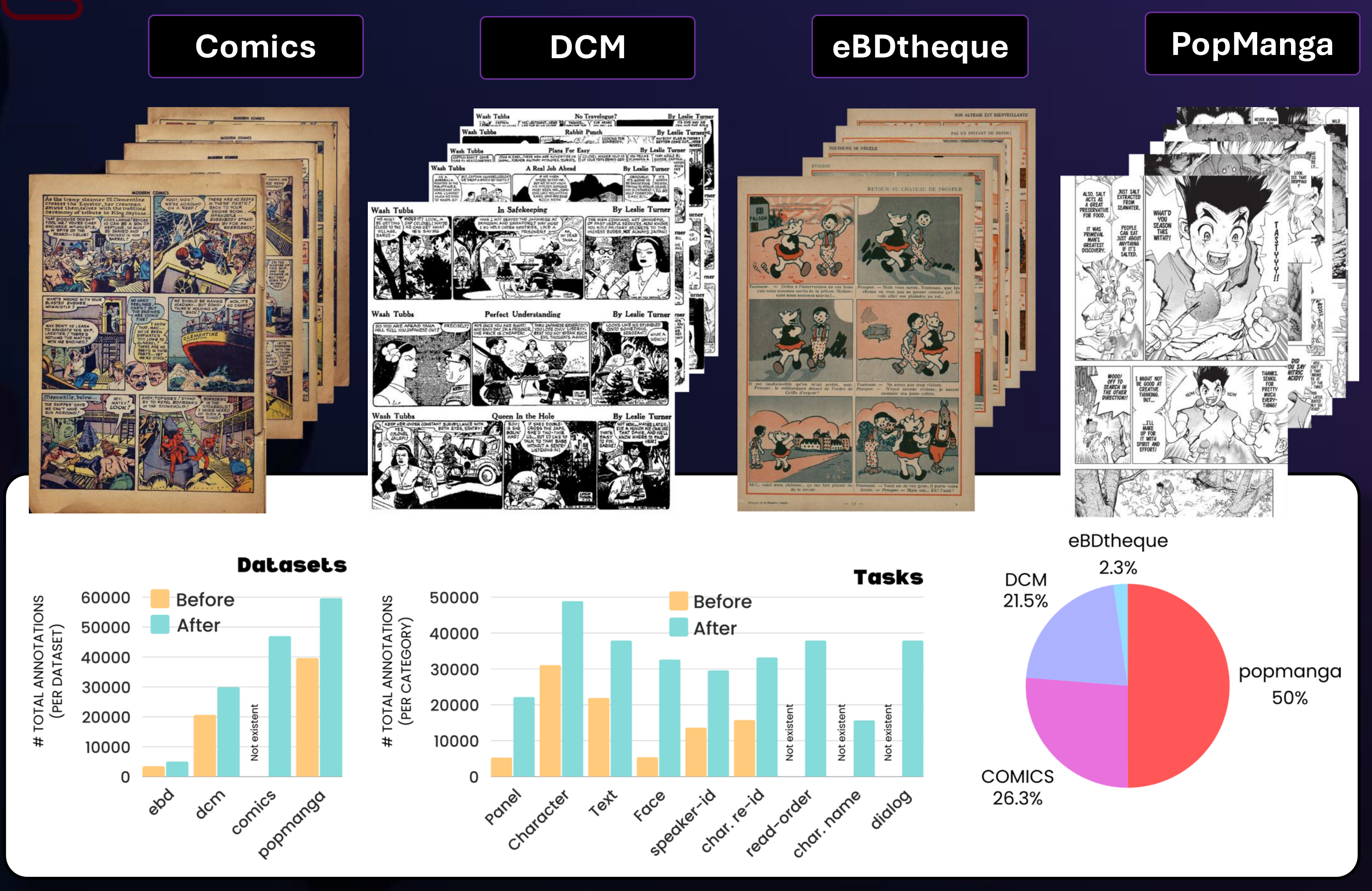
Comics datasets only tackle **simple tasks**, are **small** or of a **single style** (manga vs. comics)

Dataset	Release	Avail	Tasks	Years	Style	Books	Pages
eBDtheque	2013	✓	d,t2c	1905-2012	mix	28	100
COMICS	2017	✓	c	1938-1954	comics	3948	198k
GCN	2017	✗	d,t2c	1978-2013	comics	*253	*38k
DCM772	2018	✓	d	1938-1954	comics	27	772
Manga109	2018	✓	d,t2c,c2c	1970-2010	manga	109	10k
BCBId	2022	✓	-	-	bangla	64	3k
VLRC	2023	✗	-	1940-now	-	*376	*7k
PopManga	2024	✓	d,t2c,c2c	2010-2023	manga	25	1.8k
CoMix (our)	2024	✓	d,t2c,c2c,N,D	1938-2023	mix	100	3.8k

We create a densely annotated dataset: ***CoMix***

Legend			
d: Detection	t2c: Text-to-Character	N: Character naming
c: Classification	c2c: Char-to-Char	D: Dialog generation

Design choices



Annotations

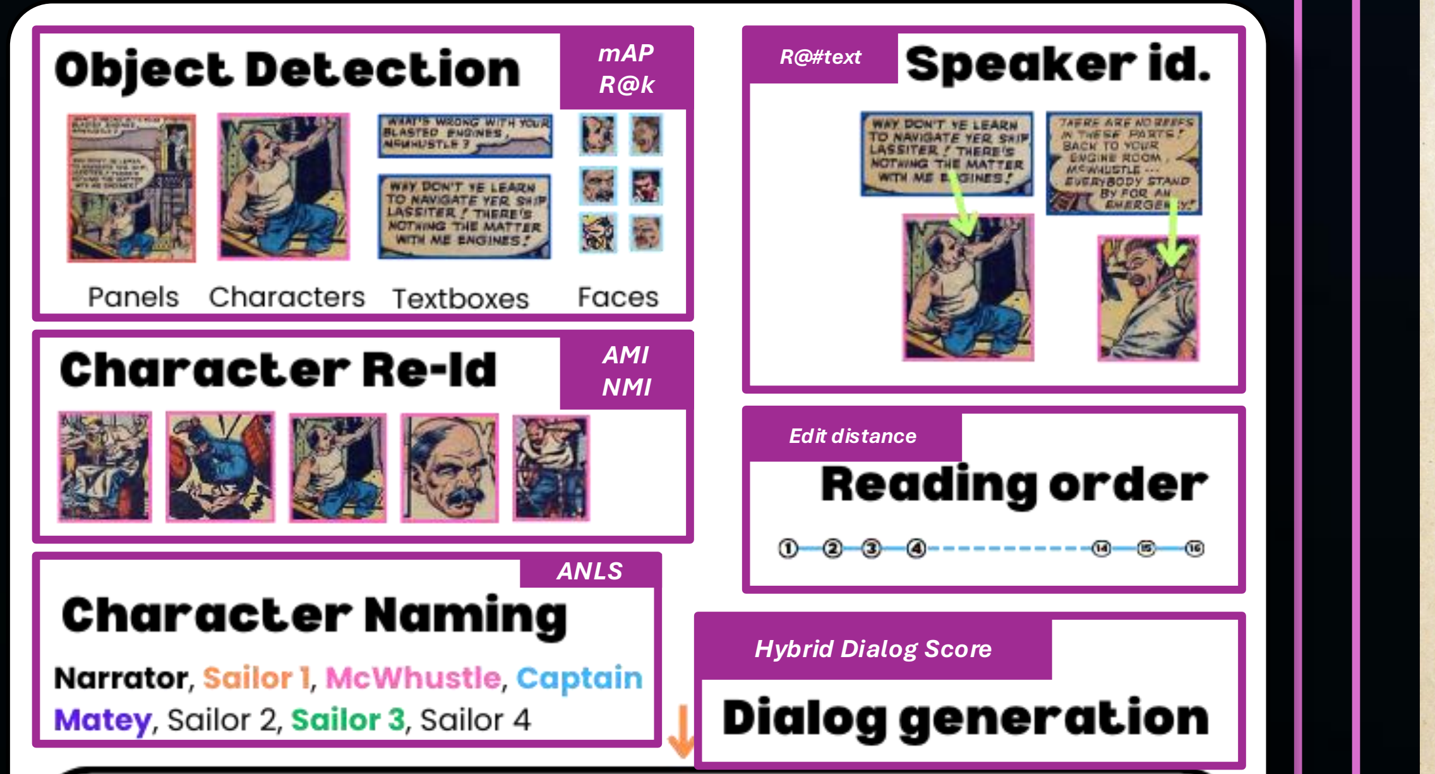
We improve the quality of existing annotations

And provide annotations for new tasks

BEFORE
AFTER
BEFORE
AFTER
BEFORE
AFTER

Tasks

WHAT'S WRONG WITH YOUR BLASTED ENGINES, MCWHUSTLE ?

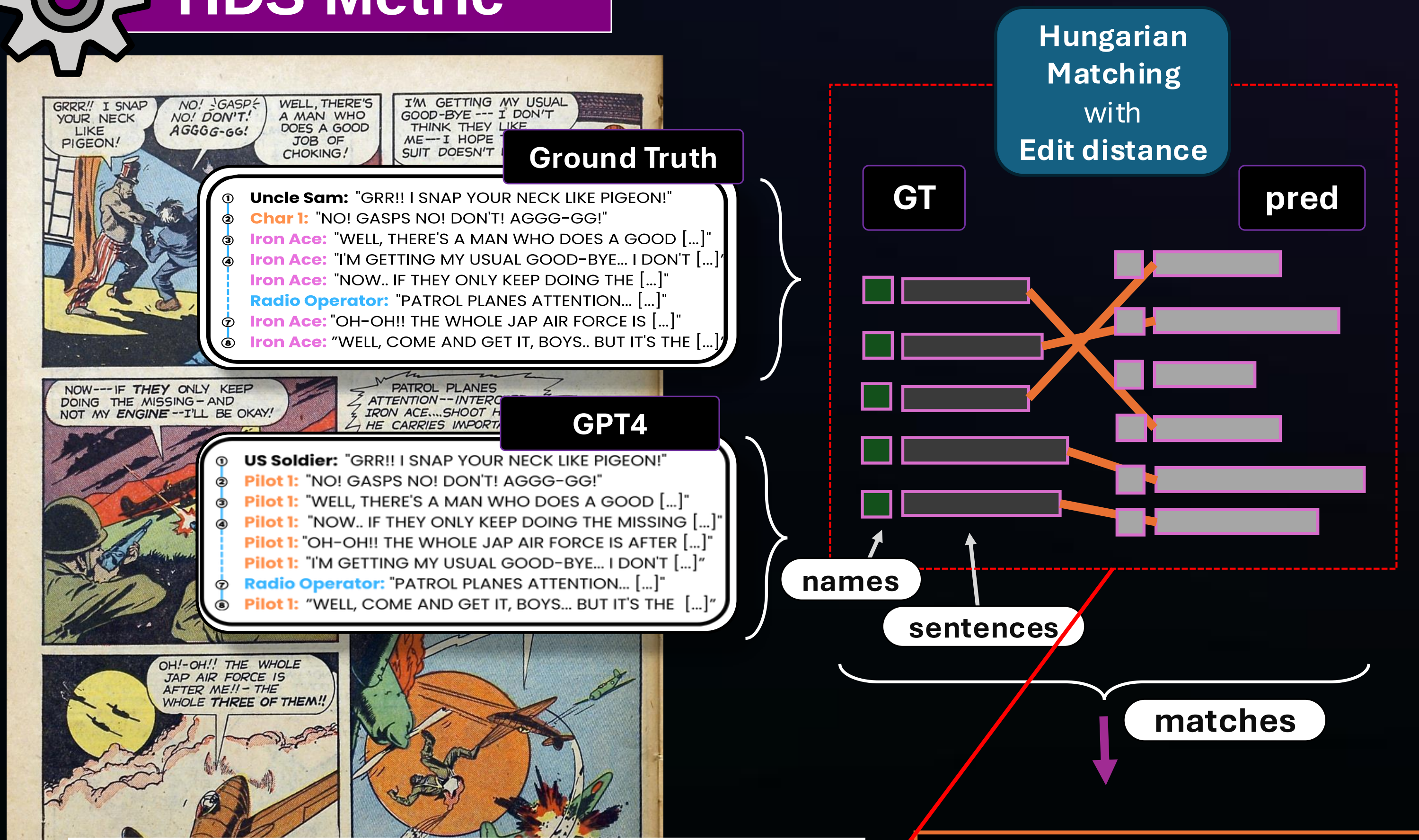


Narrator, Sailor 1, McWhustle, Captain
Matey, Sailor 2, Sailor 3, Sailor 4

1. **Narrator:** "As the tramp steamer SS. Clementine crosses the Equator [...]"
2. **Sailor 1:** "THAT DISGUISE DOESN'T FOOL ME! YOU'RE CHIEF ENGINEER [...]"
3. **McWhustle:** "HOOT, LADDIE! BEFORE YE CAN BE A SON OF NEPTUNE, YE [...]"
4. **McWhustle:** "HOOT, MON! WE'RE AGROUND ON A REEF!"

Captain: "THERE ARE NO REEFS IN THESE PARTS! BACK TO YOUR ENGINE [...]"
Captain: "WHAT'S WRONG WITH YOUR BLASTED ENGINES, MCWHUSTLE ?"
McWhustle: "WHY DON'T YE LEARN TO NAVIGATE YER SHIP, LASSITER! [...]"
McWhustle: "NO HARD FEELINGS, CAP'N! BUT THE ENGINES ARE DOING FINE!"
Captain: "I KNOW THAT, MAC... WE'VE BEEN SHIPMATES TOO LONG TO [...]"
Captain: "WE SHOULD BE MAKING HEADWAY... BUT SOMETHING'S [...]"
McWhustle: "MON, IT'S NO CANNY!"
Narrator: "Meanwhile, below..."
Matey: "THE SKIPPER' SAYS WE CAN'T HAVE RUN AGROUND!"
14. Sailor 2: "HEY, MATEY! LOOK!"
15. **Sailor 3:** "AHOY, TOPSIDES! STAND BY TO REPEL BOARDERS IN THE [...]"
16. Sailor 4: "BOARDERS IN THE STOKEHOLD? I NEVER HEARD OF SUCH A THING."

HDS Metric


GRRR!! I SNAP YOUR NECK LIKE PIGEON!
NO! GASP! NO! DON'T! AGGGG-GG!
WELL, THERE'S A MAN WHO DOES A GOOD JOB OF CHOKING!
I'M GETTING MY USUAL GOOD-BYE --- I DON'T THINK THEY LIKE ME --- I HOPE SUIT DOESN'T
NOW---IF THEY ONLY KEEP DOING THE MISSING-AND NOT MY ENGINE--I'LL BE OKAY!
PATROL PLANES ATTENTION--INTERC IRON ACE....SHOOT HE CARRIES IMPORT
OH!-OH!! THE WHOLE JAP AIR FORCE IS AFTER ME!!-THE WHOLE THREE OF THEM!!
Hungarian Matching with Edit distance
GT
pred
names
sentences
matches

Ground Truth

1. **Uncle Sam:** "GRR!! I SNAP YOUR NECK LIKE PIGEON!"
2. **Char 1:** "NO! GASPS NO! DON'T! AGGG-GG!"
3. **Iron Ace:** "WELL, THERE'S A MAN WHO DOES A GOOD [...]"
4. **Iron Ace:** "I'M GETTING MY USUAL GOOD-BYE... I DON'T [...]"

Iron Ace: "NOW.. IF THEY ONLY KEEP DOING THE [...]"
Radio Operator: "PATROL PLANES ATTENTION... [...]"

7. **Iron Ace:** "OH-OH!! THE WHOLE JAP AIR FORCE IS [...]"
8. **Iron Ace:** "WELL, COME AND GET IT, BOYS.. BUT IT'S THE [...]"

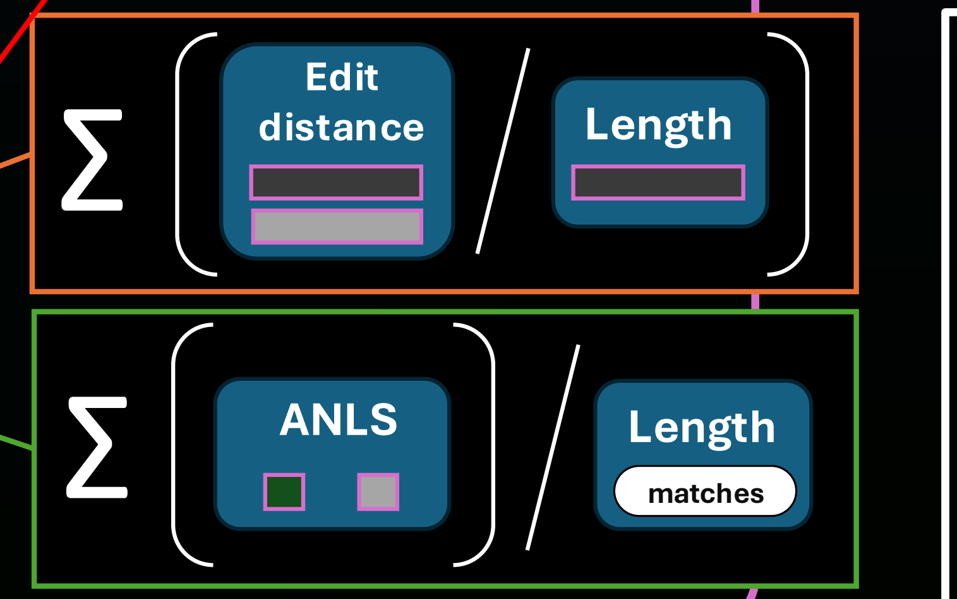
GPT4

1. **US Soldier:** "GRR!! I SNAP YOUR NECK LIKE PIGEON!"
2. **Pilot 1:** "NO! GASPS NO! DON'T! AGGG-GG!"
3. **Pilot 1:** "WELL, THERE'S A MAN WHO DOES A GOOD [...]"
4. **Pilot 1:** "NOW.. IF THEY ONLY KEEP DOING THE MISSING [...]"

Pilot 1: "OH-OH!! THE WHOLE JAP AIR FORCE IS AFTER [...]"
Pilot 1: "I'M GETTING MY USUAL GOOD-BYE... I DON'T [...]"

7. **Radio Operator:** "PATROL PLANES ATTENTION... [...]"
8. **Pilot 1:** "WELL, COME AND GET IT, BOYS... BUT IT'S THE [...]"

```
Algorithm 2 Hybrid Dialog Score
procedure EVALUATETRANSCRIPTION($model\_output, ground\_truth$)
    $matches \leftarrow$ find optimal matches($model\_output, ground\_truth$)
    $tot\_ed, char\_name\_score \leftarrow 0, 0, 0$
    for each $(mo, gt)$ in $matches$ do
        $edit\_dist \leftarrow$ calculate edit distance($mo.text, gt.text$)
        $tot\_ed \leftarrow tot\_ed + edit\_dist$ / len($gt.text$)
        $anls\_score \leftarrow$ calculate ANLS($mo.name, gt.name$)
        $char\_name\_score \leftarrow char\_name\_score + anls\_score$
    end for
    $tot\_ed \leftarrow 1 - tot\_ed$
    $char\_name\_score \leftarrow char\_name\_score$ / len($matches$)
    return $tot\_ed, char\_name\_score$
end procedure
```



Benchmarks

We evaluate models on a "per-task" setting as no model can perform all the 6 tasks together:

Task	Output	Metric	Baseline	Score
Object detection	box detection	mAP - R@100	Magi	78.6 - 67.9
Speaker identification	object indexes	R@#text	heuristic	0.68
Character Re-Id	cluster ids	AMI - NMI	DINOv2	0.29 - 0.51
Character Naming	names	ANLS	GPT-4	47.11
Dialog generation	list of tuples	HDS	GPT-4	93.14

This is Lele
if you like Comics/Manga and
Vision-Language, come and
talk to me!

and... read our SURVEY